What is Machine Learning?

Neil D. Lawrence

Department of Computer Science, University of Sheffield, U.K.

16th April 2012

Outline

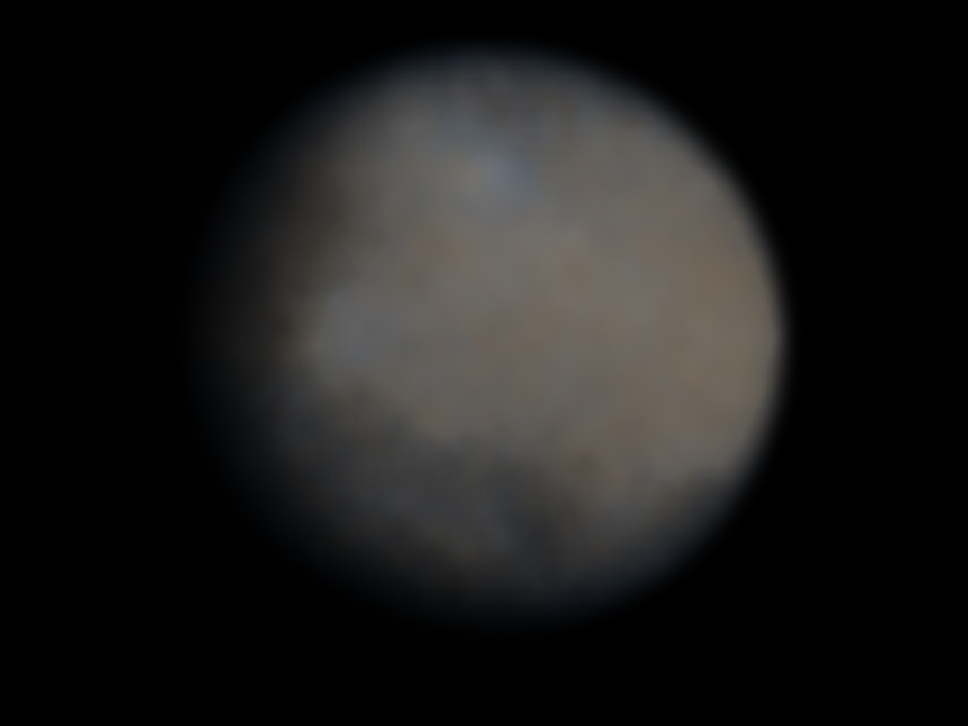
Introduction

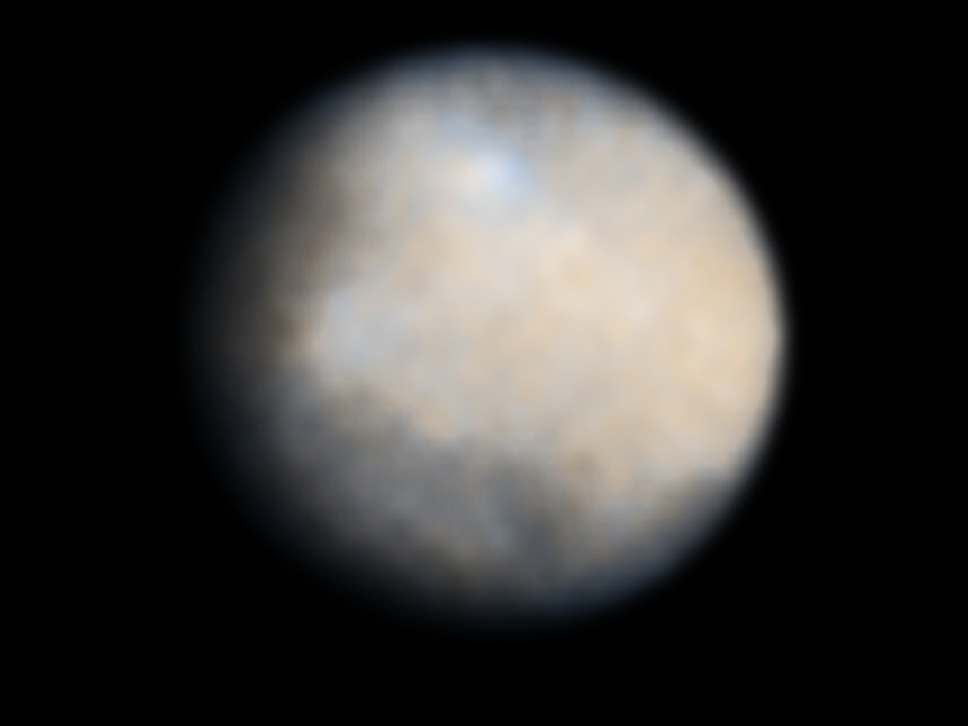
ML Motivation

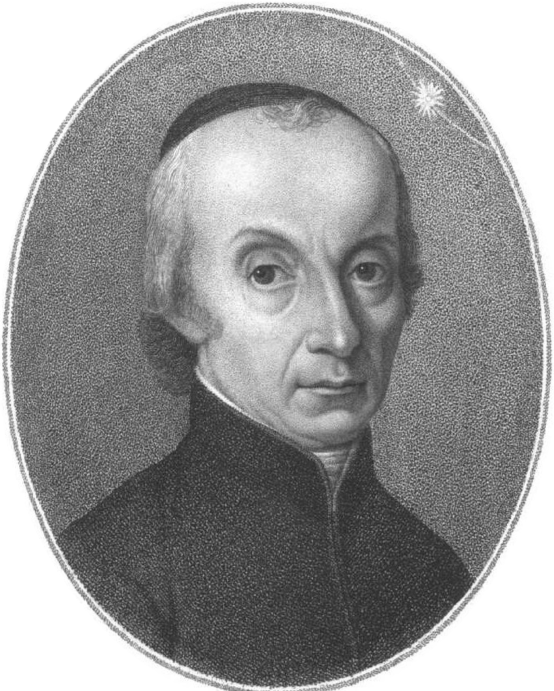
Outline

Introduction

ML Motivation

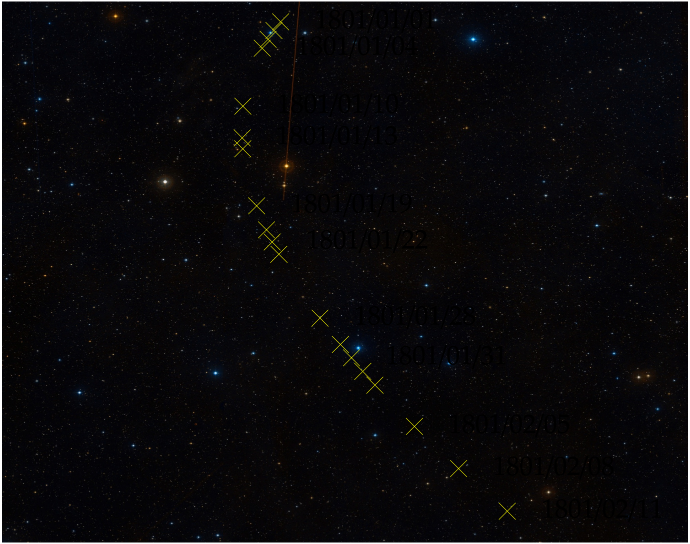






Beobachtungen des zu Palermo d. 1. Jan. 1801 von Prof. *Piazzi* neu entdeckten Gestirns.

1801	Mittlere Sonnen-Zeit	Gerade Aufsteig. in Zeit	Gerade Aufsteigung in Graden	Nördl. Abweich.	Geocentrische Länge	Geocentr. Breite	Ort der Sonne + 20″ Aberration	Logar. d. Distanz ⊙ ♁
	St ′ ″	St ′ ″	° ′ ″	° ′ ″	Z ° ′ ″	° ′ ″	Z ° ′ ″	
Jan. 1	8 43 27,8	3 27 11,25	51 47 48,8	15 37 43,5	1 23 22 58,3	3 6 42,1	9 11 1 30,9	9,9926156
2	8 39 4,6	3 26 53,85	51 43 27,8	15 41 5,5	1 23 19 44,3	3 2 24,9	9 12 2 28,6	9,9926317
3	8 34 53,3	3 26 38,4[illegible]	51 39 36,0	15 44 31,6	1 23 16 58,6	2 58 9,9	9 13 3 26,6	9,9926324
4	8 30 42,1	3 26 23,16	51 35 47,3	15 47 57,6	1 23 14 15,5	2 53 55,6	9 14 4 24,9	9,9926418
10	8 6 15,8	3 25 32,1[illegible]	51 23 1,5	16 10 32,0	1 23 7 59,1	2 29 0,6	9 20 10 17,5	9,9927641
11	8 2 17,5	3 25 29,73	51 22 26,0	. . . .	. . . .	. . . . .	. . . .	. . . . .
13	7 54 26,2	3 25 30,30	51 22 34,5	16 22 49,5	1 23 10 27,6	2 16 59,7	9 23 12 13,8	9,9928490
14	7 50 31,7	3 25 31,72	51 22 55,8	16 27 5,7	1 23 12 1,2	2 12 56,7	9 24 14 13,5	9,9928809
17	. . . . . .	. . . . . .	. . . . . .	16 40 13,0	. . . .	. . . . .	. . . .	. . . . .
18	7 35 11,3	3 25 55,[illegible]	51 28 45,0	. . . .	. . . .	. . . . .	. . . .	. . . . .
19	7 31 28,5	3 26 8,15	51 32 2,3	16 49 16,1	1 23 25 59,2	1 53 38,2	9 29 19 53,8	9,9930607
21	7 24 2,7	3 26 34,27	51 38 34,1	16 58 35,9	1 23 34 21,3	1 46 6,0	10 1 20 40,3	9,9931434
22	7 20 21,7	3 26 49,42	51 42 21,3	17 3 18,5	1 23 39 1,8	1 42 28,1	10 2 21 32,0	9,9931886
23	7 16 43,5	3 27 6,90	51 46 43,5	17 8 5,5	1 23 44 15,7	1 38 52,1	10 3 22 22,7	9,9932348
28	6 58 51,3	3 28 54,55	52 13 38,3	17 32 54,1	1 24 15 15,7	1 21 6,9	10 8 26 20,1	9,9935062
30	6 51 52,9	3 29 48,14	52 27 2,1	17 43 11,0	1 24 30 9,0	1 14 16,0	10 10 27 46,2	9,9936332
31	6 48 26,4	3 30 17,25	52 34 18,8	17 48 21,5	1 24 38 7,3	1 10 54,6	10 11 28 28,5	9,9937007
Febr. 1	6 44 59,9	3 30 47,2[illegible]	52 41 48,0	17 53 36,3	1 24 46 19,3	1 7 30,9	10 12 29 9,6	9,9937703
2	6 41 35,8	3 31 19,06	52 49 45,2	17 58 57,5	1 24 54 57,9	1 4 1[illegible],5	10 13 29 49,9	9,9938423
5	6 31 31,5	3 33 2,70	53 15 40,5	18 15 1,0	1 25 22 43,4	0 54 23,9	10 16 31 45,5	9,9940751
8	6 21 39,2	3 34 58,50	53 44 37,5	18 31 23,2	1 25 53 29,5	0 45 5,0	10 19 33 33,3	9,9943276
11	6 11 58,2	3 37 6,54	54 16 38,1	18 47 58,8	1 26 26 40,0	0 36 2,9	10 22 35 11,4	9,9945823


1801/01/01
1801/01/04
1801/01/10
1801/01/13
1801/01/19
1801/01/22
1801/01/28
1801/01/31
1801/02/05
1801/02/08
1801/02/11

hier in der Nähe der Quadratur der Einfluſs der Sonnen-Länge geringer iſt, als in andern Lagen. Dr. *Gauſs* glaubt daher, daſs es nicht undienlich wäre, wenn man die Fehler der Sonnentafeln aus ſehr genauen Beobachtungen für dieſe Zeiten beſtimmte, und die Örter der Sonne hiernach verbeſſerte. Dieſe *vierten* Elemente ſind nun folgende:

Sonnenferne . .	326° 27′ 38″	Hieraus:	
☊	81 0 44	gröſste Mittelp. Gleichung	9° 27′ 41″
Neigung . .	10 36 57		
Log. halb. gr. Axe	0,4420527	tägl. mittlere helioc. tropiſche Beweg.	770,″914
Excentricität . .	0,0825017		
Epoche 1800 31 Dec.	77° 36′ 34″		

Aus dieſen Elementen hat Dr. *Gauſs* folgende Örter der *Ceres Ferdinandea* im voraus berechnet. Die Zeit iſt mittlere für Mitternacht in *Palermo*.

1801	Geocentriſche Länge	Geocentriſche Breite nördl.	Logarith. des Abſtandes von der ♁	Logarith. des Abſtandes von der ☉	Verhältniſs der geſehenen Helligk.
	Z ° ′	° ′			
Nov. 25	5 20 16	9 25	0, 42181	0, 40468	0, 6102
Dec. 1	5 22 15	9 48	0, 40940	0, 40472	0, 6459
7	5 24 7	10 12	0, 39643	0, 40479	0, 6855
13	5 25 51	10 37	0, 38296	0, 40488	0, 7290
19	5 27 27	11 4	0, 36902	0, 40499	0, 7770
25	5 28 53	11 32	0, 35468	0, 40512	0, 8295
31	6 0 10	12 1	0, 34000	0, 40528	0, 8869

Sollte man den Ort des Planeten nach dieſen Elementen genauer, oder auf eine längere Zeit berechnen wollen: ſo ſetzen wir zu dieſem Behufe noch folgende Formeln hierher:

1) Zur

Excentricität . . 0,08[illegible] | [illegible]tropische Beweg. 770, 914
Epoche 1800 31 Dec. 77° 36′ 34″ |

Aus diesen Elementen hat Dr. *Gauss* folgende Örter der *Ceres Ferdinandea* im voraus berechnet. Die Zeit ist mittlere für Mitternacht in *Palermo*.

1801	Geocentrische Länge	Geocentrische Breite nördl.	Logarith. des Abstandes von der ♁	Logarith. des Abstandes von der ☉	Verhältniss der gesehenen Helligk.
	Z ° ′	° ′			
Nov. 25	5 20 16	9 25	0, 42181	0, 40468	0, 6102
Dec. 1	5 22 15	9 48	0, 40940	0, 40472	0, 6459
7	5 24 7	10 12	0, 39643	0, 40479	0, 6855
13	5 25 51	10 37	0, 38296	0, 40488	0, 7290
19	5 27 27	11 4	0, 36902	0, 40499	0, 7770
25	5 28 53	11 32	0, 35468	0, 40512	0, 8295
31	6 0 10	12 1	0, 34000	0, 40528	0, 8869

Sollte man den Ort des Planeten nach diesen Elementen genauer, oder auf eine längere Zeit berechnen wollen: so setzen wir zu diesem Behufe noch folgende Formeln hierher:

DELLA SCOPERTA

DEL NUOVO PIANETA

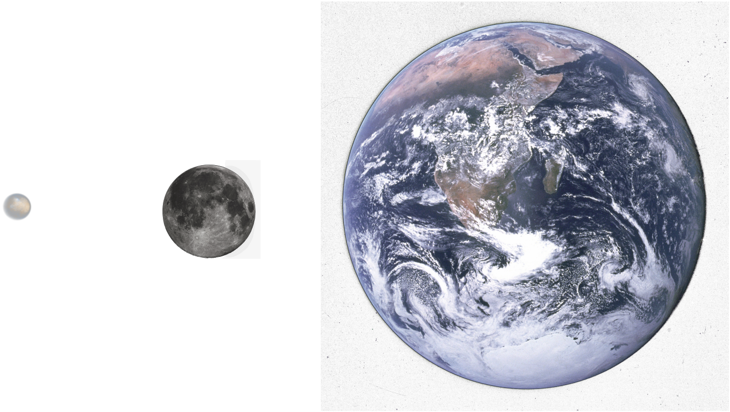
CERERE FERDINANDEA

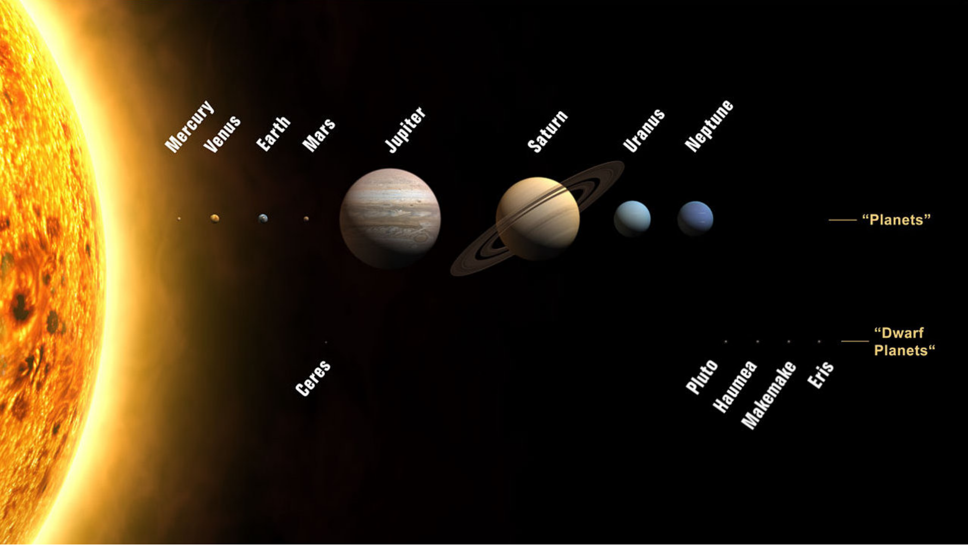
OTTAVO TRA I PRIMARJ DEL NOSTRO SISTEMA SOLARE.

PALERMO

1802

NELLA STAMPERIA REALE.




Mercury
Venus
Earth
Mars
Jupiter
Saturn
Uranus
Neptune
“Planets”
“Dwarf Planets“
Ceres
Pluto
Haumea
Makemake
Eris

What is Machine Learning?

data

- data: observations, could be actively or passively acquired (meta-data).

What is Machine Learning?

data +

- data: observations, could be actively or passively acquired (meta-data).

What is Machine Learning?

data + model

- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

What is Machine Learning?

data + model =

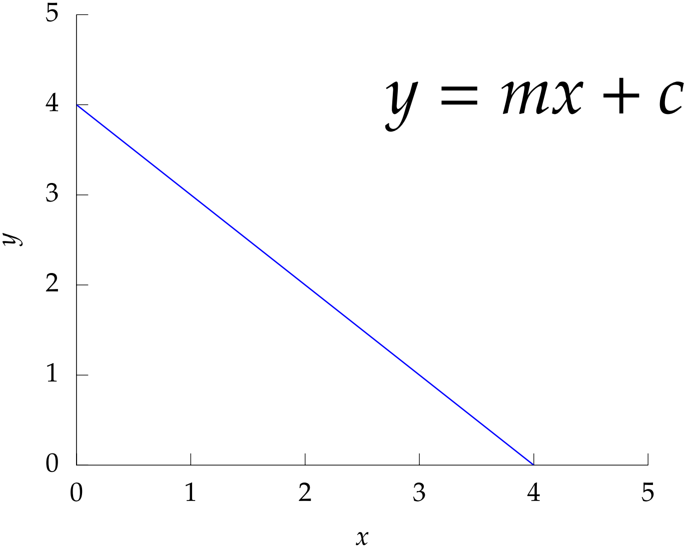
- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

What is Machine Learning?

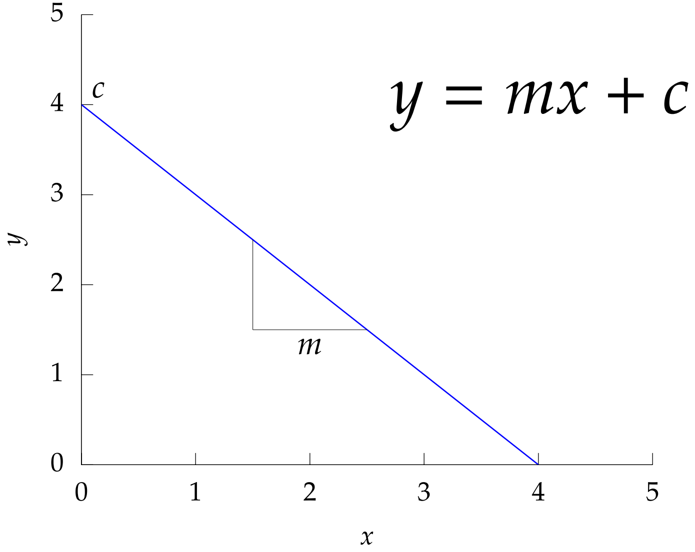
data + model = prediction

- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.
- prediction: an action to be taken or a categorization or a quality score.

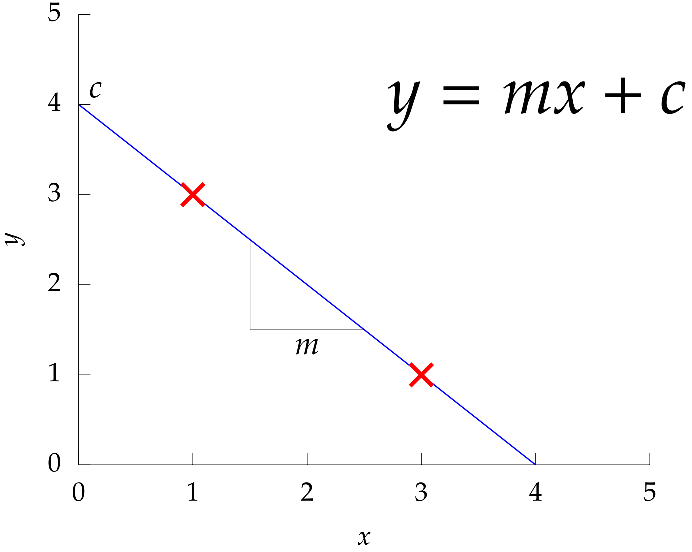
$$y = mx + c$$

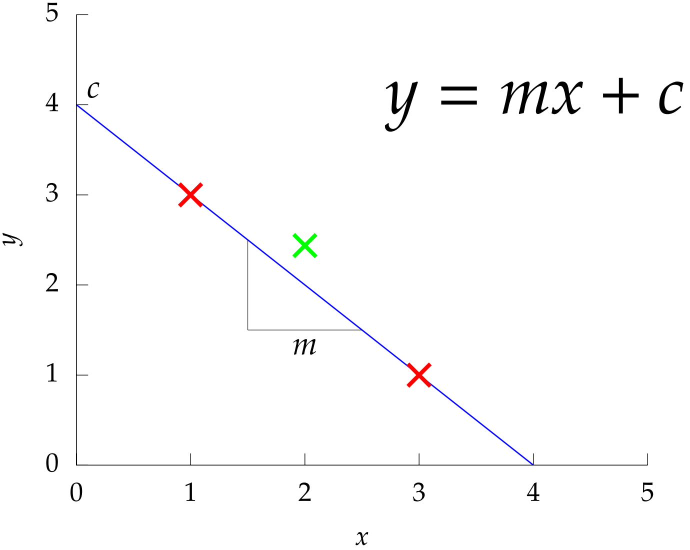


$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
x
y

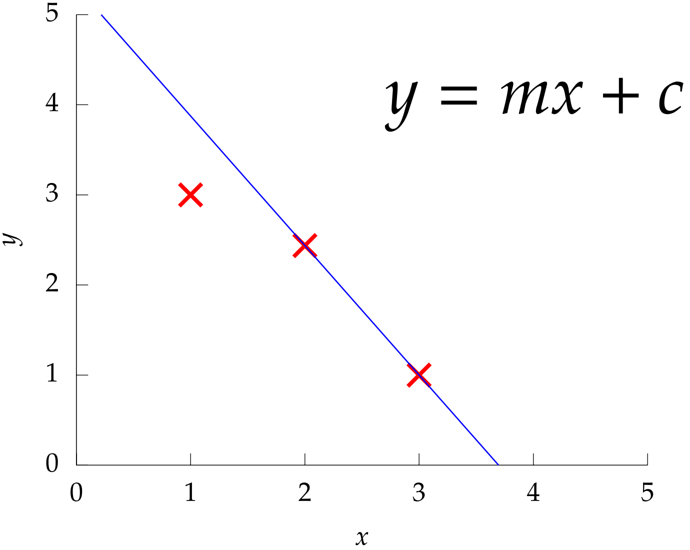


$y = mx + c$
c
m
0
1
2
3
4
5
0
1
2
3
4
5
y
x

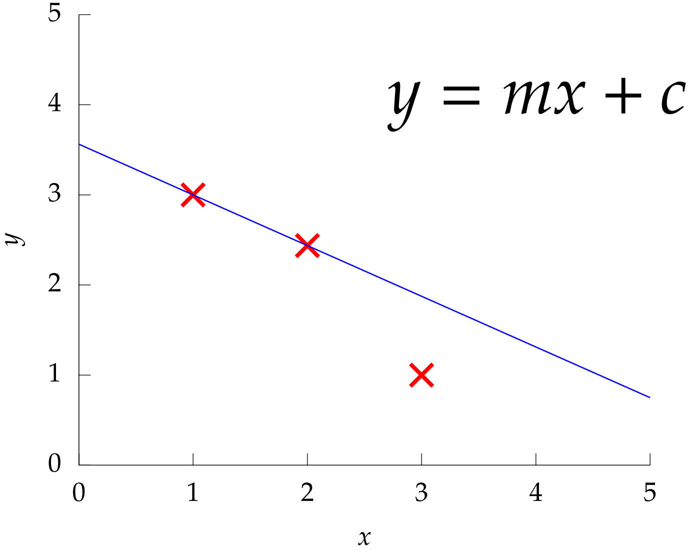

$y = mx + c$
c
m
x
y
0
1
2
3
4
5
0
1
2
3
4
5



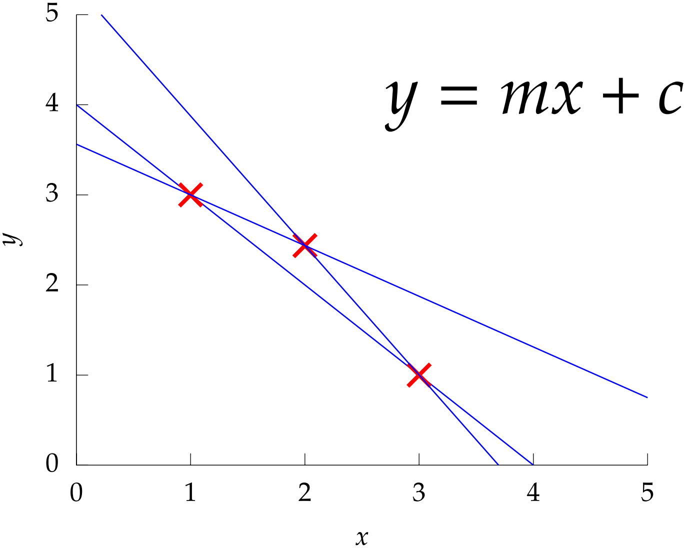
$y = mx + c$
c
m
0
1
2
3
4
5
x
y



$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
x
y



$y = mx + c$
5
4
3
2
1
0
y
0
1
2
3
4
5
x


$y = mx + c$
y
x
0
1
2
3
4
5
0
1
2
3
4
5

$y = mx + c$

point 1: $x = 1,\ y = 3$

$$3 = m + c$$

point 2: $x = 3,\ y = 1$

$$1 = 3m + c$$

point 3: $x = 2,\ y = 2.5$

$$2.5 = 2m + c$$

point 1: $x = 1$, $y = 3$

$$3 = m + c + \epsilon_1$$

point 2: $x = 3$, $y = 1$

$$1 = 3m + c + \epsilon_2$$

point 3: $x = 2$, $y = 2.5$

$$2.5 = 2m + c + \epsilon_3$$

Applications of Machine Learning

Handwriting Recognition : Recognising handwritten characters. For example LeNet http://bit.ly/d26fwK.

Friend Indentification : Suggesting friends on social networks https://www.facebook.com/help/501283333222485

Ranking : Learning relative skills of on line game players, the TrueSkill system http://research.microsoft.com/en-us/projects/trueskill/.

Collaborative Filtering : Prediction of user preferences for items given purchase history. For example the Netflix Prize http://www.netflixprize.com/.

Internet Search : For example Ad Click Through rate prediction http://bit.ly/a7XLH4.

News Personalisation : For example Zite http://www.zite.com/.

Game Play Learning : For example, learning to play Go http://bit.ly/cV77zM

- Arises from the Connectionist movement in AI.
  http://en.wikipedia.org/wiki/Connectionism

- Arises from the Connectionist movement in AI.
  http://en.wikipedia.org/wiki/Connectionism
- Early Connectionist research focused on models of the brain.

Frank Rosenblatt's Perceptron

- Rosenblatt's perceptron (**?**) based on simple model of a neuron (**?**) and a learning algorithm.

Figure: Frank Rosenblatt in 1950 (source: Cornell University Library)

Vladmir Vapnik's Statistical Learning Theory

- Later machine learning research focused on theoretical foundations of such models and their capacity to learn (**?**).

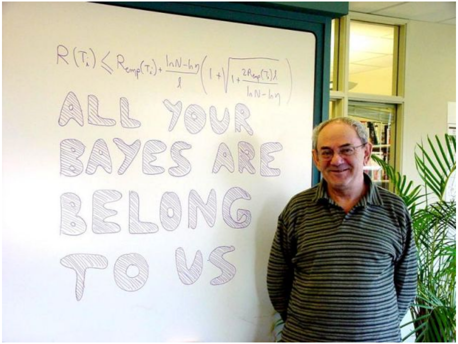


Figure: Vladimir Vapnik "All Your Bayes ..." (source http://lecun.com/ex/fun/index.html), see also http://bit.ly/qfd2mU.

Personal View

- Machine learning benefited greatly by incorporating ideas from psychology, but not being afraid to incorporate rigorous theory.

Machine Learning Today

An extension of statistics?

- Early machine learning viewed with scepticism by statisticians.

Machine Learning Today

An extension of statistics?

- Early machine learning viewed with scepticism by statisticians.
- Modern machine learning and statistics interact to both communities benefits.

Machine Learning Today

An extension of statistics?

- Early machine learning viewed with scepticism by statisticians.
- Modern machine learning and statistics interact to both communities benefits.
- *Personal view*: statistics and machine learning are fundamentally different. Statistics aims to provide a human with the tools to analyze data. Machine learning wants to replace the human in the processing of data.

- For the moment the two overlap strongly. But they are not the same field!

- For the moment the two overlap strongly. But they are not the same field!
- Machine learning also has overlap with Cognitive Science.

- For the moment the two overlap strongly. But they are not the same field!
- Machine learning also has overlap with Cognitive Science.
- Mathematical formalisms of a problem are helpful, but they can hide facts: i.e. the fallacy that "aerodynamically a bumble bee can't fly". Clearly a limitation of the model rather than fact.

Machine Learning Today

Mathematics and Bumblebees

- For the moment the two overlap strongly. But they are not the same field!
- Machine learning also has overlap with Cognitive Science.
- Mathematical formalisms of a problem are helpful, but they can hide facts: i.e. the fallacy that "aerodynamically a bumble bee can't fly". Clearly a limitation of the model rather than fact.
- Mathematical foundations are still very important though: they help us understand the capabilities of our algorithms.

- For the moment the two overlap strongly. But they are not the same field!
- Machine learning also has overlap with Cognitive Science.
- Mathematical formalisms of a problem are helpful, but they can hide facts: i.e. the fallacy that "aerodynamically a bumble bee can't fly". Clearly a limitation of the model rather than fact.
- Mathematical foundations are still very important though: they help us understand the capabilities of our algorithms.
- But we mustn't restrict our ambitions to the limitations of current mathematical formalisms. That is where humans give inspiration.

Statistics

What's in a Name?

- Early statistics had great success with the idea of statistical proof.

- Early statistics had great success with the idea of statistical proof.

  Question: I computed the mean of these two tables of numbers (a statistic). They are different. Does this "prove" anything?

- Early statistics had great success with the idea of statistical proof.

  Question: I computed the mean of these two tables of numbers (a statistic). They are different. Does this "prove" anything?

  Answer: it depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.

- Hypothesis testing: questions you can ask about your data are quite limiting.

- Early statistics had great success with the idea of statistical proof.

  Question: I computed the mean of these two tables of numbers (a statistic). They are different. Does this "prove" anything?

  Answer: it depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.

- Hypothesis testing: questions you can ask about your data are quite limiting.
- This can have the affect of limiting science too.

Statistics

What's in a Name?

- Early statistics had great success with the idea of statistical proof.

  Question: I computed the mean of these two tables of numbers (a statistic). They are different. Does this "prove" anything?

  Answer: it depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.

- Hypothesis testing: questions you can ask about your data are quite limiting.
- This can have the affect of limiting science too.
- Many successes: crop fertilization, clinical trials, brewing, polling.

Statistics
What's in a Name?

- Early statistics had great success with the idea of statistical proof.

  Question: I computed the mean of these two tables of numbers (a statistic). They are different. Does this "prove" anything?

  Answer: it depends on how the numbers are generated, how many there are and how big the difference. Randomization is important.
- Hypothesis testing: questions you can ask about your data are quite limiting.
- This can have the affect of limiting science too.
- Many successes: crop fertilization, clinical trials, brewing, polling.
- Many open questions: e.g. causality.

Early 20th Century Statistics

- Many statisticians were Edwardian English gentleman.

Figure: William Sealy Gosset in 1908

Statistics and Machine Learning

Statisticians want to turn humans into computers. Machine learners want to turn computers into humans. We meet somewhere in the middle.

NDL 2012/06/16

Statistics

- Cricket and Baseball are two games with a lot of “statistics”.

Statistics

- Cricket and Baseball are two games with a lot of "statistics".
- The study of the meaning behind these numbers is "mathematical statistics" often abbreviated to "statistics".

Machine Learning and Probability

- The world is an *uncertain* place.

Machine Learning and Probability

- The world is an *uncertain* place.

  Epistemic uncertainty: uncertainty arising through lack of knowledge. (What colour socks is that person wearing?)

Machine Learning and Probability

- The world is an *uncertain* place.

  Epistemic uncertainty: uncertainty arising through lack of knowledge. (What colour socks is that person wearing?)

  Aleatoric uncertainty: uncertainty arising through an underlying stochastic system. (Where will a sheet of paper fall if I drop it?)

Probability: A Framework to Characterise Uncertainty

- We need a framework to characterise the uncertainty.

Probability: A Framework to Characterise Uncertainty

- We need a framework to characterise the uncertainty.
- In this course we make use of probability theory to characterise uncertainty.

Richard Price

- Welsh philosopher and essay writer.
- Edited **Thomas Bayes**'s essay which contained foundations of Bayesian philosophy.

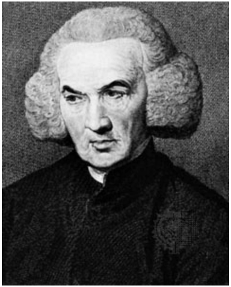


Figure: Richard Price, 1723–1791. (source Wikipedia)

Laplace

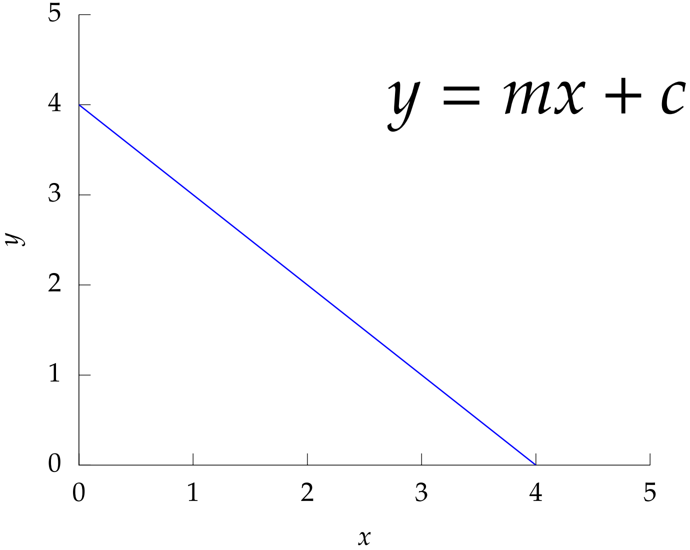
- French Mathematician and Astronomer.

Figure: Pierre-Simon Laplace, 1749–1827. (source Wikipedia)

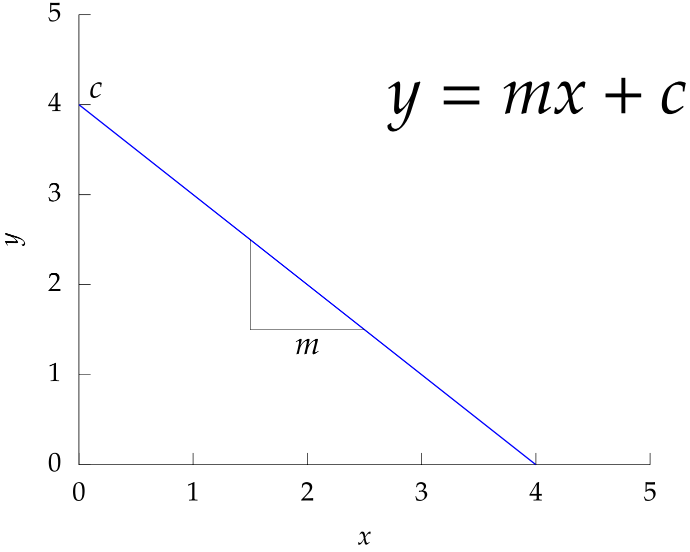

Tony Ryan


Hot
100°
0°
cold
Sheffield Festival of
& Engineering

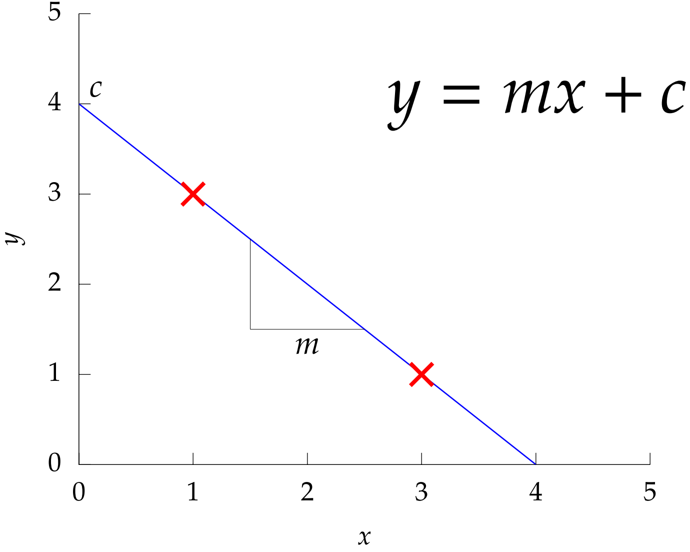
$$y = mx + c$$



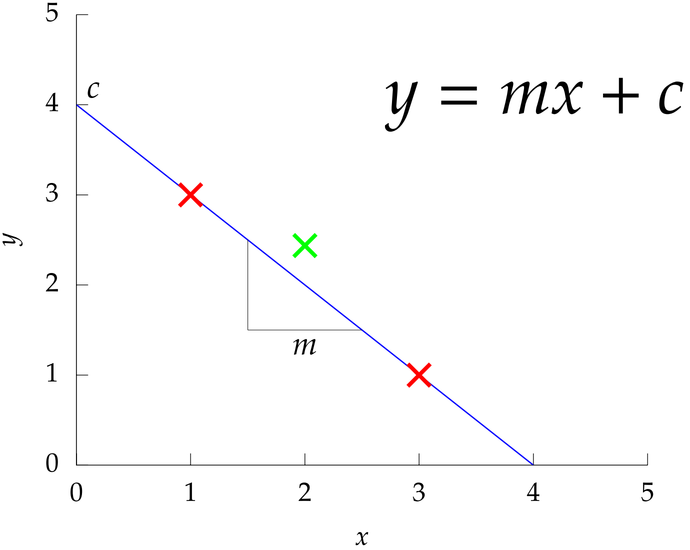
$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
x
y



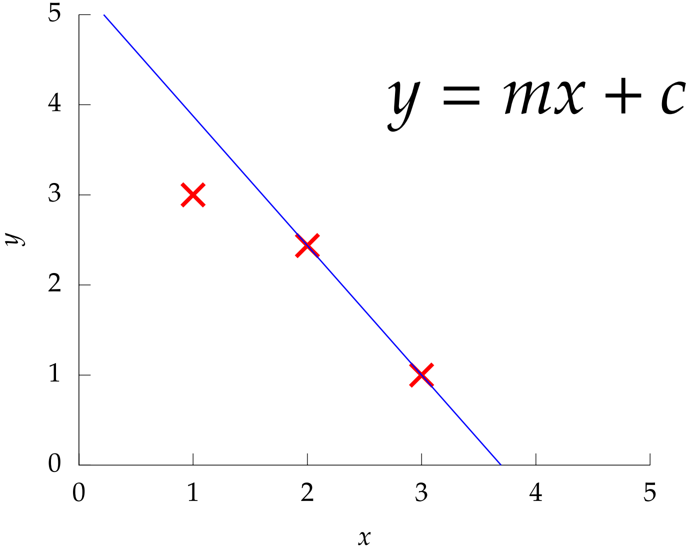
$y = mx + c$
c
m
0
1
2
3
4
5
0
1
2
3
4
5
x
y



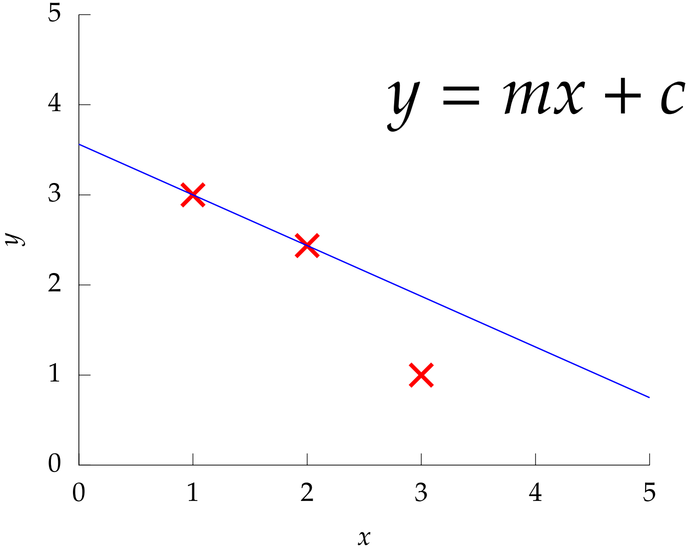
$y = mx + c$
c
m
x
y
0
1
2
3
4
5



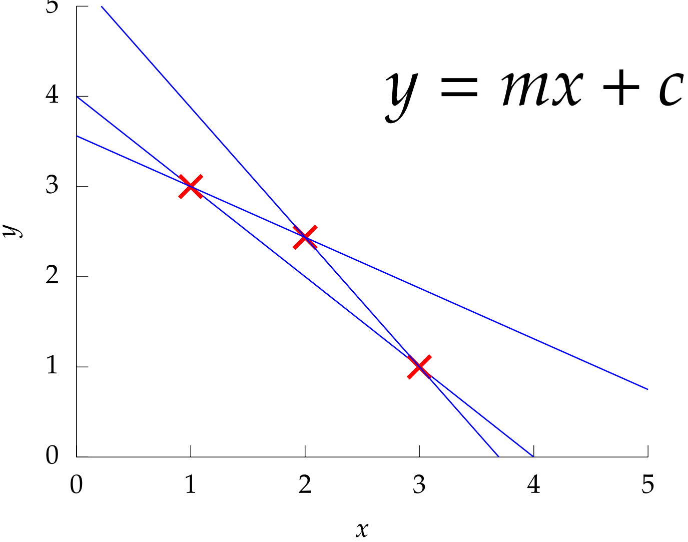
$y = mx + c$
c
m
y
x
0
1
2
3
4
5
0
1
2
3
4
5



$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
y
x



$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
x
y



$y = mx + c$
0
1
2
3
4
5
0
1
2
3
4
5
x
y

point 1: $x = 1$, $y = 3$

$$3 = m + c$$

point 2: $x = 3$, $y = 1$

$$1 = 3m + c$$

point 3: $x = 2$, $y = 2.5$

$$2.5 = 2m + c$$

point 1: $x = 1$, $y = 3$

$$3 = m + c + \epsilon_1$$

point 2: $x = 3$, $y = 1$

$$1 = 3m + c + \epsilon_2$$

point 3: $x = 2$, $y = 2.5$

$$2.5 = 2m + c + \epsilon_3$$

other, we say that its choice is an effect without a cause. It is then, says Leibnitz, the blind chance of the Epicureans. The contrary opinion is an illusion of the mind, which, losing sight of the evasive reasons of the choice of the will in indifferent things, believes that choice is determined of itself and without motives.

We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. The human mind offers, in the perfection which it has been able to give to astronomy, a feeble idea of this intelligence. Its discoveries in mechanics and geometry, added to that of universal gravity, have enabled it to comprehend in the same analytical expressions the past and future states of the system of the world. Applying the same method to some other objects of its knowledge, it has succeeded in referring to general laws observed phenomena and in foreseeing those which given circumstances ought to produce. All these efforts in the search for truth tend to lead it back continually to the vast intelligence which we have just mentioned, but from which it will always remain infinitely removed. This tendency, peculiar to the human race, is that which renders it superior to animals; and their progress choice is determined of itself and without motives.

We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. The human mind offers, in the perfection which it has

w.ted.com/talks/neil_burgess_how_your_brain_tells_you_where_you height: "The day will come when, by study pursued through several ages, the things now concealed will appear with evidence; and posterity will be astonished that truths so clear had escaped us." Clairaut then undertook to submit to analysis the perturbations which the comet had experienced by the action of the two great planets, Jupiter and Saturn; after immense calculations he fixed its next passage at the perihelion toward the beginning of April, 1759, which was actually verified by observation. The regularity which astronomy shows us in the movements of the comets doubtless exists also in all phenomena.

The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance.

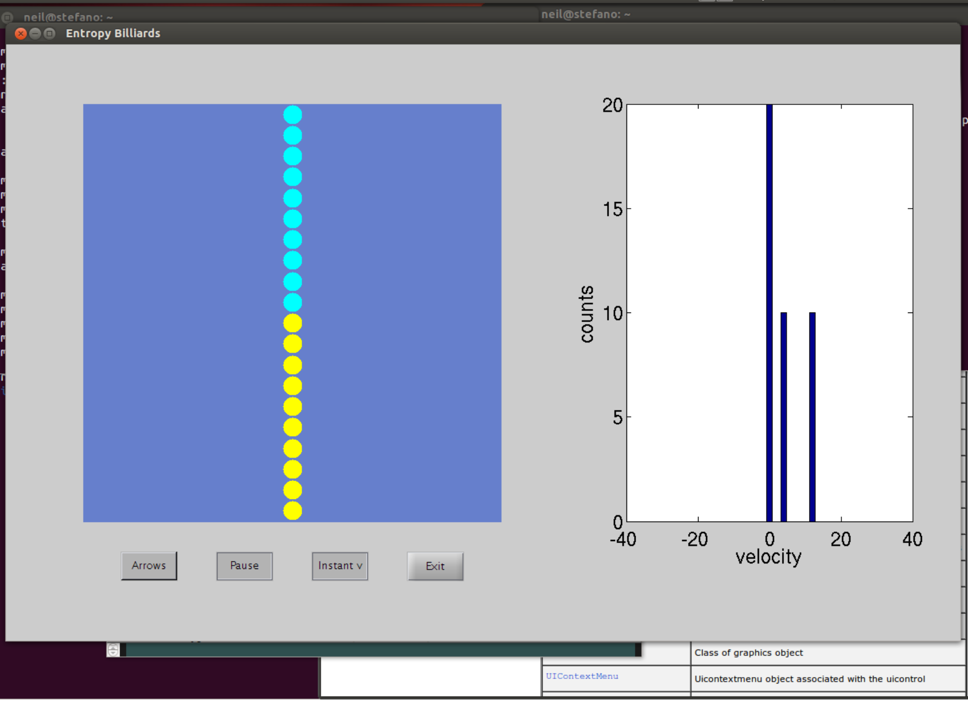
Probability is relative, in part to this ignorance, in part to our knowledge. We know that of three or a greater number of events a single one ought to occur; but nothing induces us to believe that one of them will occur rather than the others. In this state of indecision it is impossible for us to announce their occurrence with certainty. It is, however, probable that one of these events, chosen at will, will not occur because we see several cases equally possible which exclude its occurrence, while only a single one favors it.

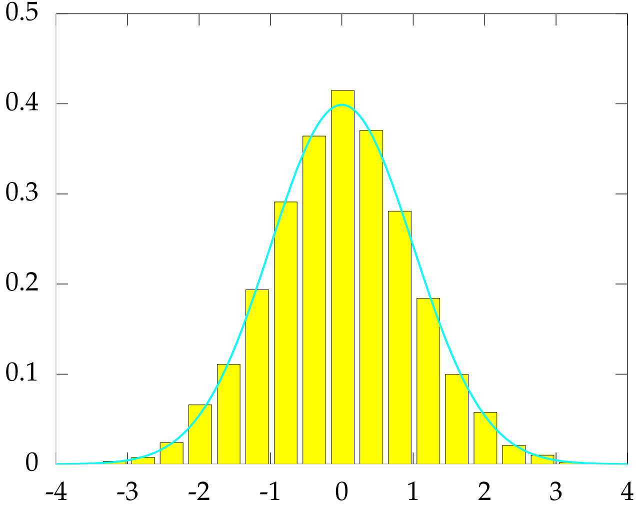
The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of shows us in the movements of the comets doubtless exists also in all phenomena.

The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance.

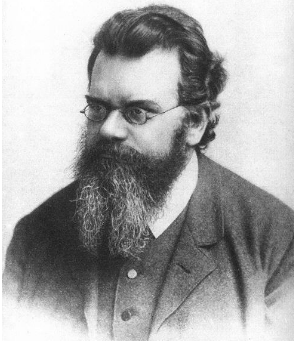
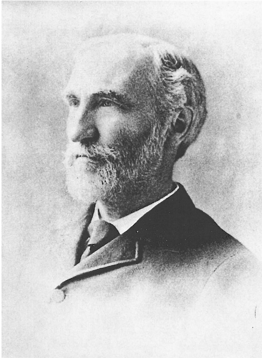
Probability is relative, in part to this ignorance, in part to our knowledge. We know that of three or a

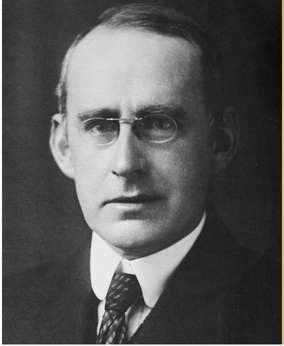

DANIEL BERNOVLLIVS
Med. D. Professor honorarius Academiæ Imper.
Petropolitanæ, Anatomiæ et Botanicæ P.P.O. in
Academia Basileensi.
Nat. d. 29. Ian. A.S.R. MDCC.
Dec. III.


neil@stefano: ~
neil@stefano: ~
Entropy Billiards
20
15
10
5
0
counts
-40
-20
0
20
40
velocity
Arrows
Pause
Instant v
Exit
Class of graphics object
UIContextMenu
Uicontextmenu object associated with the uicontrol



0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4



THE NATURE
OF THE
PHYSICAL WORLD

by

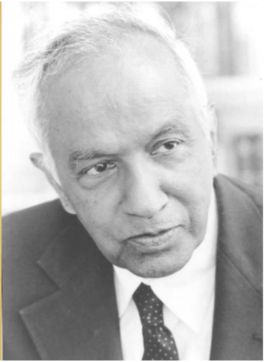
A. S. EDDINGTON
M.A., LL.D., D.SC., F.R.S.
Plumian Professor of Astronomy
in the
University of Cambridge

THE
GIFFORD LECTURES
1927

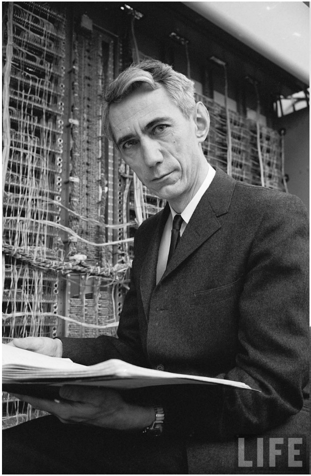
NEW YORK:
THE MACMILLAN COMPANY
CAMBRIDGE, ENGLAND:
AT THE UNIVERSITY PRESS
1929

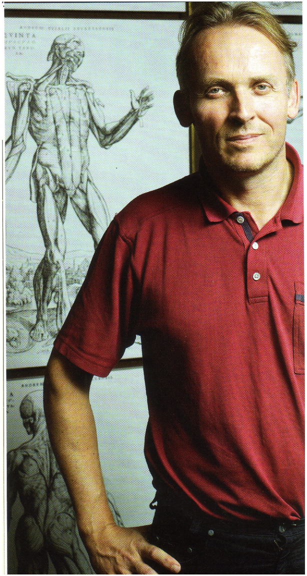
The uniform march of a regiment is not the only form of organised motion; the organised evolutions of a stage chorus have their natural analogue in sound waves. A common measure can now be applied to all forms of organisation. Any loss of organisation is equitably measured by the chance against its recovery by an accidental coincidence. The chance is absurd regarded as a contingency, but it is precise as a measure.

The practical measure of the random element which can increase in the universe but can never decrease is called *entropy*. Measuring by entropy is the same as measuring by the chance explained in the last paragraph, only the unmanageably large numbers are transformed (by a simple formula) into a more convenient scale of reckoning. Entropy continually increases. We can, by isolating parts of the world and postulating rather idealised conditions in our problems, arrest the increase, but we cannot turn it into a decrease. That would involve something much worse than a violation of an ordinary law of Nature, namely, an improbable coincidence. The law that entropy always increases—the second law of thermodynamics—holds, I think, the supreme position among the laws of Nature. If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations—then so much the worse for Maxwell's equations. If it is found to be contradicted by observation—well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation. This exaltation of the second law is not unreasonable. There are other laws which we have strong reason to believe in, and we feel that a hypothesis which violates them is highly

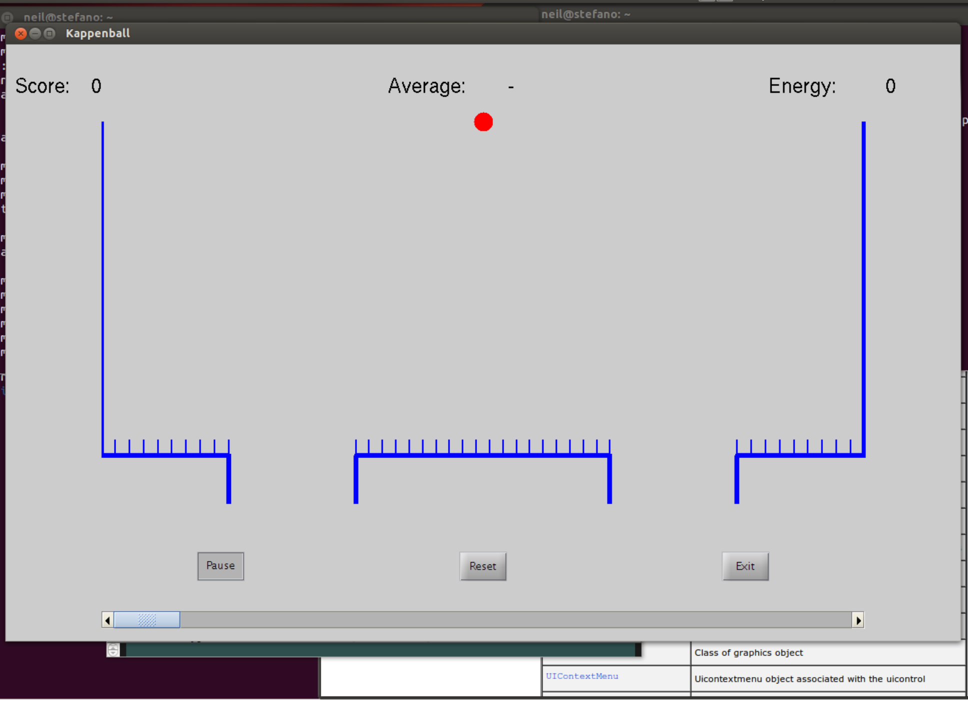


ordinary law of Nature, namely, an improbable coincidence. The law that entropy always increases—the second law of thermodynamics—holds, I think, the supreme position among the laws of Nature. If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations—then so much the worse for Maxwell's equations. If it is found to be contradicted by observation—well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation. This exaltation


LIFE

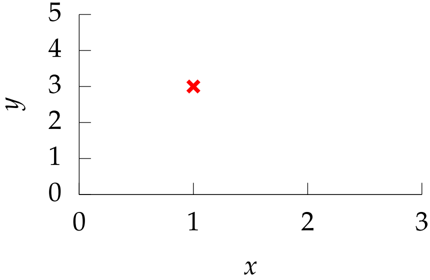
http://videolectures.net/aispds08_kappen_easop/


neil@stefano: ~
neil@stefano: ~
Kappenball
Score: 0
Average: -
Energy: 0
Pause
Reset
Exit
Class of graphics object
UIContextMenu
Uicontextmenu object associated with the uicontrol

Underdetermined System

What about two unknowns and *one* observation?

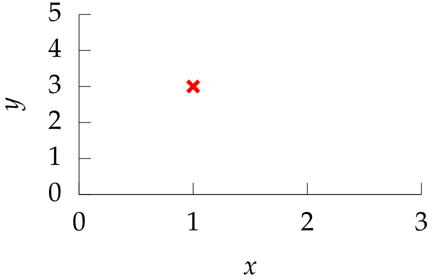
$$y_1 = mx_1 + c$$



Underdetermined System

Can compute m given c.

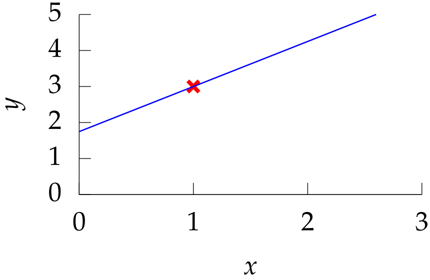
$$m = \frac{y_1 - c}{x}$$



Underdetermined System

Can compute m given c.

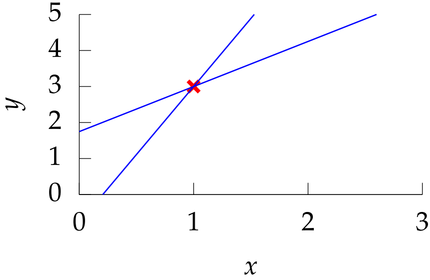
$$c = 1.75 \implies m = 1.25$$



Underdetermined System

Can compute m given c.

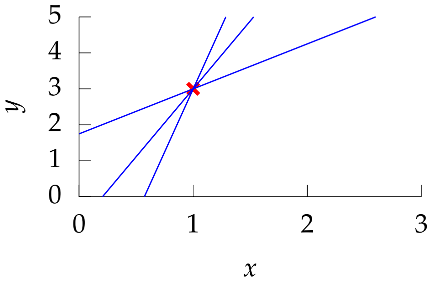
$$c = -0.777 \implies m = 3.78$$



Underdetermined System

Can compute m given c.

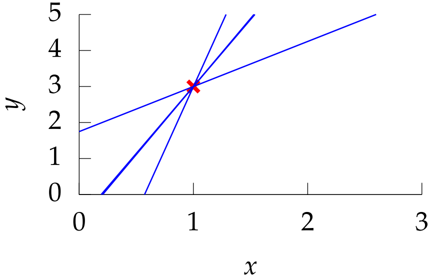
$$c = -4.01 \implies m = 7.01$$



Underdetermined System

Can compute m given c.

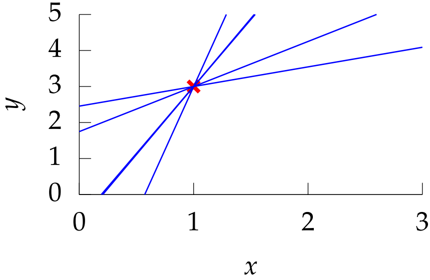
$$c = -0.718 \Longrightarrow m = 3.72$$



Underdetermined System

Can compute m given c.

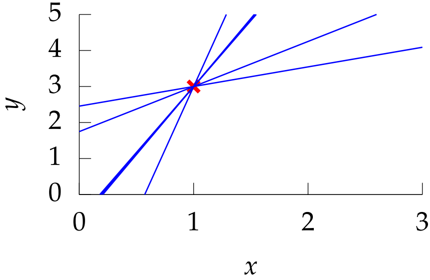
$$c = 2.45 \implies m = 0.545$$



Underdetermined System

Can compute m given c.

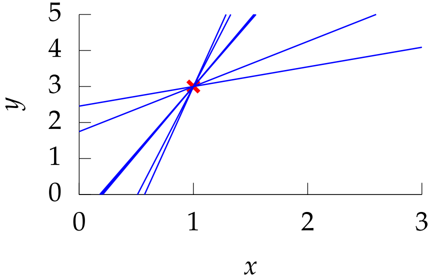
$$c = -0.657 \implies m = 3.66$$



Underdetermined System

Can compute m given c.

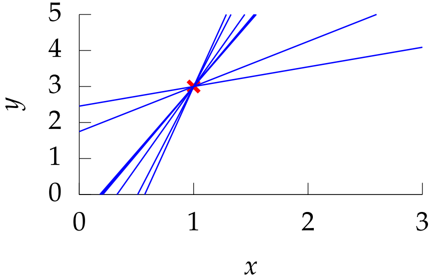
$$c = -3.13 \Longrightarrow m = 6.13$$



Underdetermined System

Can compute m given c.

$$c = -1.47 \Longrightarrow m = 4.47$$

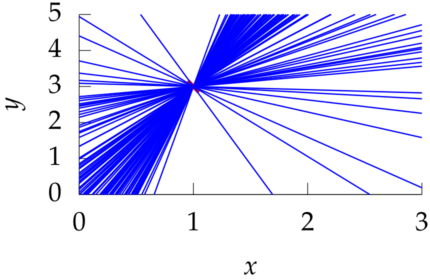


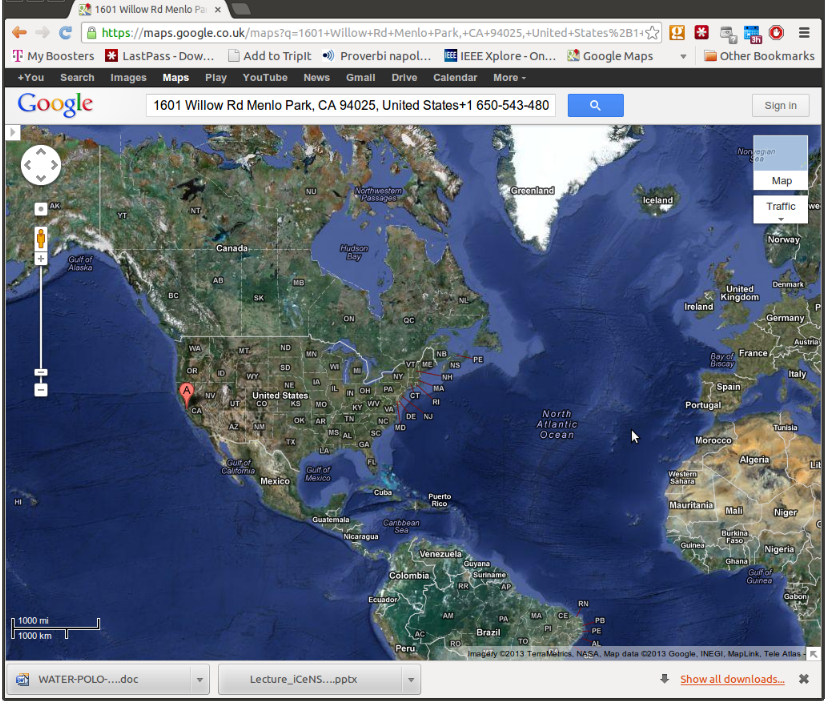
Underdetermined System

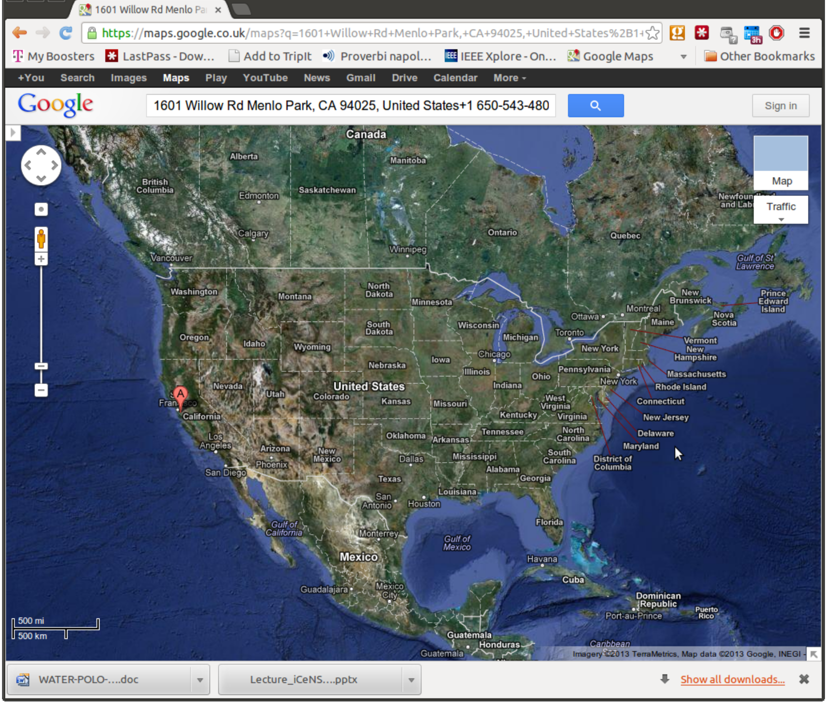
Can compute m given c. Assume

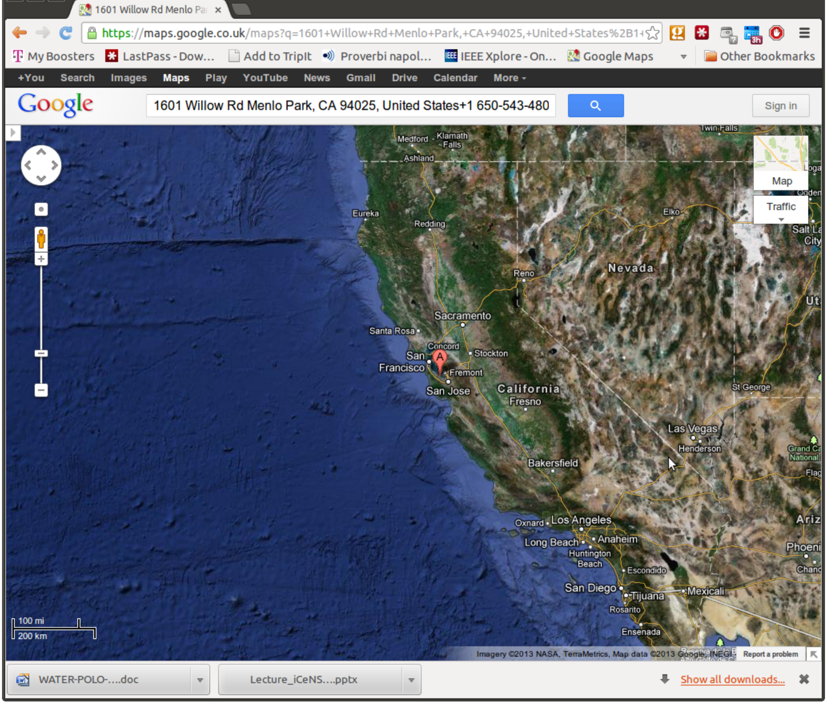
$$c \sim \mathcal{N}(0, 4),$$

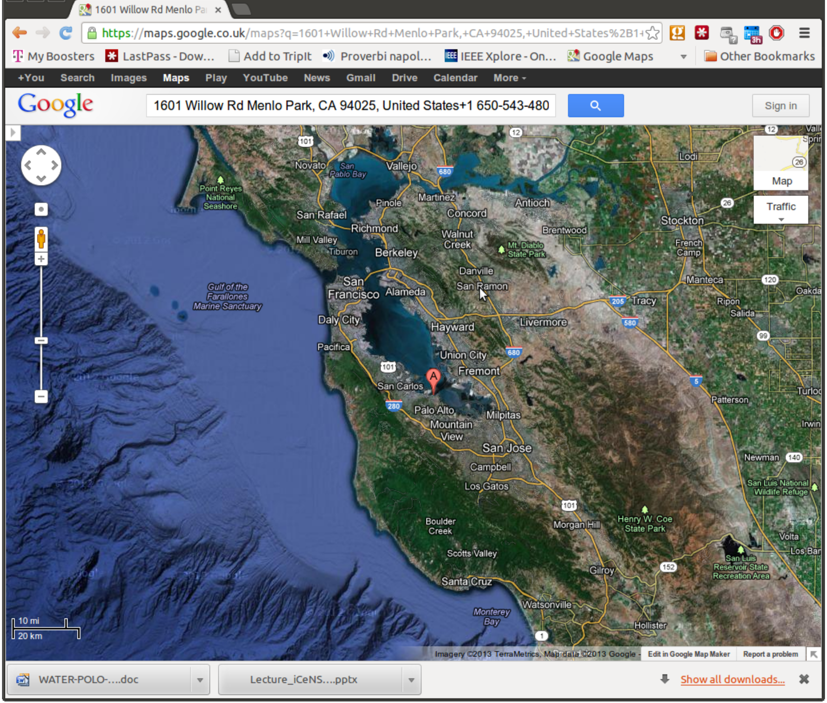
we find a distribution of solutions.

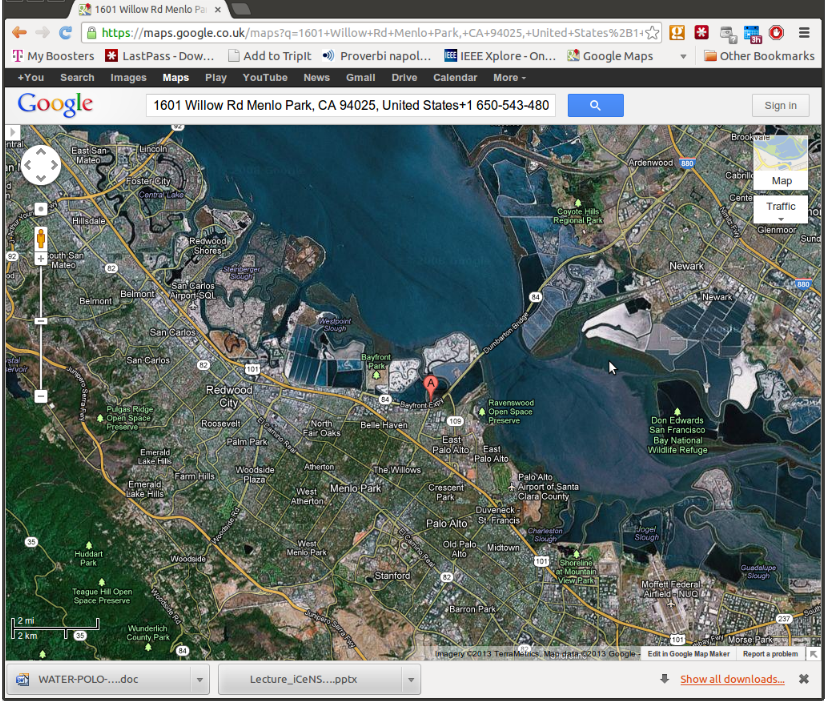


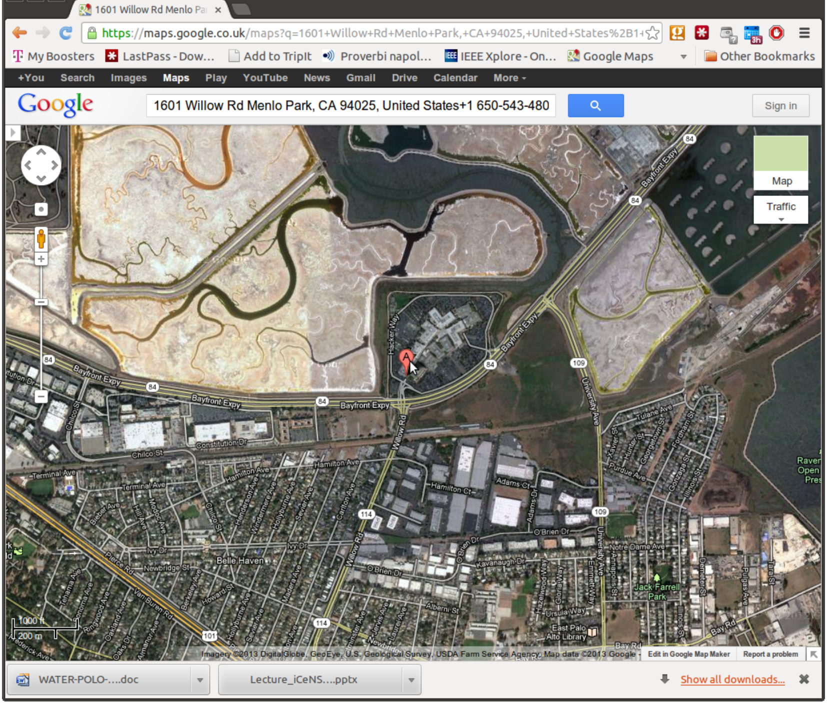

1601 Willow Rd Menlo Pa
https://maps.google.co.uk/maps?q=1601+Willow+Rd+Menlo+Park,+CA+94025,+United+States%2B1+
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+You
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
1601 Willow Rd Menlo Park, CA 94025, United States+1 650-543-480
Sign in
Map
Traffic
Greenland
Iceland
Norway
Canada
Hudson Bay
Northwestern Passages
Gulf of Alaska
United States
North Atlantic Ocean
United Kingdom
Ireland
Denmark
Germany
France
Bay of Biscay
Spain
Portugal
Italy
Austria
Tunisia
Morocco
Algeria
Western Sahara
Mauritania
Mali
Niger
Burkina Faso
Guinea
Ghana
Nigeria
Gulf of Guinea
Gabon
Gulf of California
Gulf of Mexico
Mexico
Cuba
Puerto Rico
Caribbean Sea
Guatemala
Nicaragua
Venezuela
Guyana
Suriname
Colombia
Ecuador
Peru
Brazil
1000 mi
1000 km
Imagery ©2013 TerraMetrics, NASA, Map data ©2013 Google, INEGI, MapLink, Tele Atlas -
WATER-POLO-....doc
Lecture_iCeNS....pptx
Show all downloads...

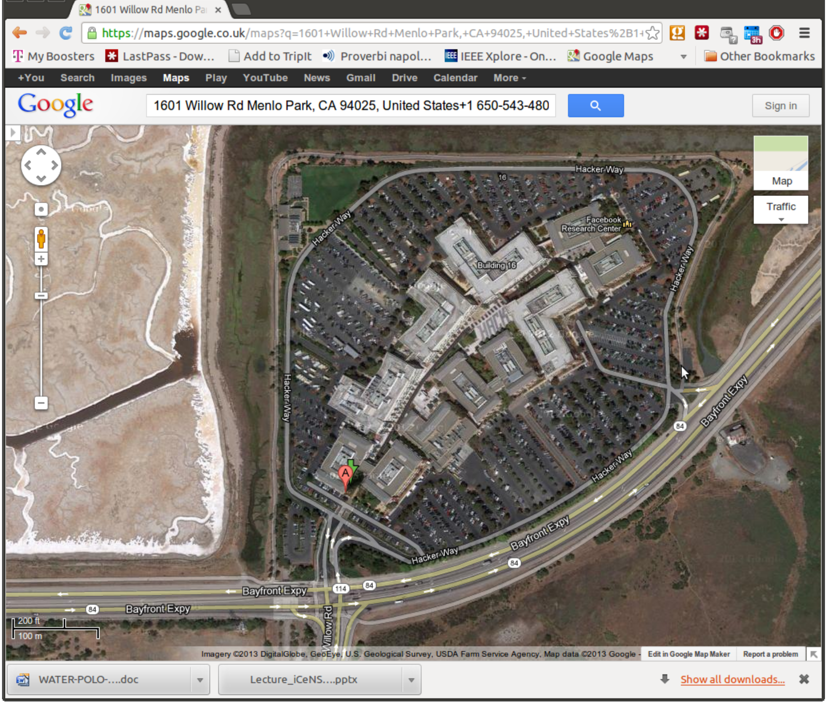

1601 Willow Rd Menlo Pa
https://maps.google.co.uk/maps?q=1601+Willow+Rd+Menlo+Park,+CA+94025,+United+States%2B1
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+You
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
1601 Willow Rd Menlo Park, CA 94025, United States+1 650-543-480
Sign in
Map
Traffic
Canada
Alberta
Manitoba
British Columbia
Edmonton
Saskatchewan
Calgary
Winnipeg
Ontario
Quebec
Vancouver
Gulf of St Lawrence
Washington
Montana
North Dakota
Minnesota
New Brunswick
Prince Edward Island
Nova Scotia
Montreal
Ottawa
Maine
Oregon
Idaho
Wyoming
South Dakota
Wisconsin
Michigan
Toronto
Vermont
New Hampshire
New York
Chicago
Iowa
Nebraska
Illinois
Indiana
Ohio
Pennsylvania
Massachusetts
Rhode Island
Connecticut
New Jersey
Delaware
Maryland
District of Columbia
Nevada
Utah
United States
Colorado
Kansas
Missouri
West Virginia
Virginia
Kentucky
California
Tennessee
North Carolina
South Carolina
Los Angeles
Arizona
New Mexico
Oklahoma
Arkansas
Mississippi
Alabama
Georgia
Phoenix
San Diego
Dallas
Texas
Louisiana
San Antonio
Houston
Florida
Gulf of California
Monterrey
Gulf of Mexico
Mexico
Havana
Cuba
Guadalajara
Mexico City
Dominican Republic
Port-au-Prince
Puerto Rico
Guatemala
Honduras
Caribbean
500 mi
500 km
Imagery ©2013 TerraMetrics, Map data ©2013 Google, INEGI -
WATER-POLO-....doc
Lecture_iCeNS....pptx
Show all downloads...

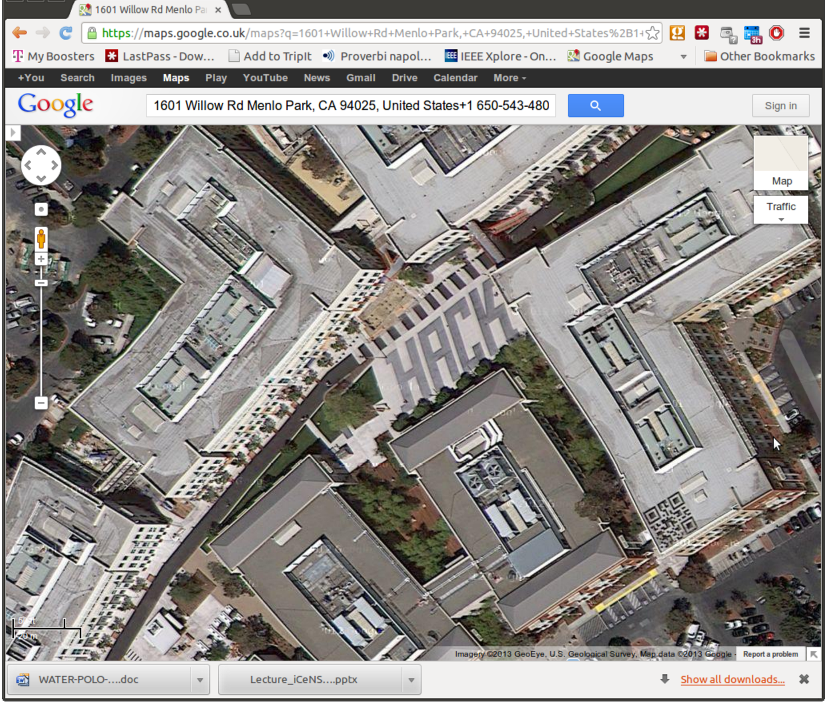

1601 Willow Rd Menlo Pa
https://maps.google.co.uk/maps?q=1601+Willow+Rd+Menlo+Park,+CA+94025,+United+States%2B1+
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+You
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
1601 Willow Rd Menlo Park, CA 94025, United States+1 650-543-480
Sign in
Map
Traffic
Medford
Klamath Falls
Ashland
Eureka
Redding
Elko
Nevada
Reno
Sacramento
Santa Rosa
Concord
Stockton
San Francisco
Fremont
San Jose
California
Fresno
St George
Las Vegas
Henderson
Bakersfield
Oxnard
Los Angeles
Long Beach
Anaheim
Huntington Beach
Escondido
San Diego
Tijuana
Mexicali
Rosarito
Ensenada
100 mi
200 km
Imagery ©2013 NASA, TerraMetrics, Map data ©2013 Google, INEGI
Report a problem
WATER-POLO-....doc
Lecture_iCeNS....pptx
Show all downloads...


1601 Willow Rd Menlo Pa
https://maps.google.co.uk/maps?q=1601+Willow+Rd+Menlo+Park,+CA+94025,+United+States%2B1+
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+You
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
1601 Willow Rd Menlo Park, CA 94025, United States+1 650-543-480
Sign in
Map
Traffic
Novato
San Pablo Bay
Vallejo
Point Reyes National Seashore
San Rafael
Pinole
Martinez
Concord
Antioch
Stockton
Lodi
Richmond
Mill Valley
Tiburon
Berkeley
Walnut Creek
Mt. Diablo State Park
Brentwood
French Camp
Danville
San Ramon
Manteca
San Francisco
Alameda
Tracy
Ripon
Salida
Gulf of the Farallones Marine Sanctuary
Daly City
Hayward
Livermore
Pacifica
Union City
Fremont
San Carlos
Palo Alto
Mountain View
Milpitas
San Jose
Patterson
Campbell
Los Gatos
Newman
San Luis National Wildlife Refuge
Boulder Creek
Morgan Hill
Henry W. Coe State Park
Scotts Valley
Santa Cruz
Gilroy
San Luis Reservoir State Recreation Area
Monterey Bay
Watsonville
Hollister
10 mi
20 km
Imagery ©2013 TerraMetrics, Map data ©2013 Google
Edit in Google Map Maker
Report a problem
WATER-POLO-....doc
Lecture_iCeNS....pptx
Show all downloads...


1601 Willow Rd Menlo Pa
https://maps.google.co.uk/maps?q=1601+Willow+Rd+Menlo+Park,+CA+94025,+United+States%2B1
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+You
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
1601 Willow Rd Menlo Park, CA 94025, United States+1 650-543-480
Sign in
Map
Traffic
East San Mateo
Lincoln
Foster City
Central Lake
Hillsdale
South San Mateo
Redwood Shores
Steinberger Slough
Belmont
San Carlos Airport-SQL
San Carlos
Westpoint Slough
Bayfront Park
Redwood City
Roosevelt
North Fair Oaks
Belle Haven
Bayfront Expy
Dumbarton Bridge
Ravenswood Open Space Preserve
Pulgas Ridge Open Space Preserve
Palm Park
Emerald Lake Hills
Farm Hills
Woodside Plaza
Atherton
West Atherton
Menlo Park
The Willows
East Palo Alto
Crescent Park
Palo Alto
Palo Alto Airport of Santa Clara County
Duveneck - St. Francis
Charleston Slough
Old Palo Alto
Midtown
Woodside
West Menlo Park
Stanford
Huddart Park
Teague Hill Open Space Preserve
Wunderlich County Park
Barron Park
Shoreline at Mountain View Park
Moffett Federal Airfield - NUQ
Jogel Slough
Guadalupe Slough
Don Edwards San Francisco Bay National Wildlife Refuge
Coyote Hills Regional Park
Ardenwood
Newark
Glenmoor
Brookvale
Cabrillo
Morse Park
Juniper Serra Fwy
El Camino Real
Woodside Rd
2 mi
2 km
Imagery ©2013 TerraMetrics, Map data ©2013 Google
Edit in Google Map Maker
Report a problem
WATER-POLO-....doc
Lecture_iCeNS....pptx
Show all downloads...


1601 Willow Rd Menlo Pa
https://maps.google.co.uk/maps?q=1601+Willow+Rd+Menlo+Park,+CA+94025,+United+States%2B1+
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+You
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
1601 Willow Rd Menlo Park, CA 94025, United States+1 650-543-480
Sign in
Map
Traffic
Bayfront Expy
Hacker Way
Willow Rd
University Ave
Constitution Dr
Chilco St
Terminal Ave
Hamilton Ave
Hamilton Ct
Adams Ct
O'Brien Dr
Belle Haven
Ivy Dr
Newbridge St
Kavanaugh Dr
Notre Dame Ave
Tulane Ave
Purdue Ave
Jack Farrell Park
East Palo Alto Library
Bay Rd
Ravenswood Open Space Preserve
1000 ft
200 m
Imagery ©2013 DigitalGlobe, GeoEye, U.S. Geological Survey, USDA Farm Service Agency, Map data ©2013 Google
Edit in Google Map Maker
Report a problem
WATER-POLO-....doc
Lecture_iCeNS....pptx
Show all downloads...


1601 Willow Rd Menlo Pa
https://maps.google.co.uk/maps?q=1601+Willow+Rd+Menlo+Park,+CA+94025,+United+States%2B1
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+You
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
1601 Willow Rd Menlo Park, CA 94025, United States+1 650-543-480
Sign in
Map
Traffic
Hacker Way
Facebook Research Center
Building 16
Bayfront Expy
84
114
Willow Rd
200 ft
100 m
Imagery ©2013 DigitalGlobe, GeoEye, U.S. Geological Survey, USDA Farm Service Agency, Map data ©2013 Google -
Edit in Google Map Maker
Report a problem
WATER-POLO-....doc
Lecture_iCeNS....pptx
Show all downloads...

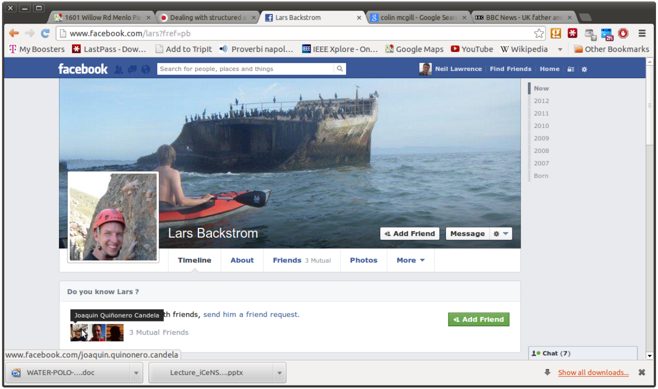

1601 Willow Rd Menlo Pa
https://maps.google.co.uk/maps?q=1601+Willow+Rd+Menlo+Park,+CA+94025,+United+States%2B1+
My Boosters
LastPass - Dow...
Add to TripIt
Proverbi napol...
IEEE Xplore - On...
Google Maps
Other Bookmarks
+You
Search
Images
Maps
Play
YouTube
News
Gmail
Drive
Calendar
More
Google
1601 Willow Rd Menlo Park, CA 94025, United States+1 650-543-480
Sign in
Map
Traffic
HACK
Imagery ©2013 GeoEye, U.S. Geological Survey, Map data ©2013 Google -
Report a problem
WATER-POLO-....doc
Lecture_iCeNS....pptx
Show all downloads...


research@fb


The HACKER
research@fb


Change Cover
Joaquin Quiñonero Candela
Update Info
Activity Log 6
Works at Facebook
Studied at Technical University of Denmark
Lives in Palo Alto, California
Married to Ines Koch
About
Friends 523
Photos 73
Map 133
Likes 79
6


Average user
130
friends
Total
100B
social connections



http://videolectures.net/eswc2011_backstrom_facebook/

Classification

- We are given data set containing "inputs", $\mathbf{X}$, and "targets", $\mathbf{y}$.
- Each data point consists of an input vector $\mathbf{x}_{i,:}$ and a class label, y_i.
- For binary classification assume y_i should be either 1 (yes) or -1 (no).
- Input vector can be thought of as features.

Classification Examples

- Classifying hand written digits from binary images (automatic zip code reading).
- Detecting faces in images (e.g. digital cameras).
- Who a detected face belongs to (e.g. Picasa).
- Classifying type of cancer given gene expression data.
- Categorization of document types (different types of news article on the internet).

The Perceptron

- Developed in 1957 by Rosenblatt.
- Take a data point at, $\mathbf{x}_i$.
- Predict it belongs to a class, $y_i = 1$ if $\sum_j w_j \mathbf{x}_{i,j} + b > 0$ i.e. $\mathbf{w}^\top \mathbf{x}_i + b > 0$. Otherwise assume $y_i = -1$.

Perceptron-like Algorithm

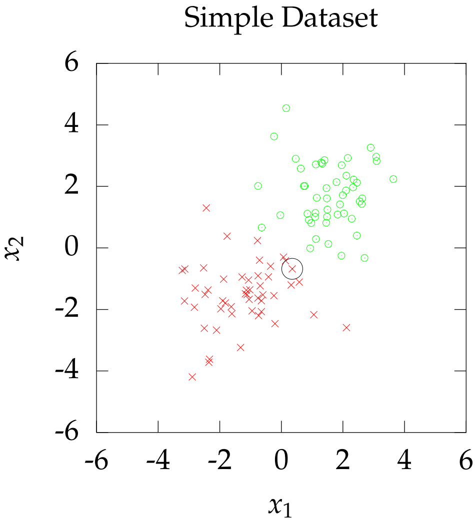
1. Select a random data point i.
2. Ensure i is correctly classified by setting $\mathbf{w} = y_i \mathbf{x}_i$.
   - i.e. $\text{sign}\left(\mathbf{w}^\top \mathbf{x}_{i,:}\right) = \text{sign}\left(y_i \mathbf{x}_{i,:}^\top \mathbf{x}_{i,:}\right) = \text{sign}\left(y_i\right) = y_i$

Perceptron Iteration

1. Select a misclassified point, i.
2. Set $\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_{i,:}$.
    - If η is large enough this will guarantee this point becomes correctly classified.
3. Repeat until there are no misclassified points.

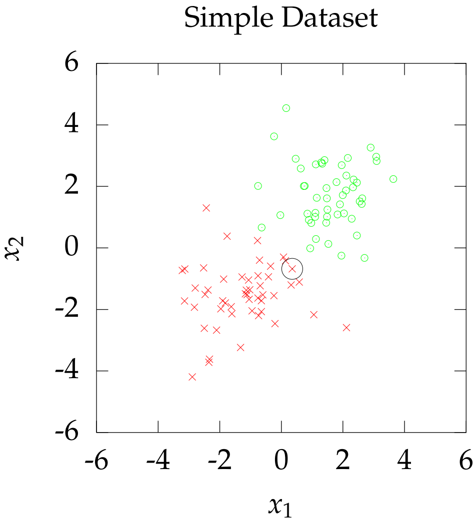
Perceptron Algorithm

- Iteration 1 data no 29



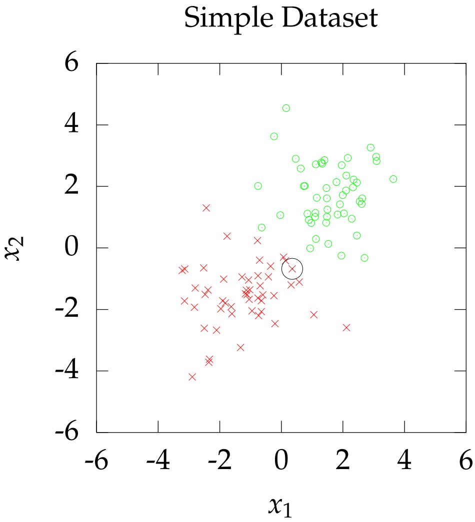
Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$



Perceptron Algorithm

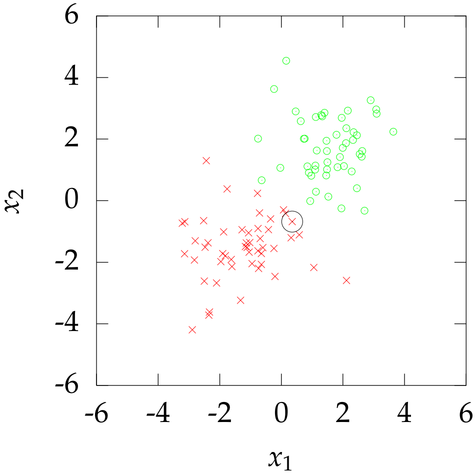
- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration



Perceptron Algorithm

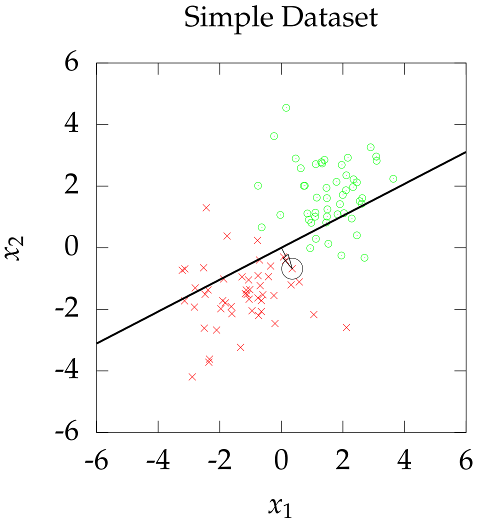
- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration
- Set weight vector to data point.

Simple Dataset



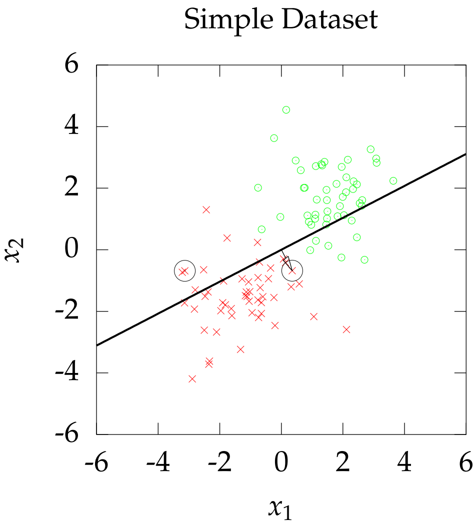
Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration
- Set weight vector to data point.
- $\mathbf{w} = y_{29}\mathbf{x}_{29,:}$



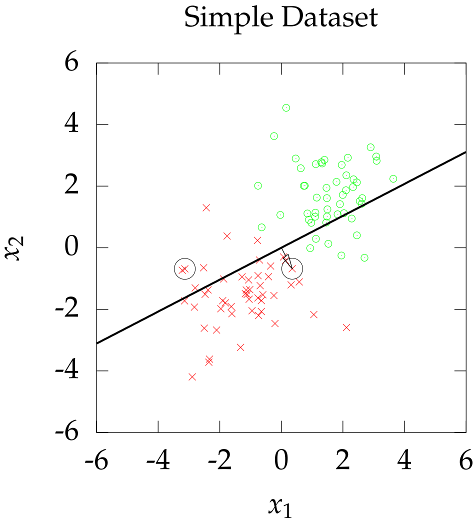
Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration
- Set weight vector to data point.
- $\mathbf{w} = y_{29}\mathbf{x}_{29,:}$
- Select new incorrectly classified data point.



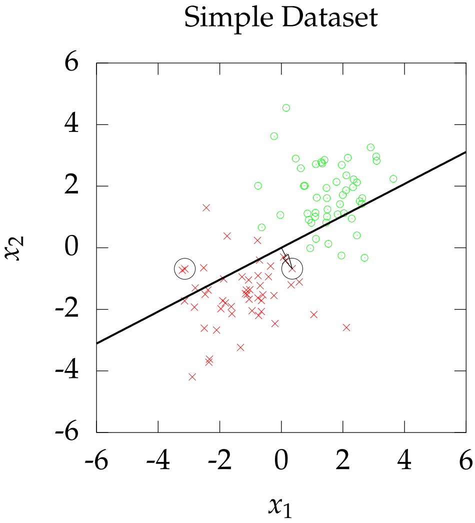
Perceptron Algorithm

- Iteration 2 data no 16



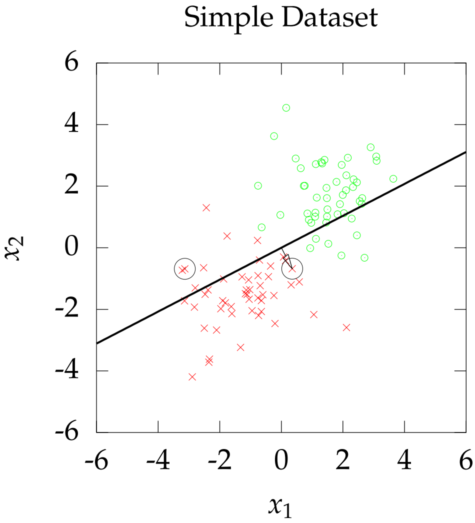
Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$



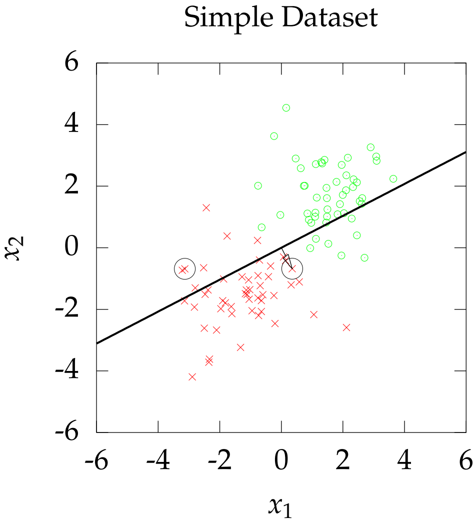
Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$
- Incorrect classification



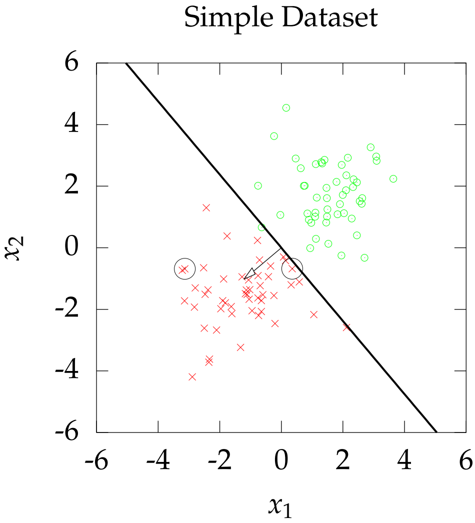
Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$
- Incorrect classification
- Adjust weight vector with new data point.



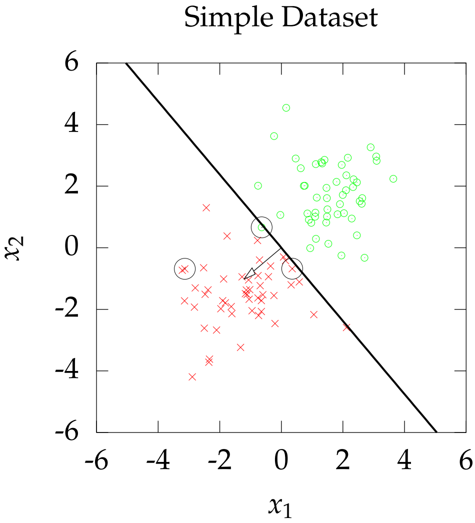
Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{16} \mathbf{x}_{16,:}$



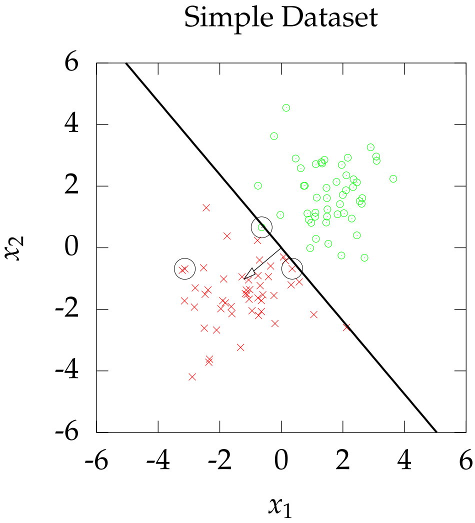
Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{16} \mathbf{x}_{16,:}$
- Select new incorrectly classified data point.



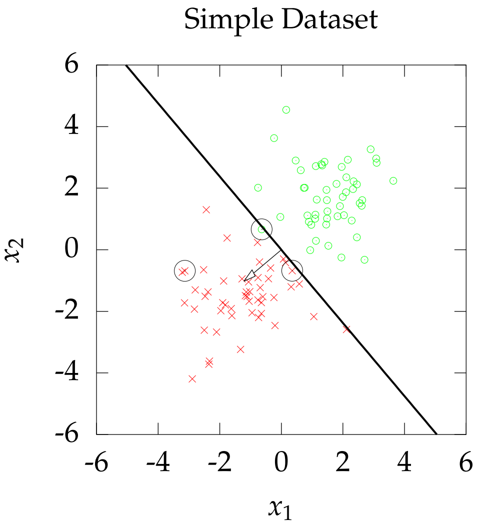
Perceptron Algorithm

- Iteration 3 data no 58



Perceptron Algorithm

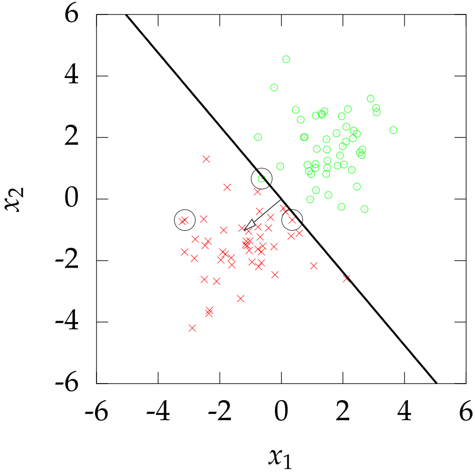
- Iteration 3 data no 58
- $w_1 = -1.2143$,
  $w_2 = -1.0217$



Perceptron Algorithm

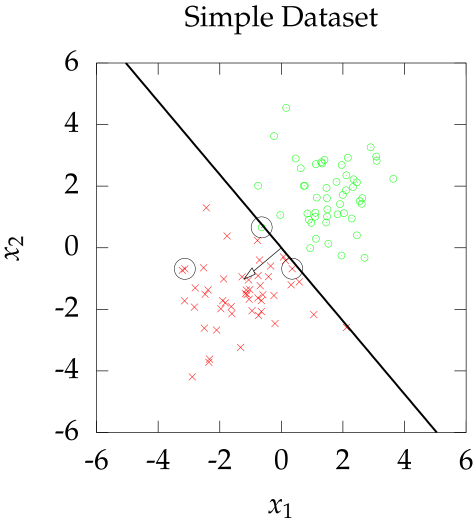
- Iteration 3 data no 58
- $w_1 = -1.2143$,
  $w_2 = -1.0217$
- Incorrect classification

Simple Dataset



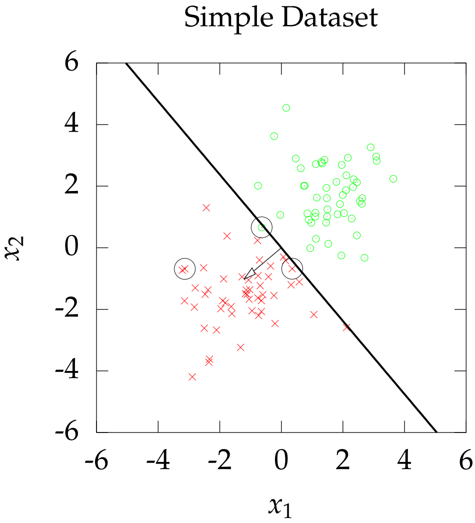
Perceptron Algorithm

- Iteration 3 data no 58
- $w_1 = -1.2143$,
  $w_2 = -1.0217$
- Incorrect classification
- Adjust weight vector with new data point.



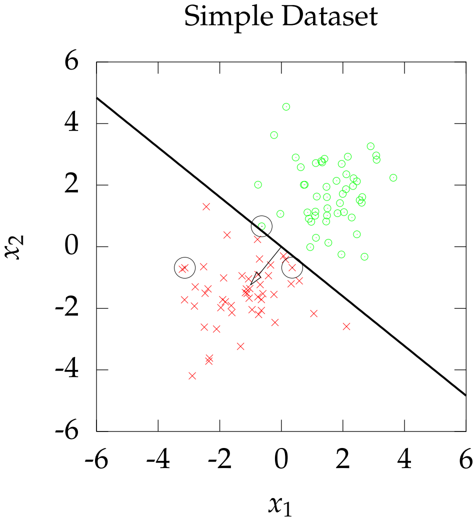
Perceptron Algorithm

- Iteration 3 data no 58
- $w_1 = -1.2143$,
  $w_2 = -1.0217$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{58} \mathbf{x}_{58,:}$



Perceptron Algorithm

- Iteration 3 data no 58
- $w_1 = -1.2143$,
  $w_2 = -1.0217$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{58} \mathbf{x}_{58,:}$
- All data correctly classified.



Regression Examples

- Predict a real value, y_i given some inputs $\mathbf{x}_i$.
- Predict quality of meat given spectral measurements (Tecator data).
- Radiocarbon dating, the C14 calibration curve: predict age given quantity of C14 isotope.
- Predict quality of different Go or Backgammon moves given expert rated training data.

Linear Regression

Is there an equivalent learning rule for regression?

- Predict a real value y given x.
- We can also construct a learning rule for regression.
  - Define our prediction
  $$f(x) = mx + c.$$
  - Define an error
  $$\Delta y_i = y_i - f(x_i).$$

Updating Bias/Intercept

- c represents bias. Add portion of error to bias.

$$c \rightarrow c + \eta \Delta y_i.$$

$$\Delta y_i = y_i - m x_i - c.$$

1. For +ve error, c and therefore $f(x_i)$ become larger and error magnitude becomes smaller.
2. For -ve error, c and therefore $f(x_i)$ become smaller and error magnitude becomes smaller.

Updating Slope

- m represents Slope. Add portion of error $\times$ input to slope.

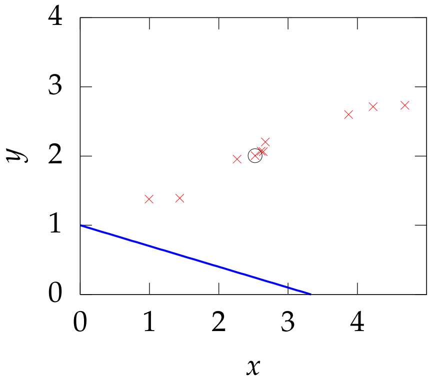
$$m \to m + \eta \Delta y_i x_i.$$

$$\Delta y_i = y_i - m x_i - c.$$

1. For +ve error and +ve input, m becomes larger and $f(x_i)$ becomes larger: error magnitude becomes smaller.
2. For +ve error and -ve input, m becomes smaller and $f(x_i)$ becomes larger: error magnitude becomes smaller.
3. For -ve error and -ve slope, m becomes larger and $f(x_i)$ becomes smaller: error magnitude becomes smaller.
4. For -ve error and +ve input, m becomes smaller and $f(x_i)$ becomes smaller: error magnitude becomes smaller.

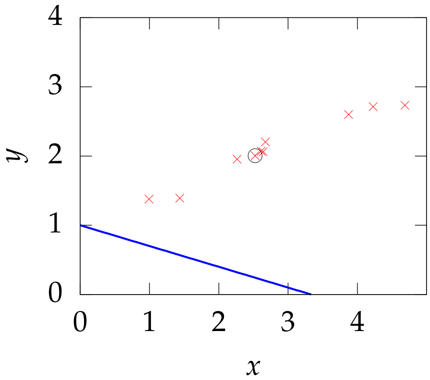
Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$ $\hat{c} = 1$



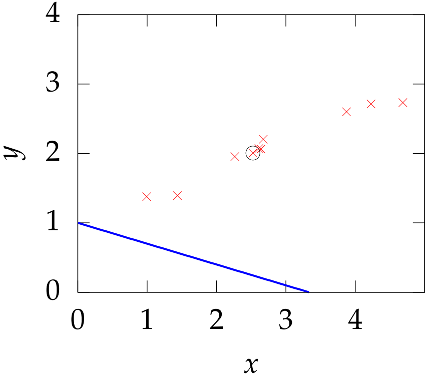
Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$ $\hat{c} = 1$
  - Present data point 4



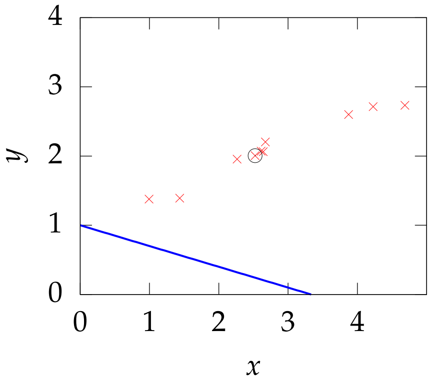
Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$ $\hat{c} = 1$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$



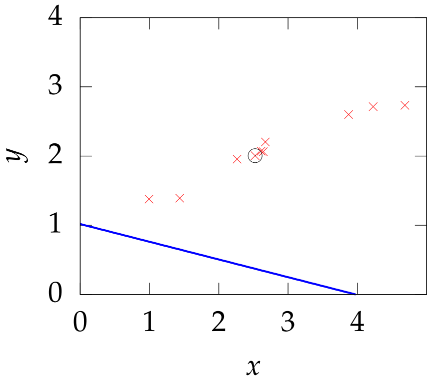
Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$ $\hat{c} = 1$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m} x_4 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_4$



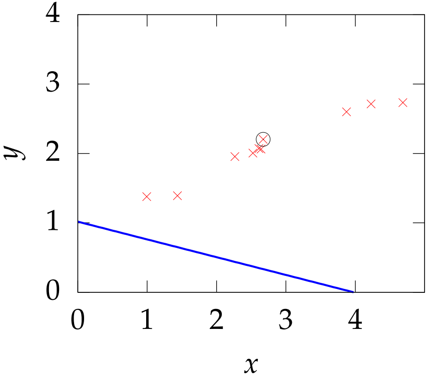
Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$ $\hat{c} = 1$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_4$
- Updated values
  $\hat{m} = -0.25593$ $\hat{c} = 1.0175$



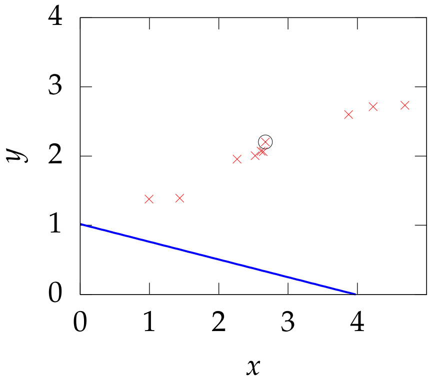
Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$



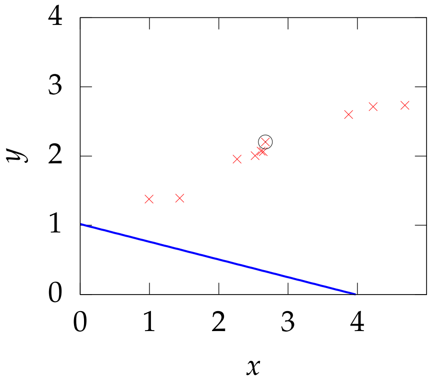
Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$
  - Present data point 7



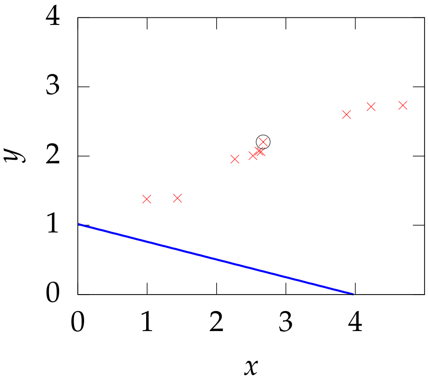
Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m} x_7 - \hat{c})$



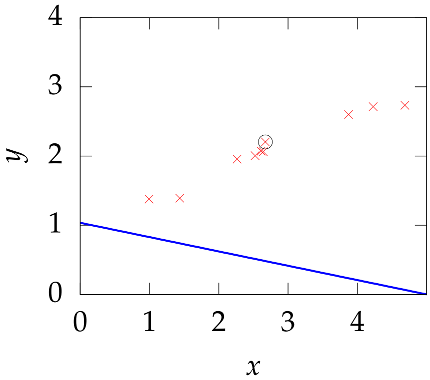
Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$



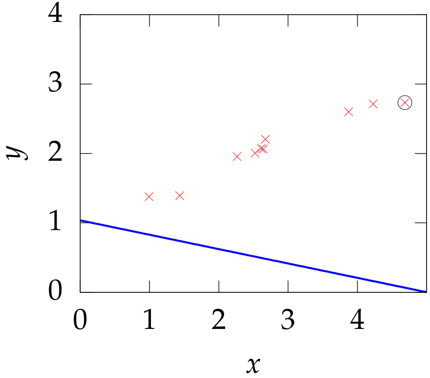
Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$
- Updated values
  $\hat{m} = -0.20693$ $\hat{c} = 1.0358$



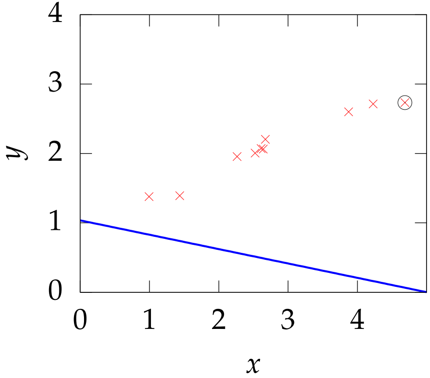
Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$



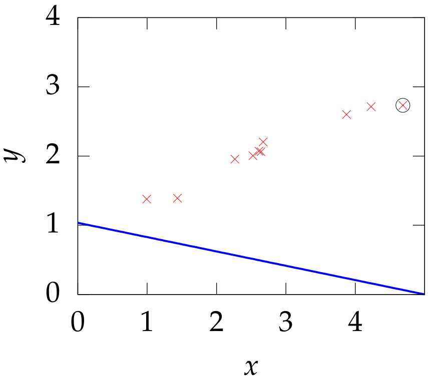
Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$
  - Present data point 10



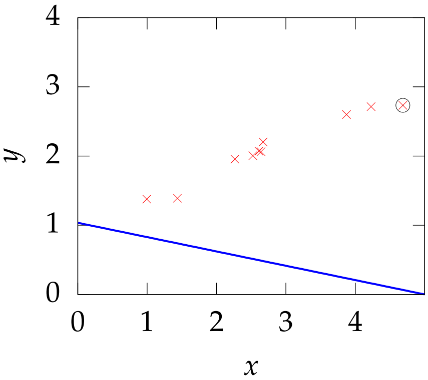
Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$



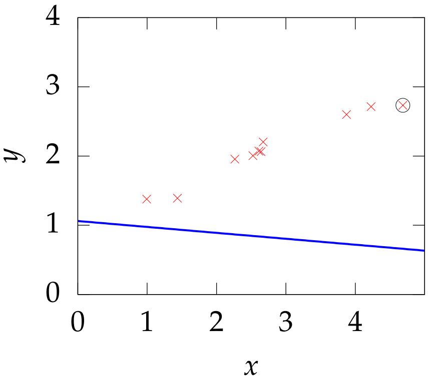
Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



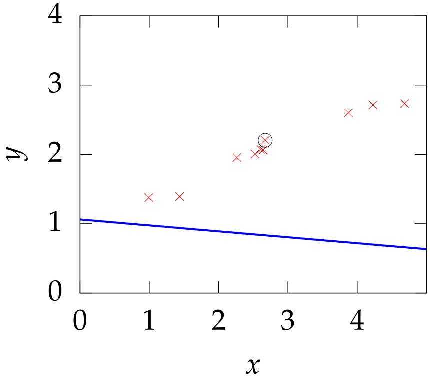
Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = -0.085591$ $\hat{c} = 1.0617$



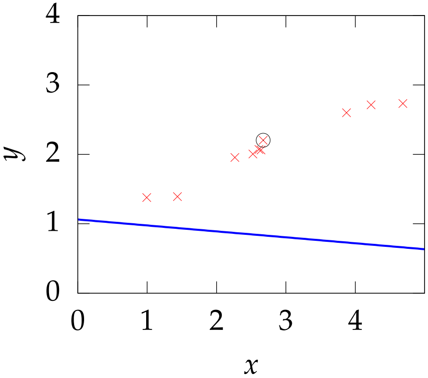
Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$



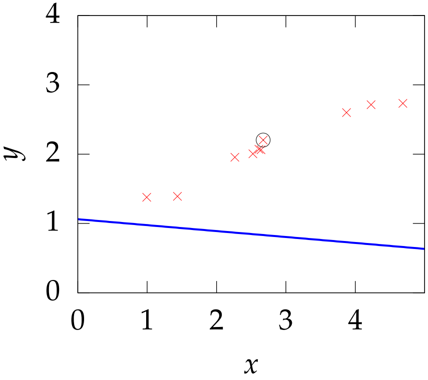
Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$
  - Present data point 7



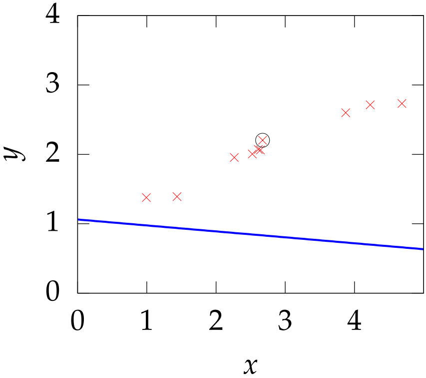
Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$



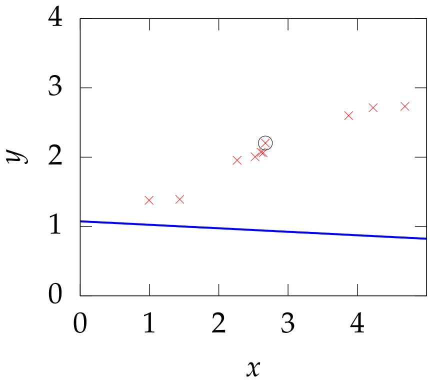
Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$



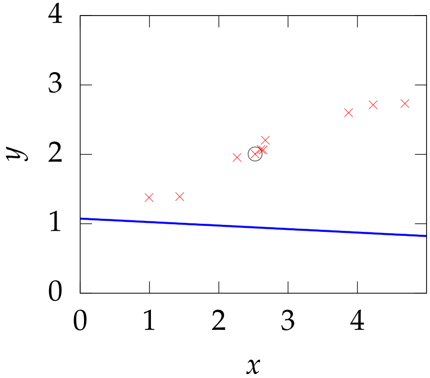
Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$
- Updated values
  $\hat{m} = -0.050355$ $\hat{c} = 1.0749$



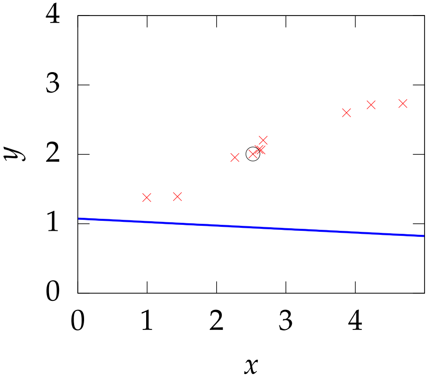
Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$



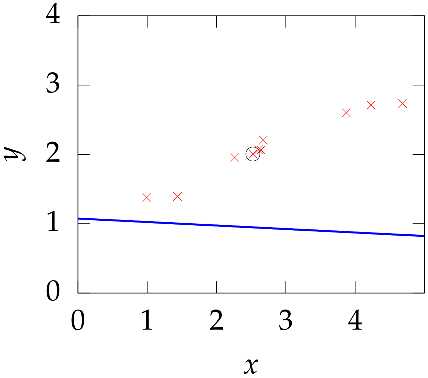
Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$
  - Present data point 4



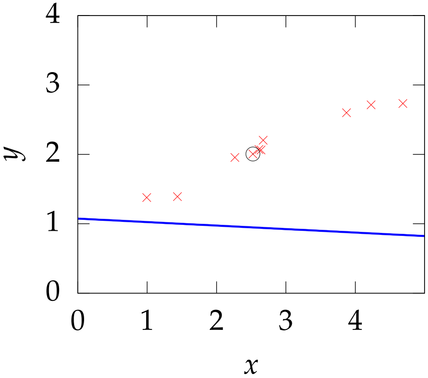
Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$



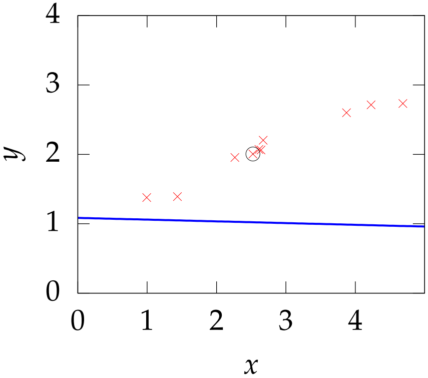
Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_4$



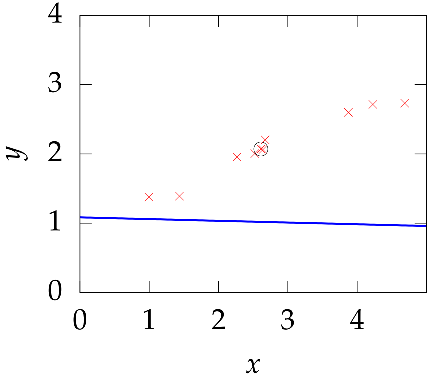
Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_4$
- Updated values
  $\hat{m} = -0.024925$ $\hat{c} = 1.0849$



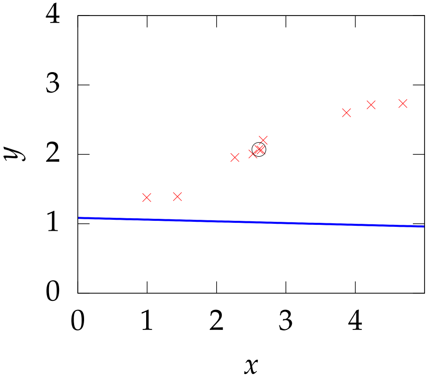
Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$



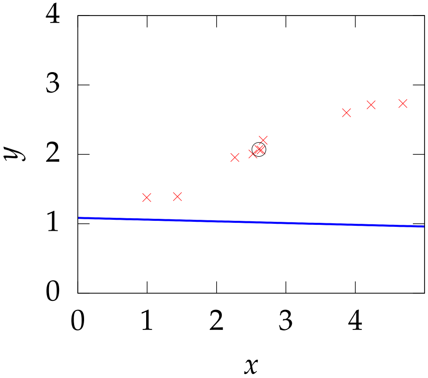
Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$
  - Present data point 5



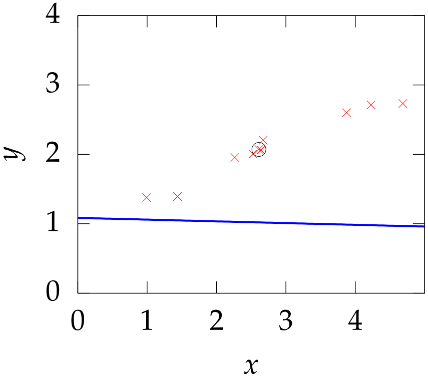
Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$
  - Present data point 5
  - $\Delta y_5 = (y_5 - \hat{m}x_5 - \hat{c})$



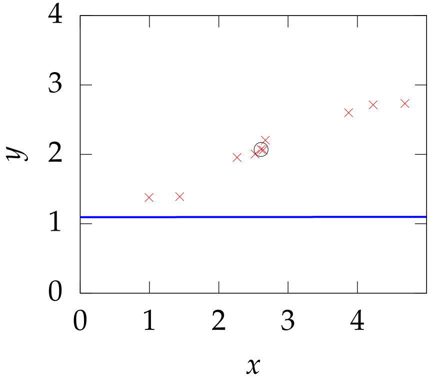
Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$
  - Present data point 5
  - $\Delta y_5 = (y_5 - \hat{m}x_5 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_5 \Delta y_5$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_5$



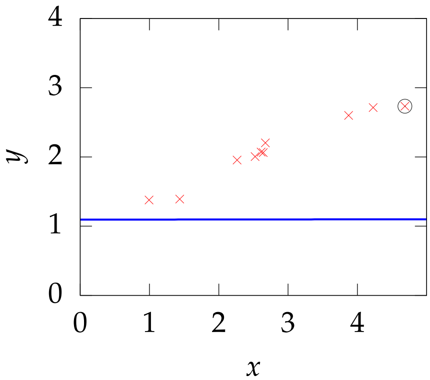
Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$
  - Present data point 5
  - $\Delta y_5 = (y_5 - \hat{m}x_5 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_5 \Delta y_5$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_5$
- Updated values
  $\hat{m} = 0.00098511$ $\hat{c} = 1.0949$



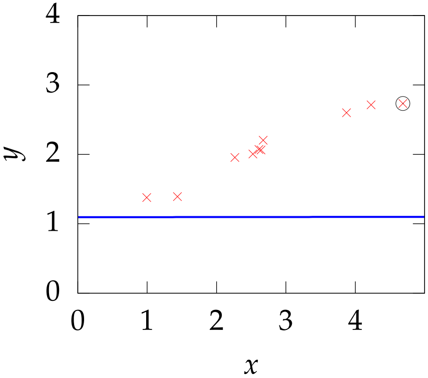
Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$



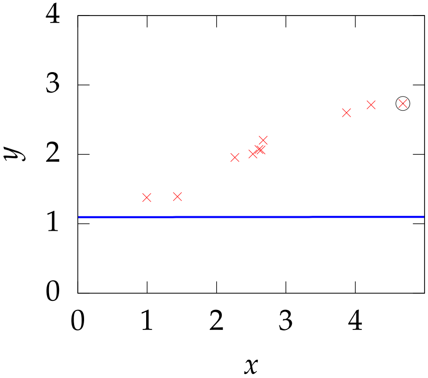
Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$
  - Present data point 10



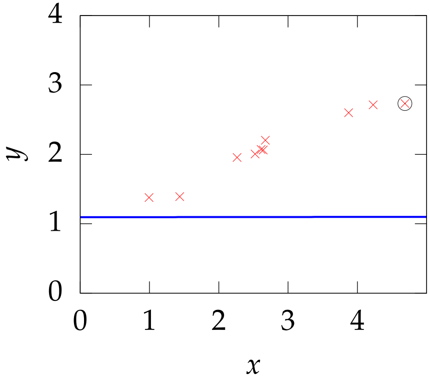
Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$



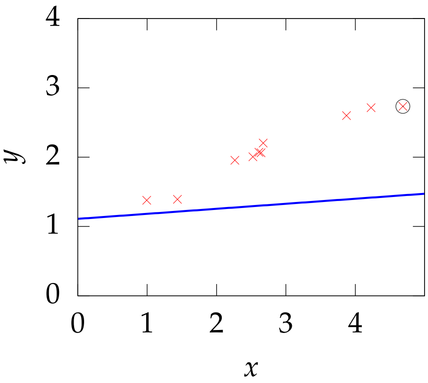
Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m} x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



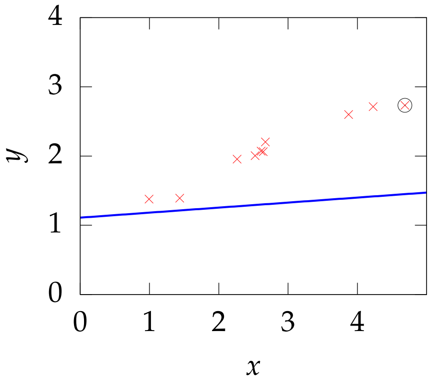
Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.072529$ $\hat{c} = 1.1101$



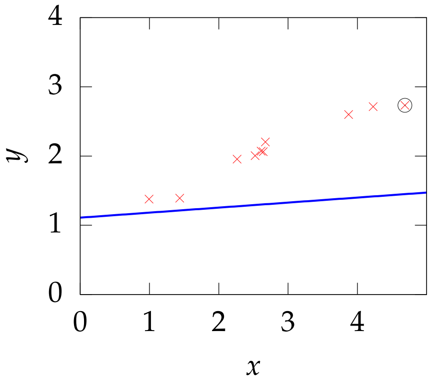
Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$



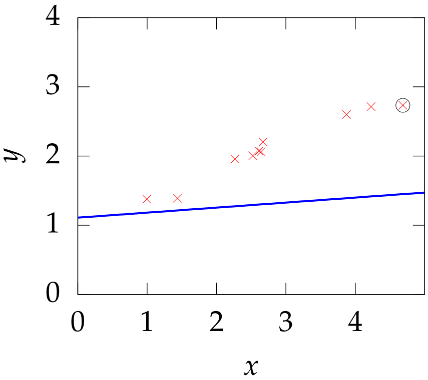
Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$
  - Present data point 10



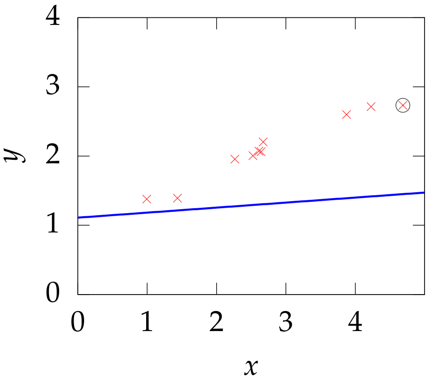
Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m} x_{10} - \hat{c})$



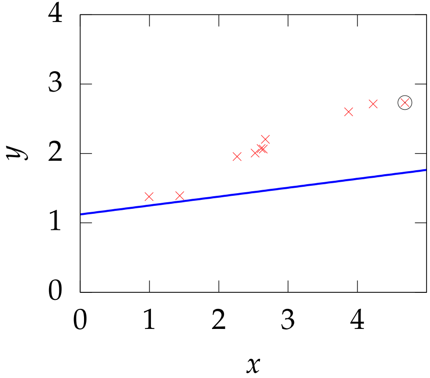
Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



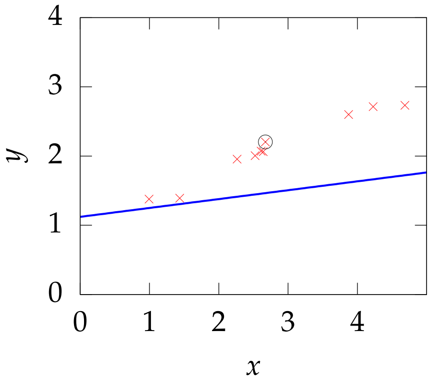
Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.1282$ $\hat{c} = 1.122$



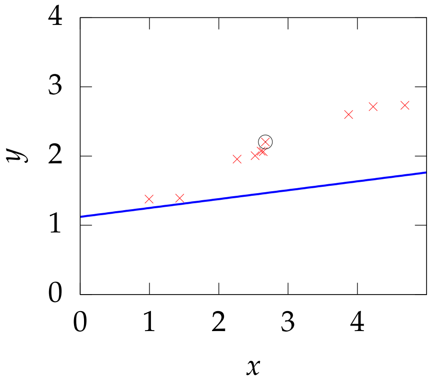
Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$



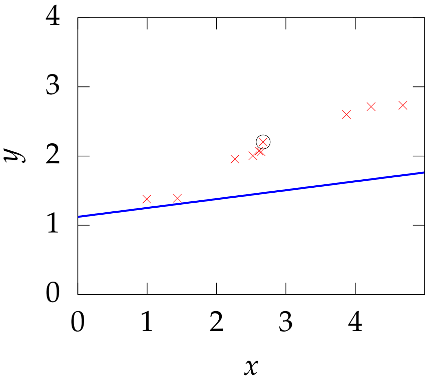
Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$
  - Present data point 7



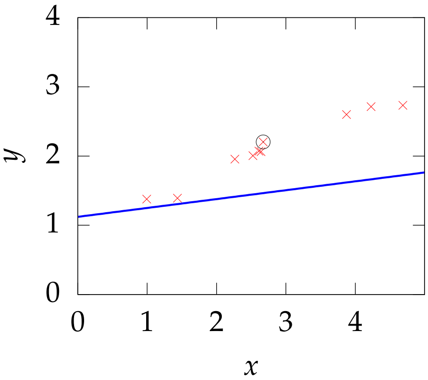
Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m} x_7 - \hat{c})$



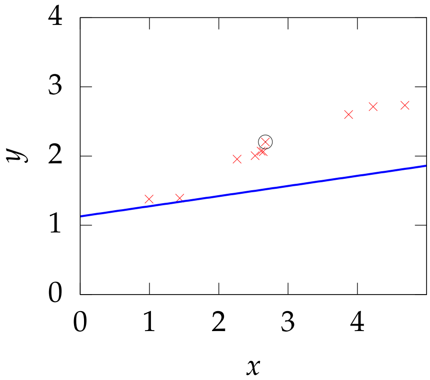
Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$



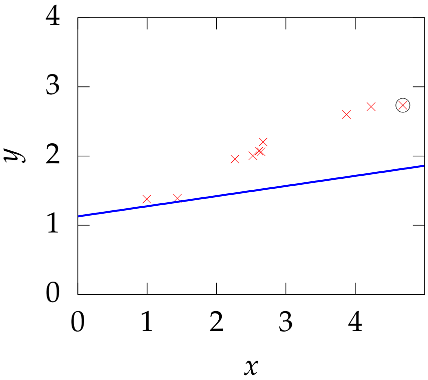
Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$
- Updated values
  $\hat{m} = 0.14634$ $\hat{c} = 1.1288$



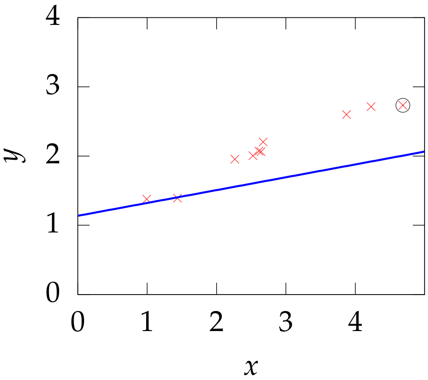
Linear Regression Example

- Iteration 10 $\hat{m} = 0.14634$ $\hat{c} = 1.1288$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



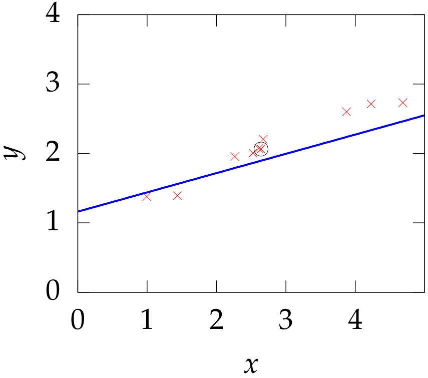
Linear Regression Example

- Iteration 10 $\hat{m} = 0.14634$ $\hat{c} = 1.1288$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.18547$ $\hat{c} = 1.1372$



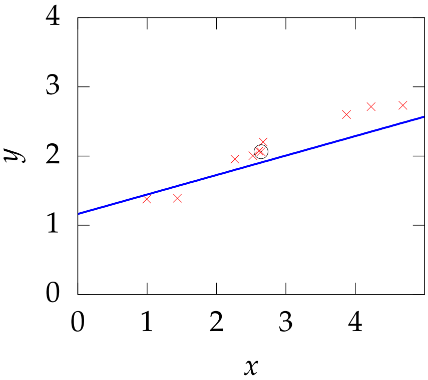
Linear Regression Example

- Iteration 20 $\hat{m} = 0.27764$ $\hat{c} = 1.1621$
  - Present data point 6
  - $\Delta y_6 = (y_6 - \hat{m}x_6 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_6 \Delta y_6$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_6$



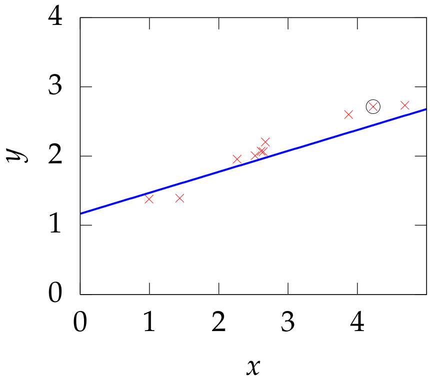
Linear Regression Example

- Iteration 20 $\hat{m} = 0.27764$ $\hat{c} = 1.1621$
  - Present data point 6
  - $\Delta y_6 = (y_6 - \hat{m}x_6 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_6 \Delta y_6$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_6$
- Updated values
  $\hat{m} = 0.28135$ $\hat{c} = 1.1635$



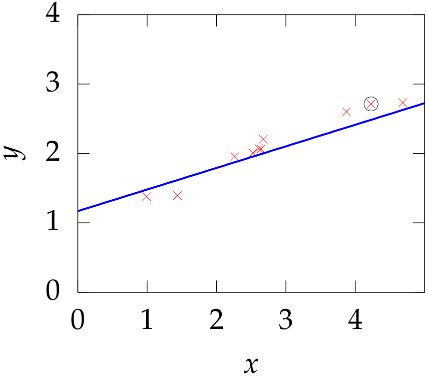
Linear Regression Example

- Iteration 30 $\hat{m} = 0.30249$ $\hat{c} = 1.1673$
  - Present data point 9
  - $\Delta y_9 = (y_9 - \hat{m} x_9 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_9 \Delta y_9$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_9$



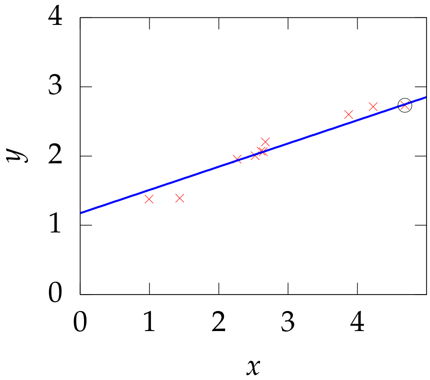
Linear Regression Example

- Iteration 30 $\hat{m} = 0.30249$ $\hat{c} = 1.1673$
  - Present data point 9
  - $\Delta y_9 = (y_9 - \hat{m}x_9 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_9 \Delta y_9$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_9$
- Updated values
  $\hat{m} = 0.31119$ $\hat{c} = 1.1693$



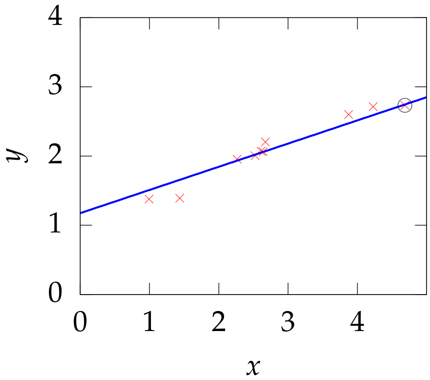
Linear Regression Example

- Iteration 40 $\hat{m} = 0.33551$ $\hat{c} = 1.1754$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10}\Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



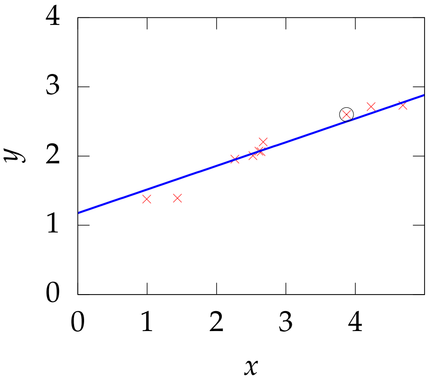
Linear Regression Example

- Iteration 40 $\hat{m} = 0.33551$ $\hat{c} = 1.1754$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.33503$ $\hat{c} = 1.1753$



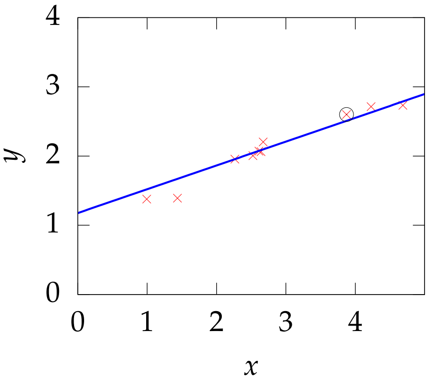
Linear Regression Example

- Iteration 50 $\hat{m} = 0.34126$ $\hat{c} = 1.1763$
  - Present data point 8
  - $\Delta y_8 = (y_8 - \hat{m}x_8 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_8 \Delta y_8$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_8$



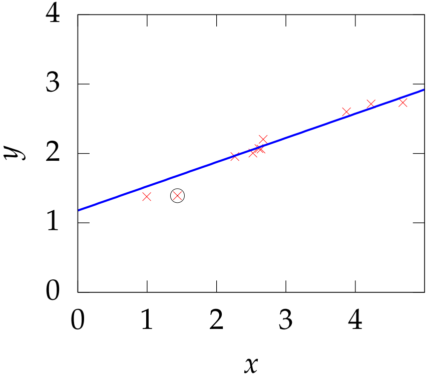
Linear Regression Example

- Iteration 50 $\hat{m} = 0.34126$ $\hat{c} = 1.1763$
  - Present data point 8
  - $\Delta y_8 = (y_8 - \hat{m}x_8 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_8 \Delta y_8$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_8$
- Updated values
  $\hat{m} = 0.3439$ $\hat{c} = 1.177$



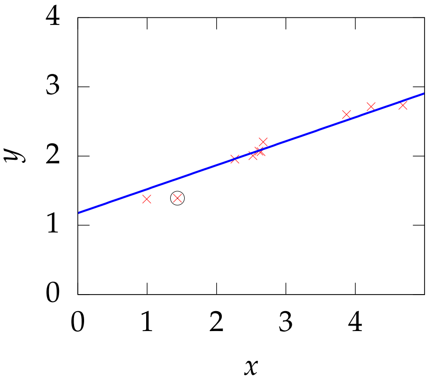
Linear Regression Example

- Iteration 60 $\hat{m} = 0.34877$ $\hat{c} = 1.1775$
  - Present data point 2
  - $\Delta y_2 = (y_2 - \hat{m} x_2 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_2 \Delta y_2$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_2$



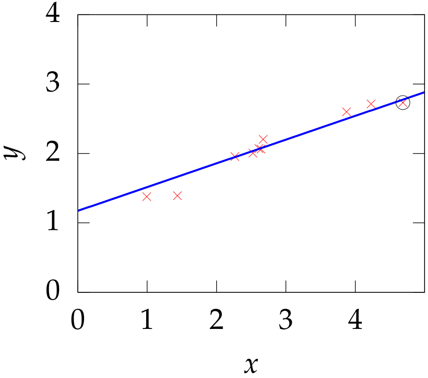
Linear Regression Example

- Iteration 60 $\hat{m} = 0.34877$ $\hat{c} = 1.1775$
  - Present data point 2
  - $\Delta y_2 = (y_2 - \hat{m} x_2 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_2 \Delta y_2$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_2$
- Updated values
  $\hat{m} = 0.34621$ $\hat{c} = 1.1757$



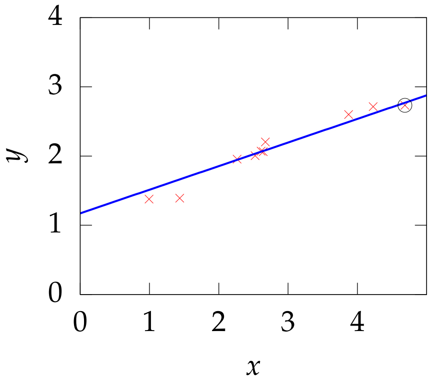
Linear Regression Example

- Iteration 70 $\hat{m} = 0.34207$ $\hat{c} = 1.1734$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



Linear Regression Example

- Iteration 70 $\hat{m} = 0.34207$ $\hat{c} = 1.1734$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.34088$ $\hat{c} = 1.1732$



Basis Functions

Nonlinear Regression

- Problem with Linear Regression—$\mathbf{x}$ may not be linearly related to $\mathbf{y}$.
- Potential solution: create a feature space: define $\phi(\mathbf{x})$ where $\phi(\cdot)$ is a nonlinear function of $\mathbf{x}$.
- Model for target is a linear combination of these nonlinear functions

$$f(\mathbf{x}) = \sum_{j=1}^{K} w_j \phi_j(\mathbf{x}) \tag{1}$$

Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

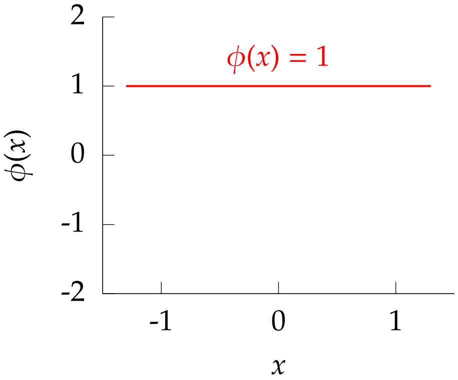


Figure: A quadratic basis.

Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

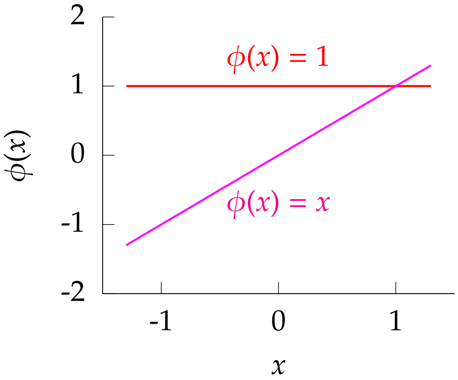


Figure: A quadratic basis.

Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

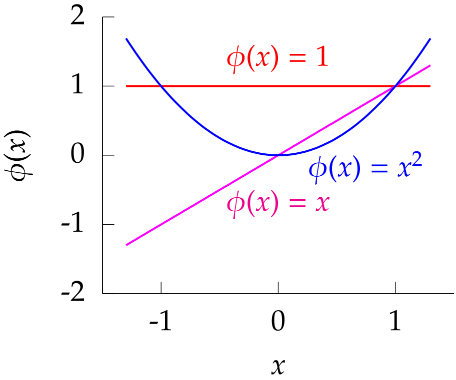


Figure: A quadratic basis.

Functions Derived from Quadratic Basis

$$f(x) = w_1 + w_2 x + w_3 x^2$$

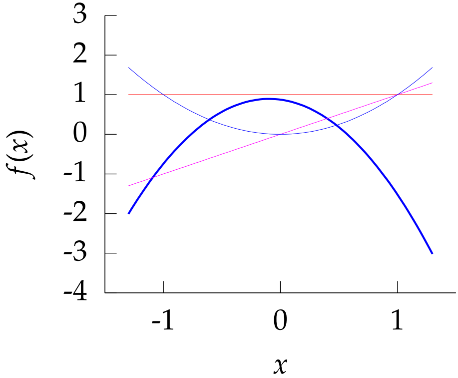


Figure: Function from quadratic basis with weights $w_1 = 0.87466$, $w_2 = -0.38835$, $w_3 = -2.0058$.

Functions Derived from Quadratic Basis

$$f(x) = w_1 + w_2 x + w_3 x^2$$

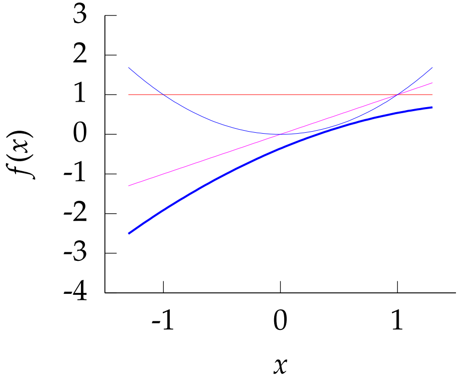


Figure: Function from quadratic basis with weights $w_1 = -0.35908$, $w_2 = 1.2274$, $w_3 = -0.32825$.

Functions Derived from Quadratic Basis

$$f(x) = w_1 + w_2 x + w_3 x^2$$

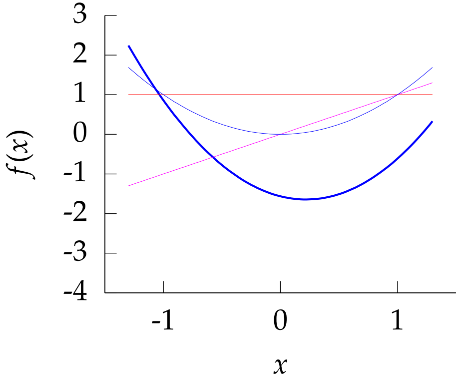


Figure: Function from quadratic basis with weights $w_1 = -1.5638$, $w_2 = -0.73577$, $w_3 = 1.6861$.

Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

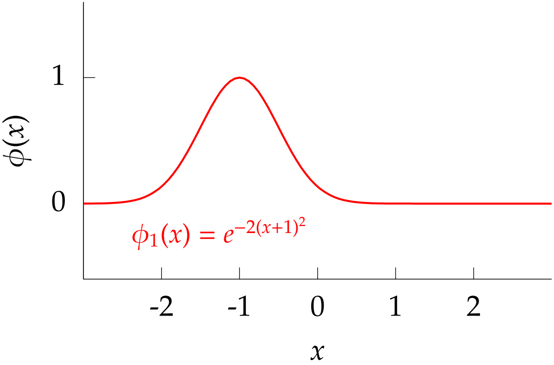


Figure: Radial basis functions.

Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

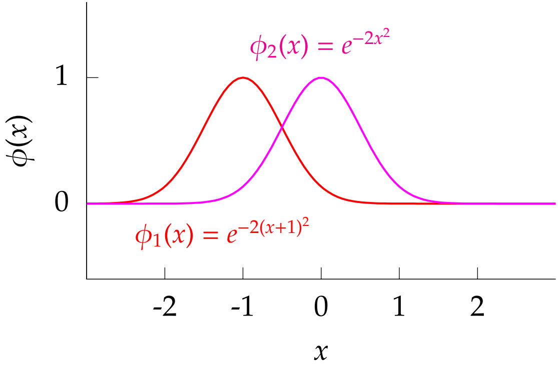


Figure: Radial basis functions.

Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

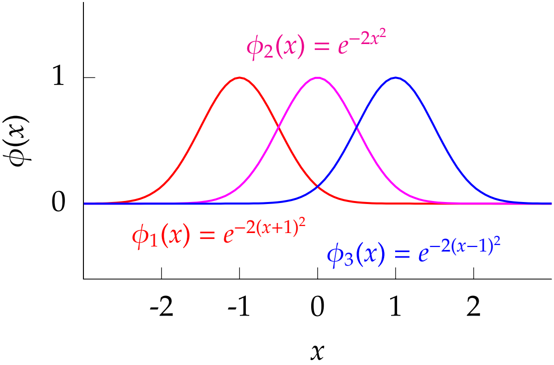


Figure: Radial basis functions.

Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

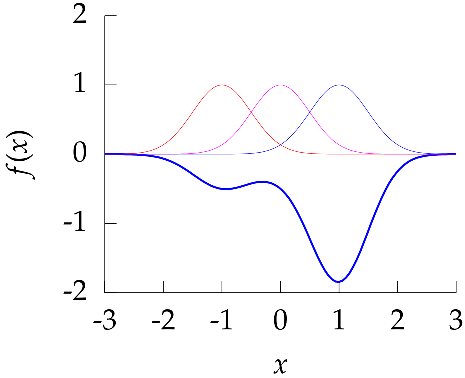


Figure: Function from radial basis with weights $w_1 = -0.47518$, $w_2 = -0.18924$, $w_3 = -1.8183$.

Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

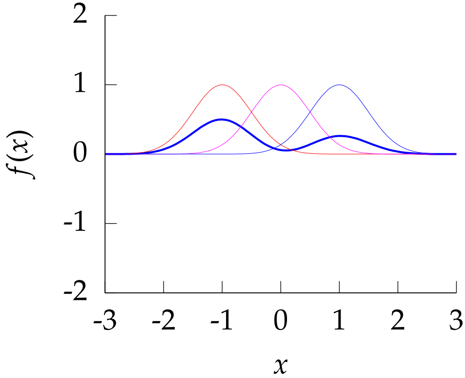


Figure: Function from radial basis with weights $w_1 = 0.50596$, $w_2 = -0.046315$, $w_3 = 0.26813$.

Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

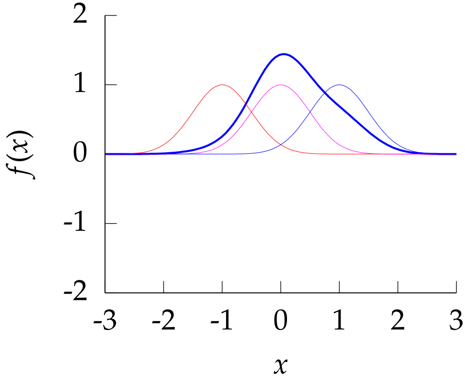
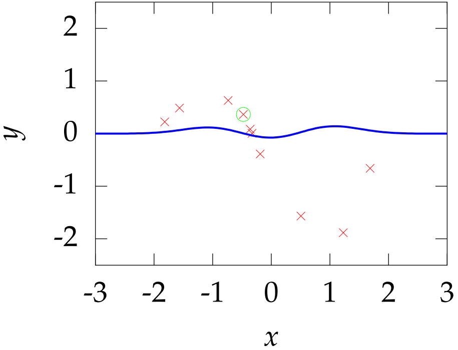


Figure: Function from radial basis with weights $w_1 = 0.07179$, $w_2 = 1.3591$, $w_3 = 0.50604$.

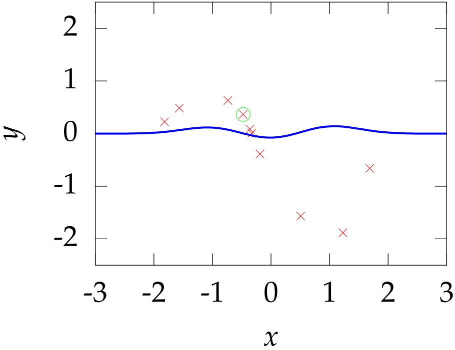
Nonlinear Regression Example

- Iteration 1
  - $w_1 = 0.13018,$
    $w_2 = -0.11355,$
    $w_3 = 0.15448$
  - Present data point 4



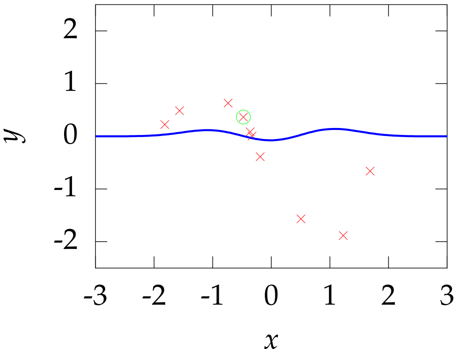
Nonlinear Regression Example

- Iteration 1
  - $w_1 = 0.13018$,
    $w_2 = -0.11355$,
    $w_3 = 0.15448$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$



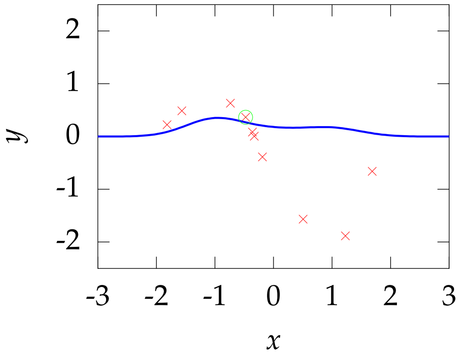
Nonlinear Regression Example

- Iteration 1
  - $w_1 = 0.13018,$
    $w_2 = -0.11355,$
    $w_3 = 0.15448$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



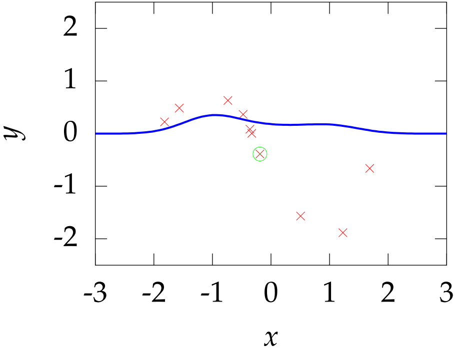
Nonlinear Regression Example

- Iteration 1
  - $w_1 = 0.13018,$
    $w_2 = -0.11355,$
    $w_3 = 0.15448$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



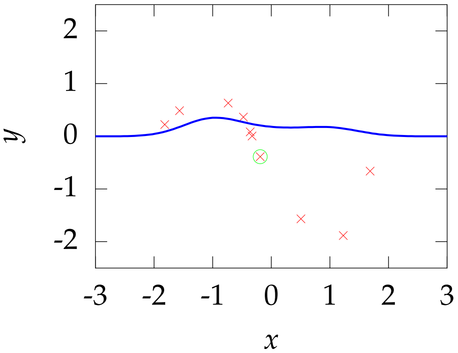
Nonlinear Regression Example

- Iteration 2
  - $w_1 = 0.33696$,
    $w_2 = 0.11481$,
    $w_3 = 0.1591$
  - Present data point 7



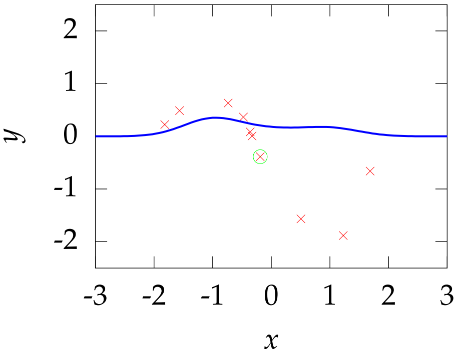
Nonlinear Regression Example

- Iteration 2
  - $w_1 = 0.33696$,
    $w_2 = 0.11481$,
    $w_3 = 0.1591$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$



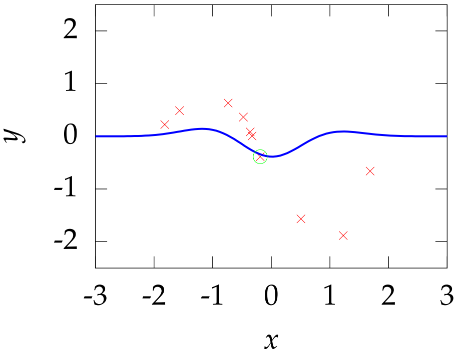
Nonlinear Regression Example

- Iteration 2
  - $w_1 = 0.33696,$
    $w_2 = 0.11481,$
    $w_3 = 0.1591$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



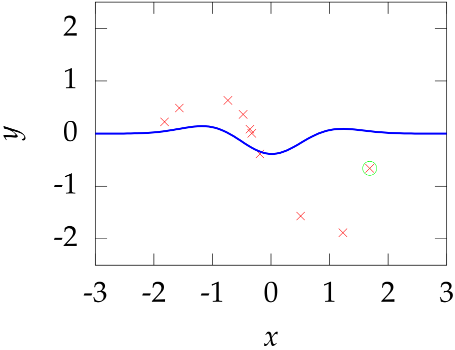
Nonlinear Regression Example

- Iteration 2
  - $w_1 = 0.33696,$
    $w_2 = 0.11481,$
    $w_3 = 0.1591$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



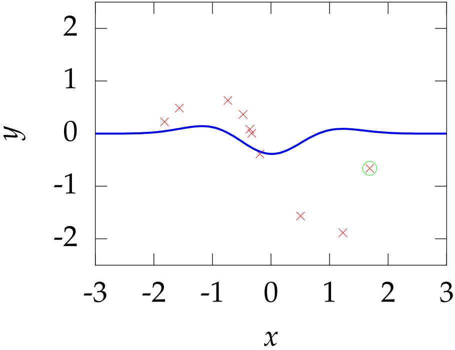
Nonlinear Regression Example

- Iteration 3
  - $w_1 = 0.18076$,
    $w_2 = -0.4266$,
    $w_3 = 0.12473$
  - Present data point 10



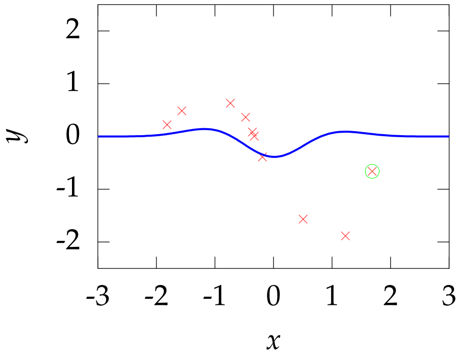
Nonlinear Regression Example

- Iteration 3
  - $w_1 = 0.18076,$
    $w_2 = -0.4266,$
    $w_3 = 0.12473$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$



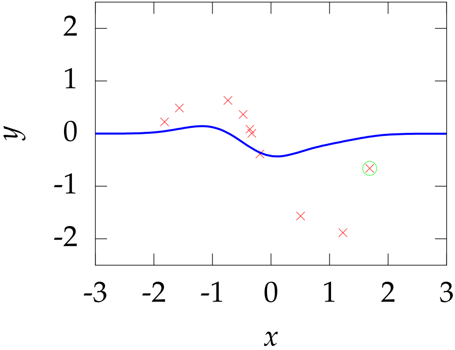
Nonlinear Regression Example

- Iteration 3
  - $w_1 = 0.18076$,
    $w_2 = -0.4266$,
    $w_3 = 0.12473$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



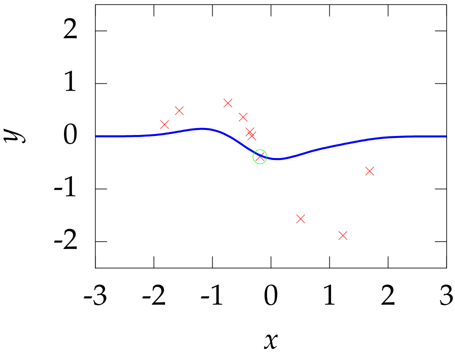
Nonlinear Regression Example

- Iteration 3
  - $w_1 = 0.18076,$
    $w_2 = -0.4266,$
    $w_3 = 0.12473$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta\boldsymbol{\phi}_{10}\Delta y_{10}$



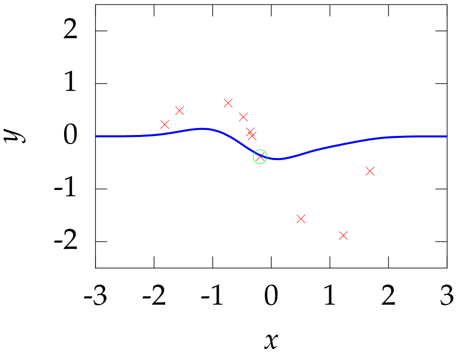
Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076$,
    $w_2 = -0.42893$,
    $w_3 = -0.14306$
  - Present data point 7



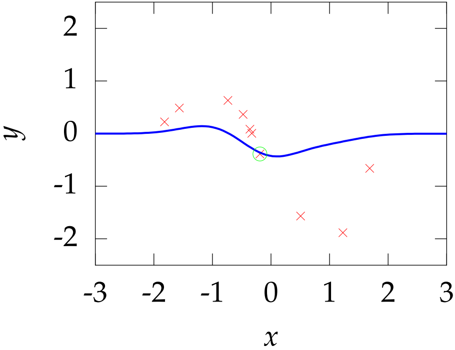
Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076$,
    $w_2 = -0.42893$,
    $w_3 = -0.14306$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$



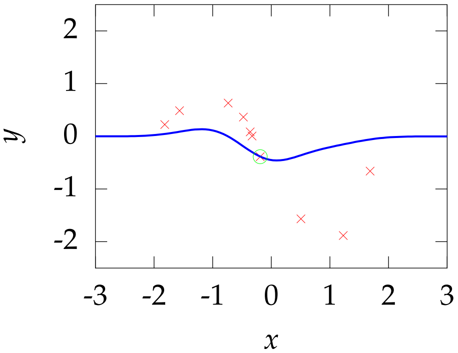
Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076$,
    $w_2 = -0.42893$,
    $w_3 = -0.14306$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



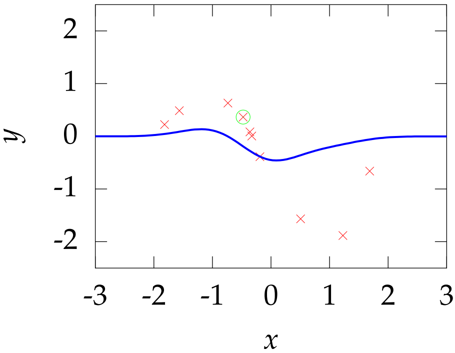
Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076,$
    $w_2 = -0.42893,$
    $w_3 = -0.14306$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



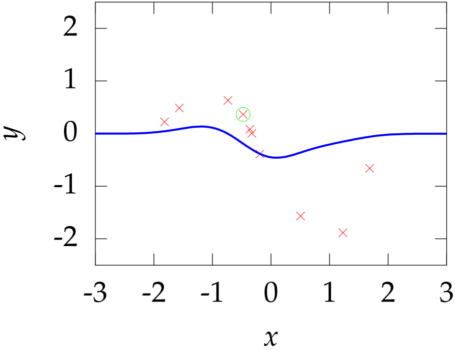
Nonlinear Regression Example

- Iteration 5
  - $w_1 = 0.17372$,
    $w_2 = -0.45335$,
    $w_3 = -0.14461$
  - Present data point 4



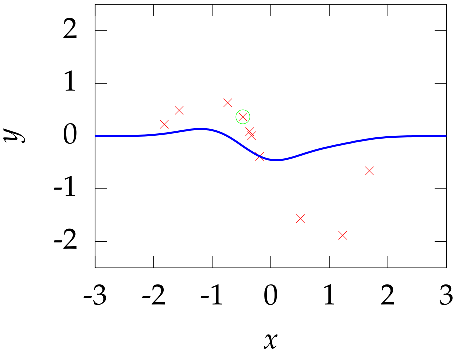
Nonlinear Regression Example

- Iteration 5
  - $w_1 = 0.17372,$
    $w_2 = -0.45335,$
    $w_3 = -0.14461$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$



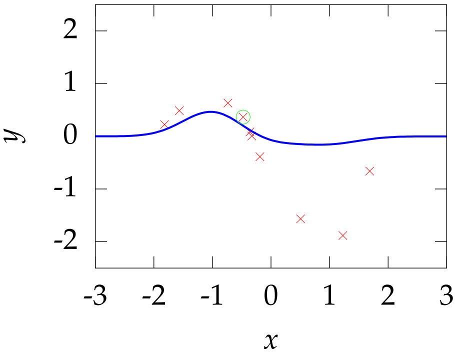
Nonlinear Regression Example

- Iteration 5
  - $w_1 = 0.17372,$
    $w_2 = -0.45335,$
    $w_3 = -0.14461$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



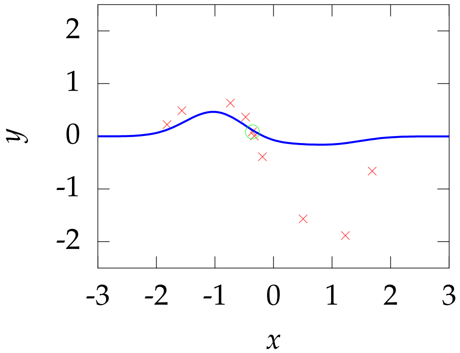
Nonlinear Regression Example

- Iteration 5
  - $w_1 = 0.17372$, $w_2 = -0.45335$, $w_3 = -0.14461$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



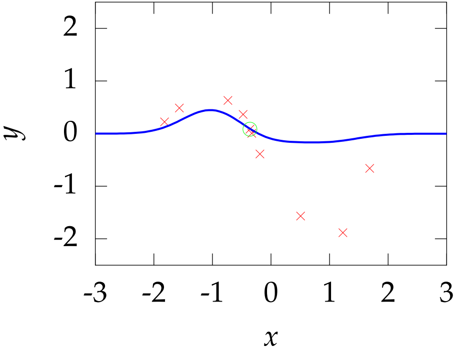
Nonlinear Regression Example

- Iteration 6
  - $w_1 = 0.47971,$
    $w_2 = -0.11541,$
    $w_3 = -0.13778$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



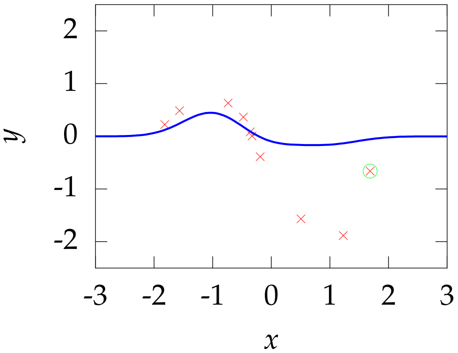
Nonlinear Regression Example

- Iteration 6
  - $w_1 = 0.47971,$
    $w_2 = -0.11541,$
    $w_3 = -0.13778$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



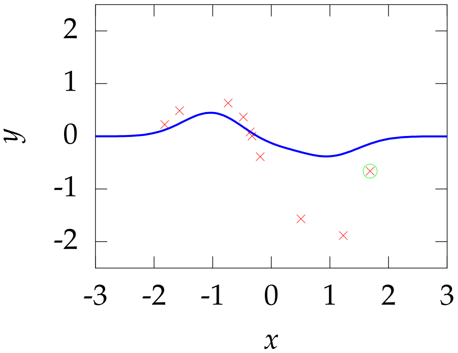
Nonlinear Regression Example

- Iteration 7
  - $w_1 = 0.46599$,
    $w_2 = -0.13952$,
    $w_3 = -0.13855$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



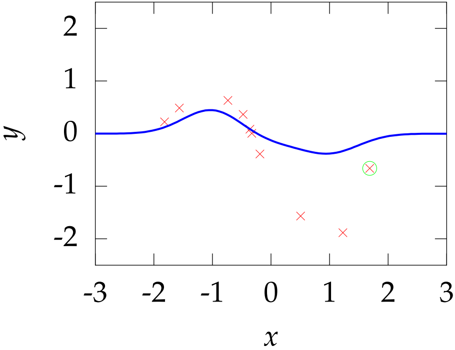
Nonlinear Regression Example

- Iteration 7
  - $w_1 = 0.46599$,
    $w_2 = -0.13952$,
    $w_3 = -0.13855$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta\boldsymbol{\phi}_{10}\Delta y_{10}$



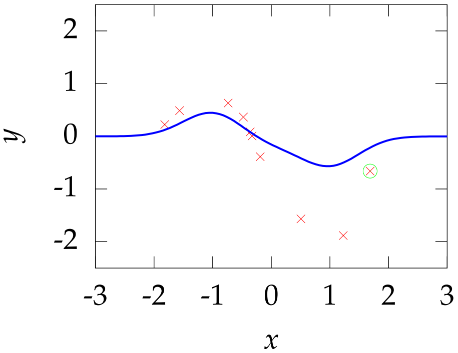
Nonlinear Regression Example

- Iteration 8
  - $w_1 = 0.46599,$
    $w_2 = -0.14144,$
    $w_3 = -0.35924$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta\boldsymbol{\phi}_{10}\Delta y_{10}$



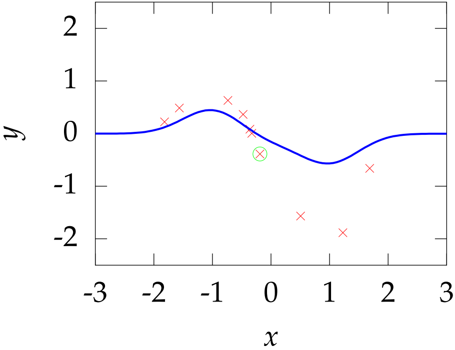
Nonlinear Regression Example

- Iteration 8
  - $w_1 = 0.46599$,
    $w_2 = -0.14144$,
    $w_3 = -0.35924$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta\boldsymbol{\phi}_{10}\Delta y_{10}$



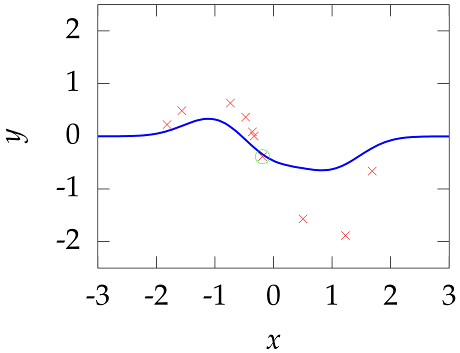
Nonlinear Regression Example

- Iteration 9
  - $w_1 = 0.46599$,
    $w_2 = -0.14307$,
    $w_3 = -0.54679$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



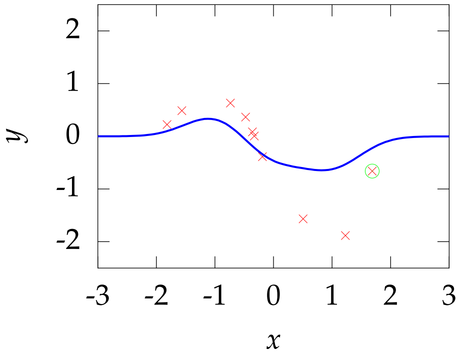
Nonlinear Regression Example

- Iteration 9
  - $w_1 = 0.46599$,
    $w_2 = -0.14307$,
    $w_3 = -0.54679$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



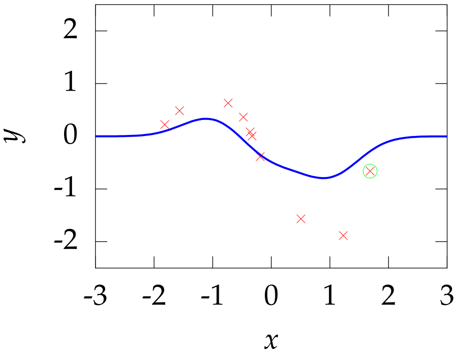
Nonlinear Regression Example

- Iteration 10
  - $w_1 = 0.38071$,
    $w_2 = -0.43867$,
    $w_3 = -0.56556$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



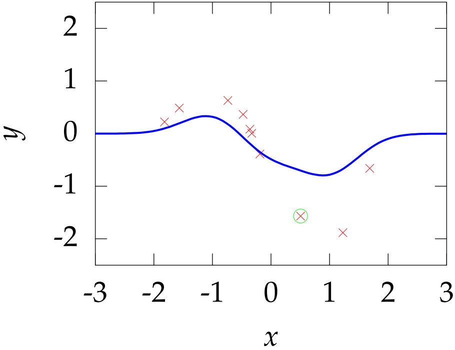
Nonlinear Regression Example

- Iteration 10
  - $w_1 = 0.38071$,
    $w_2 = -0.43867$,
    $w_3 = -0.56556$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



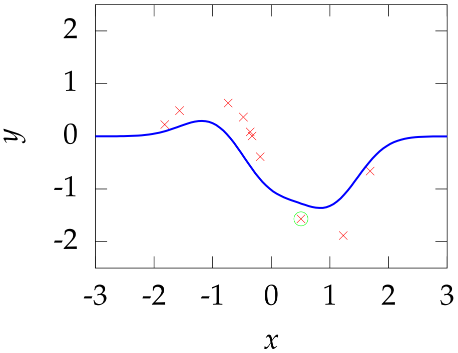
Nonlinear Regression Example

- Iteration 11
  - $w_1 = 0.38071$,
    $w_2 = -0.44002$,
    $w_3 = -0.7208$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



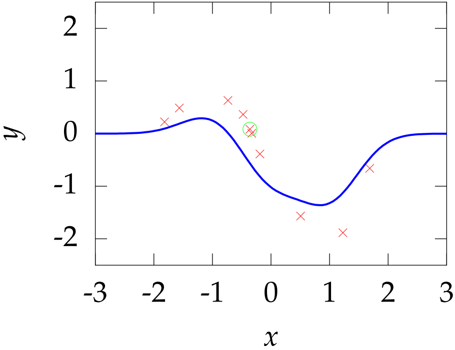
Nonlinear Regression Example

- Iteration 11
  - $w_1 = 0.38071$,
    $w_2 = -0.44002$,
    $w_3 = -0.7208$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



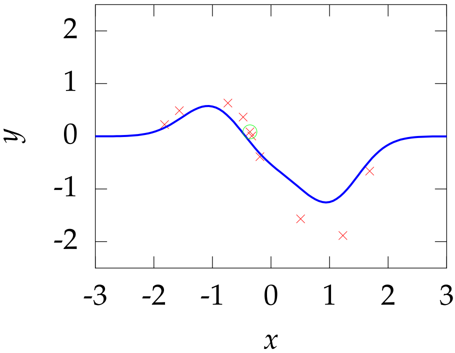
Nonlinear Regression Example

- Iteration 12
  - $w_1 = 0.37237$,
    $w_2 = -0.90666$,
    $w_3 = -1.1987$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



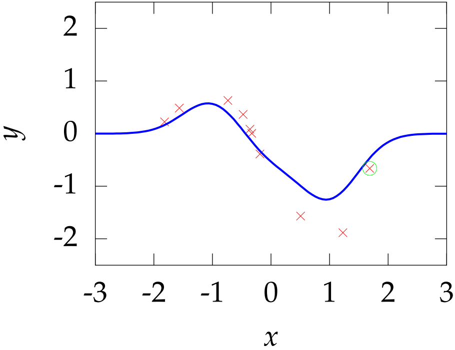
Nonlinear Regression Example

- Iteration 12
  - $w_1 = 0.37237,$
    $w_2 = -0.90666,$
    $w_3 = -1.1987$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



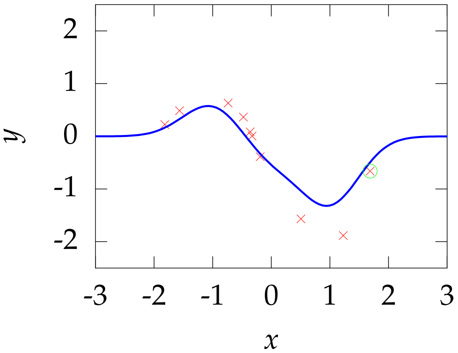
Nonlinear Regression Example

- Iteration 13
  - $w_1 = 0.62833,$
    $w_2 = -0.45691,$
    $w_3 = -1.1842$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta\boldsymbol{\phi}_{10}\Delta y_{10}$



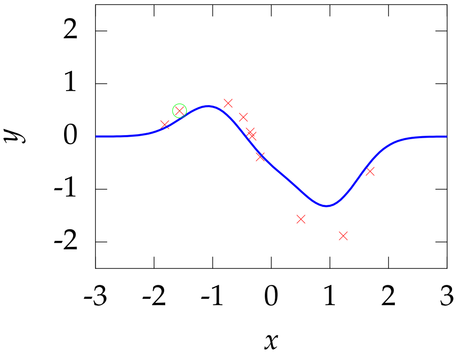
Nonlinear Regression Example

- Iteration 13
  - $w_1 = 0.62833,$
    $w_2 = -0.45691,$
    $w_3 = -1.1842$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta\boldsymbol{\phi}_{10}\Delta y_{10}$



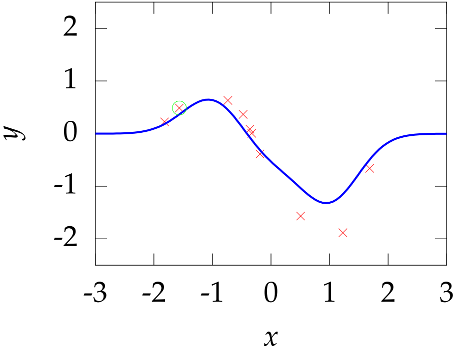
Nonlinear Regression Example

- Iteration 14
  - $w_1 = 0.62833$,
    $w_2 = -0.4575$,
    $w_3 = -1.252$
  - Present data point 2
  - $\Delta y_2 = y_2 - \phi_2^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_2 \Delta y_2$



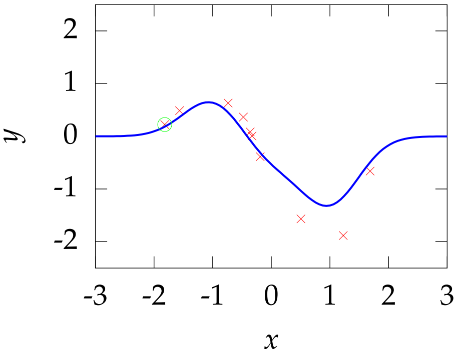
Nonlinear Regression Example

- Iteration 14
  - $w_1 = 0.62833,$
    $w_2 = -0.4575,$
    $w_3 = -1.252$
  - Present data point 2
  - $\Delta y_2 = y_2 - \phi_2^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_2 \Delta y_2$



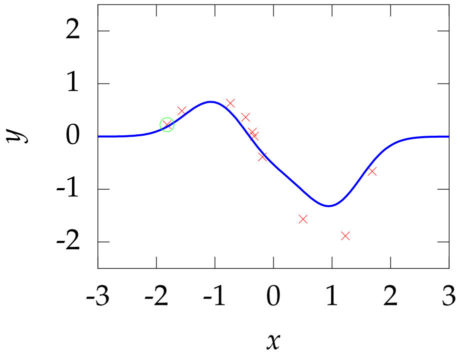
Nonlinear Regression Example

- Iteration 15
  - $w_1 = 0.7016,$
    $w_2 = -0.45646,$
    $w_3 = -1.252$
  - Present data point 1
  - $\Delta y_1 = y_1 - \phi_1^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_1 \Delta y_1$



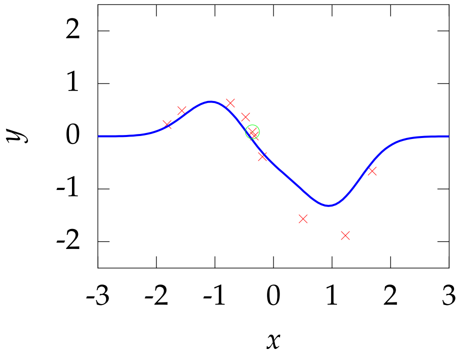
Nonlinear Regression Example

- Iteration 15
  - $w_1 = 0.7016,$
    $w_2 = -0.45646,$
    $w_3 = -1.252$
  - Present data point 1
  - $\Delta y_1 = y_1 - \phi_1^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_1 \Delta y_1$



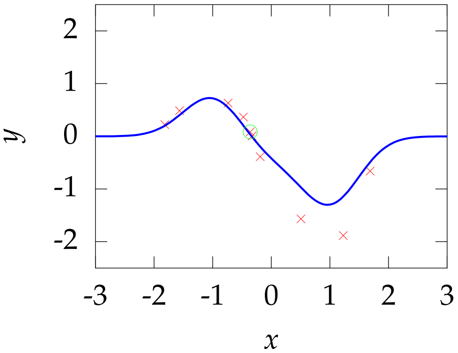
Nonlinear Regression Example

- Iteration 16
  - $w_1 = 0.7109,$
    $w_2 = -0.45641,$
    $w_3 = -1.252$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



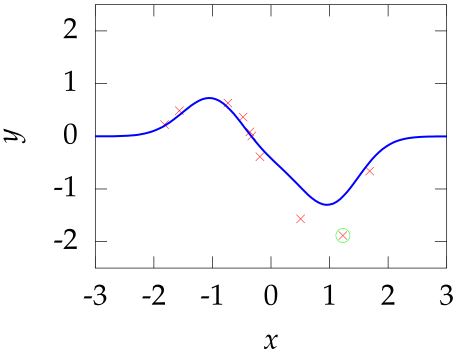
Nonlinear Regression Example

- Iteration 16
  - $w_1 = 0.7109$,
    $w_2 = -0.45641$,
    $w_3 = -1.252$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



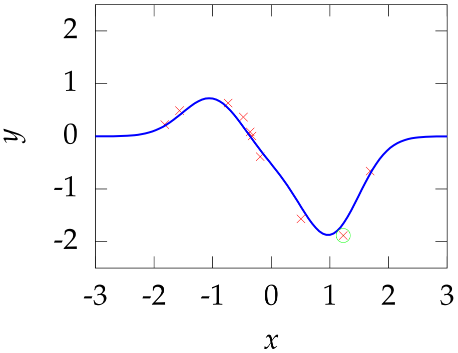
Nonlinear Regression Example

- Iteration 17
  - $w_1 = 0.77022$,
    $w_2 = -0.35219$,
    $w_3 = -1.2487$
  - Present data point 9
  - $\Delta y_9 = y_9 - \phi_9^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_9 \Delta y_9$



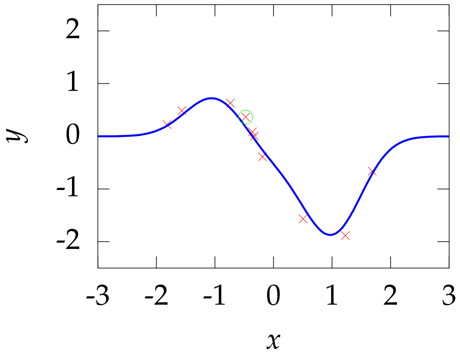
Nonlinear Regression Example

- Iteration 17
  - $w_1 = 0.77022$,
    $w_2 = -0.35219$,
    $w_3 = -1.2487$
  - Present data point 9
  - $\Delta y_9 = y_9 - \phi_9^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_9 \Delta y_9$



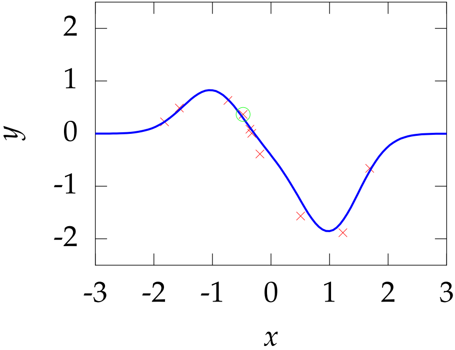
Nonlinear Regression Example

- Iteration 18
  - $w_1 = 0.77019$,
    $w_2 = -0.3832$,
    $w_3 = -1.8175$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



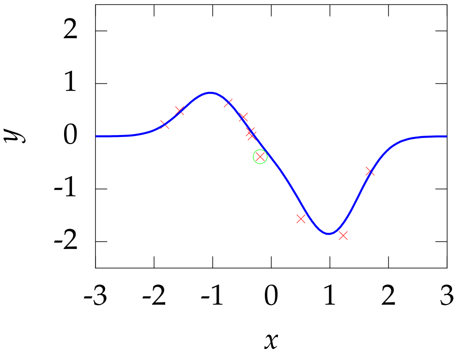
Nonlinear Regression Example

- Iteration 18
  - $w_1 = 0.77019$,
    $w_2 = -0.3832$,
    $w_3 = -1.8175$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



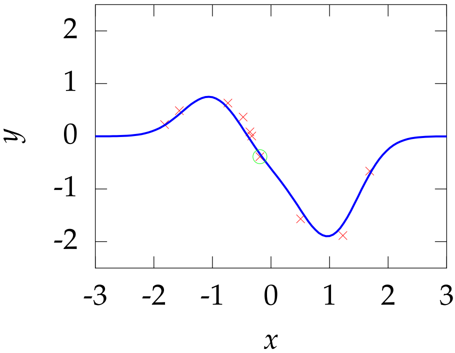
Nonlinear Regression Example

- Iteration 19
  - $w_1 = 0.86321$,
    $w_2 = -0.28046$,
    $w_3 = -1.8154$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



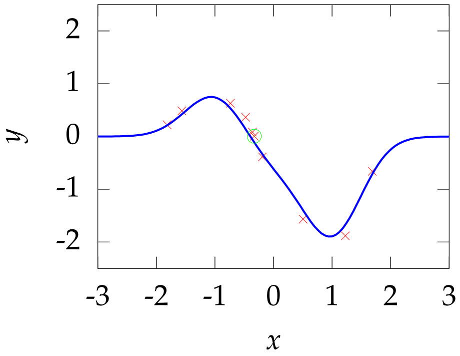
Nonlinear Regression Example

- Iteration 19
  - $w_1 = 0.86321,$
    $w_2 = -0.28046,$
    $w_3 = -1.8154$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



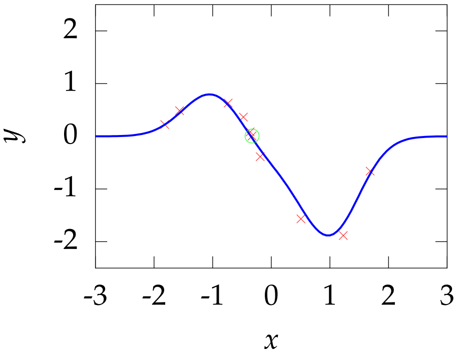
Nonlinear Regression Example

- Iteration 20
  - $w_1 = 0.80681$,
    $w_2 = -0.47597$,
    $w_3 = -1.8278$
  - Present data point 6
  - $\Delta y_6 = y_6 - \phi_6^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_6 \Delta y_6$



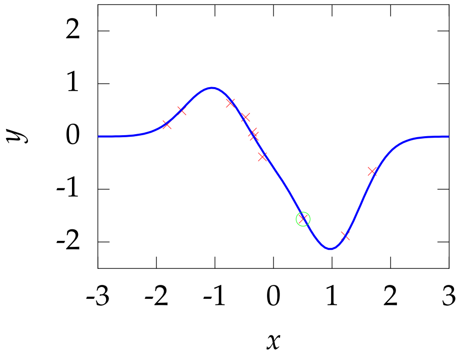
Nonlinear Regression Example

- Iteration 20
  - $w_1 = 0.80681$,
    $w_2 = -0.47597$,
    $w_3 = -1.8278$
  - Present data point 6
  - $\Delta y_6 = y_6 - \phi_6^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_6 \Delta y_6$



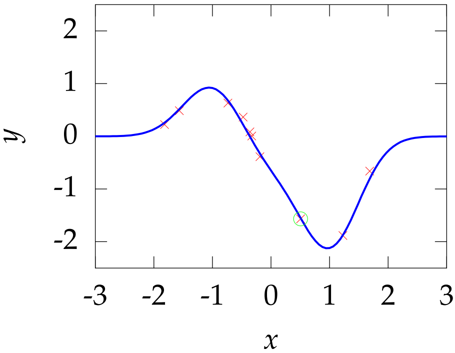
Nonlinear Regression Example

- Iteration 50
  - $w_1 = 0.9777$,
    $w_2 = -0.4076$,
    $w_3 = -2.038$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



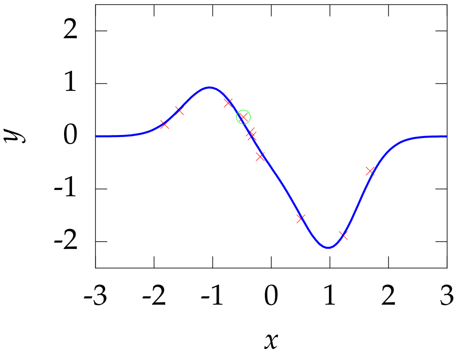
Nonlinear Regression Example

- Iteration 100
  - $w_1 = 0.98593,$
    $w_2 = -0.49744,$
    $w_3 = -2.046$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



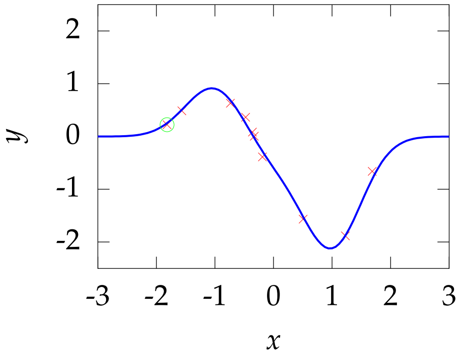
Nonlinear Regression Example

- Iteration 200
  - $w_1 = 0.95307,$
    $w_2 = -0.48041,$
    $w_3 = -2.0553$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



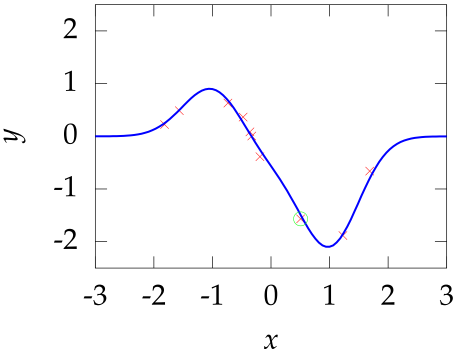
Nonlinear Regression Example

- Iteration 300
  - $w_1 = 0.97066,$
    $w_2 = -0.44667,$
    $w_3 = -2.0588$
  - Present data point 1
  - $\Delta y_1 = y_1 - \phi_1^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_1 \Delta y_1$



Nonlinear Regression Example

- Iteration 400
  - $w_1 = 0.95515,$
    $w_2 = -0.40611,$
    $w_3 = -2.0289$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



Mathematical Interpretation

- What is the mathematical interpretation?
  - There is a cost function.
  - It expresses mismatch between your prediction and reality.

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{K} w_j \phi_j(x_i) - y_i \right)^2$$

  - This is known as the sum of squares error.

Mathematical Interpretation

- What is the mathematical interpretation?
  - There is a cost function.
  - It expresses mismatch between your prediction and reality.

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left(\mathbf{w}^\top \boldsymbol{\phi}_i - y_i\right)^2$$

  - This is known as the sum of squares error.
  - Defining $\boldsymbol{\phi}_i = \left[\phi_1(x_i), \dots, \phi_K(x_i)\right]^\top$.

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Gradient of error function:

$$\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}} = -2\sum_{i=1}^{n} \boldsymbol{\phi}_i \left(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i\right)$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Gradient of error function:

$$\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}} = -2\sum_{i=1}^{n} \boldsymbol{\phi}_i \Delta y_i$$

- Where $\Delta y_i = (y_i - \mathbf{w}^\top \boldsymbol{\phi}_i)$.

Minimization via Gradient Descent

- One way of minimizing is steepest descent.
- Initialize algorithm with $\mathbf{w}$.
- Compute gradient of error function, $\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}}$.
- Change $\mathbf{w}$ by moving in steepest downhill direction.

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}}$$

Steepest Descent

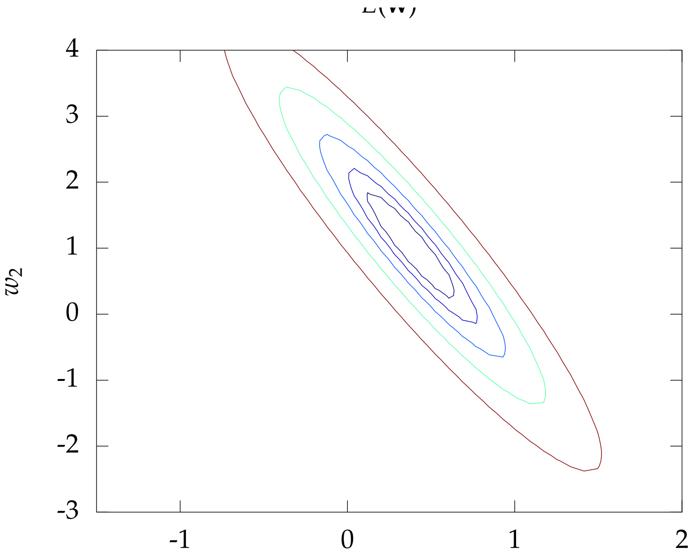


Figure: Steepest descent on a quadratic error surface.

Stochastic Gradient Descent

How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}}$$

Stochastic Gradient Descent

How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - 2\eta \sum_{i=1}^{n} \boldsymbol{\phi}_i \left(\mathbf{w}^\top \boldsymbol{\phi}_i - y_i \right)$$

Stochastic Gradient Descent

How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \left(\mathbf{w}^\top \boldsymbol{\phi}_i - y_i \right)$$

Stochastic Gradient Descent

How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \left(\mathbf{w}^\top \boldsymbol{\phi}_i - y_i\right)$$

Stochastic Gradient Descent

How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \Delta y_i$$

Stochastic Gradient Descent

How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \Delta y_i$$

- And the stochastic approximation is

$$\mathbf{w} \leftarrow \mathbf{w} + \eta' \boldsymbol{\phi}_i \Delta y_i$$

Stochastic Gradient Descent

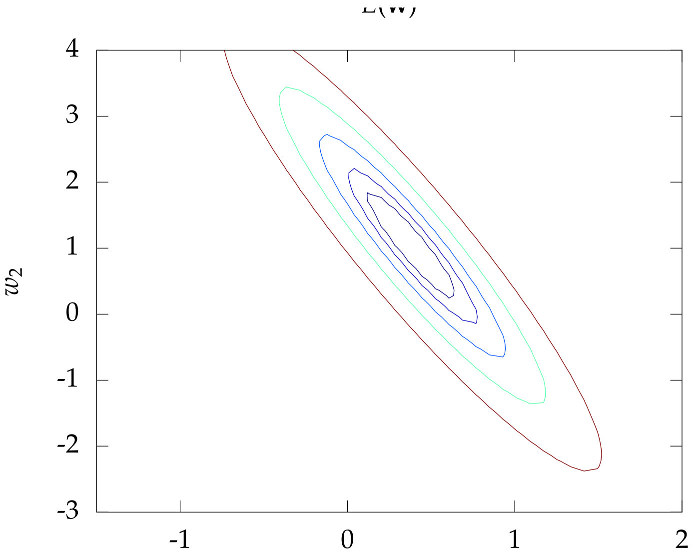


Figure: Stochastic gradient descent on a quadratic error surface.

Modern View of Error Functions

- Error function has a probabilistic interpretation (maximum likelihood).
- Error function is an actual loss function that you want to minimize (empirical risk minimization).
- For these interpretations probability and optimization theory become important.
- Much of the last 15 years of machine learning research has focused on probabilistic interpretations or clever relaxations of difficult objective functions.

Important Concepts Not Covered

- Optimization methods.
  - Second order methods, conjugate gradient, quasi-Newton and Newton.
  - Effective heuristics such as momentum.
- Local vs global solutions.

Clustering

- Divide data into discrete groups according to characteristics.
  - For example different animal species.
  - Different political parties.
- Determine the allocation to the groups and (harder) number of different groups.

K-means Clustering

An Algorithm

- *Require*: Set of K cluster centers & assignment of each point to a cluster.
  - Initialize cluster centers as data points.
  - Assign each data point to nearest cluster center.
  - Update each cluster center by setting it to the mean of assigned data points.

Objective Function

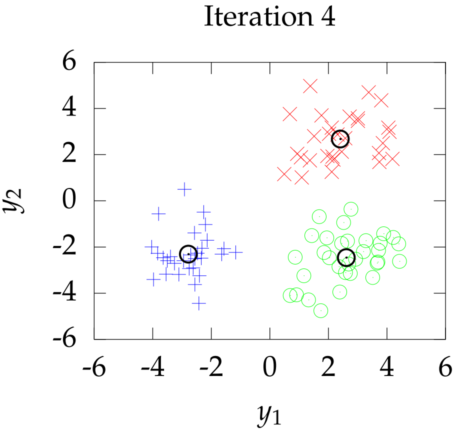
- This minimizes the objective:

$$\sum_{j=1}^{K} \sum_{i \text{ allocated to } j} \left(\mathbf{y}_{i,:} - \boldsymbol{\mu}_{j,:}\right)^{\top} \left(\mathbf{y}_{i,:} - \boldsymbol{\mu}_{j,:}\right)$$

- i.e. it minimizes the sum of Euclidean squared distances between points and their associated centers.
- The minimum is not guaranteed to be *global* or *unique*.
  - This objective is a non-convex optimization problem.

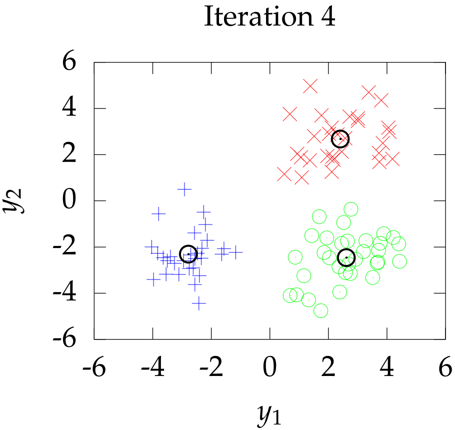
K-means Clustering

- K-means clustering.
  - Update each center by setting to the mean of the allocated points.



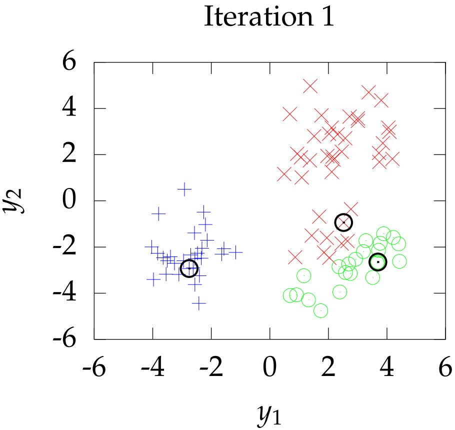
K-means Clustering

- *K*-means clustering.
  - Allocate each data point to the nearest cluster center.



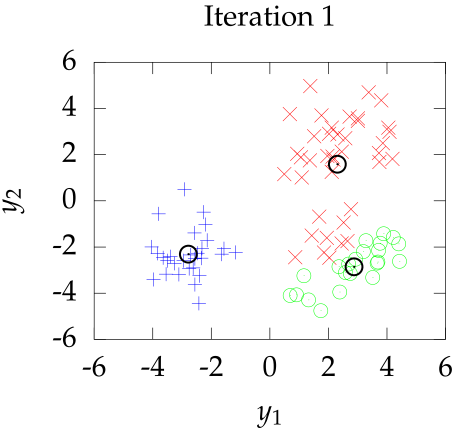
K-means Clustering

- *K*-means clustering.
  - Update each center by setting to the mean of the allocated points.



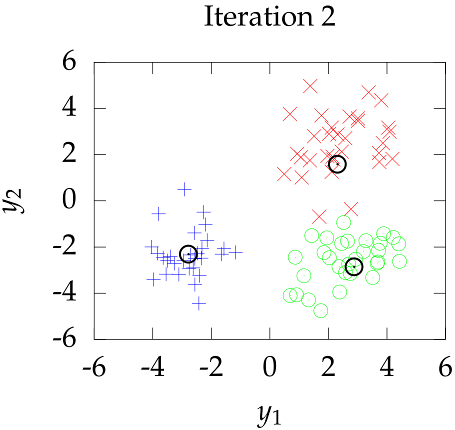
K-means Clustering

- *K*-means clustering.
  - Allocate each data point to the nearest cluster center.



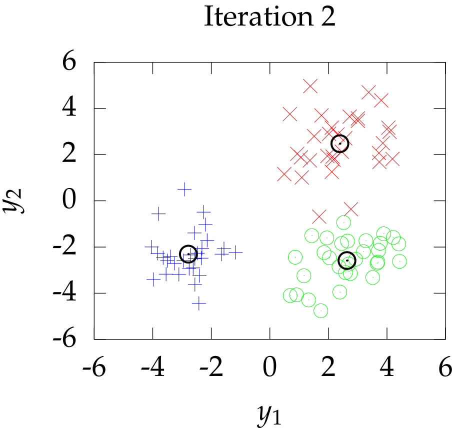
K-means Clustering

- *K*-means clustering.
  - Update each center by setting to the mean of the allocated points.



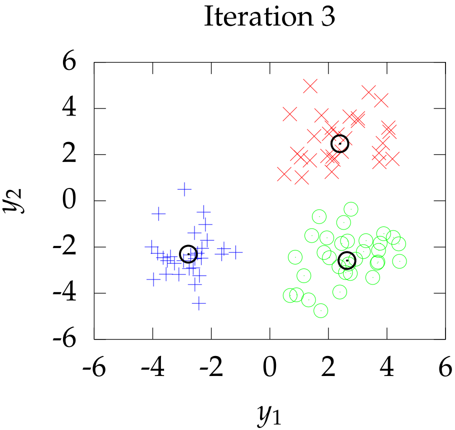
K-means Clustering

- *K*-means clustering.
  - Allocate each data point to the nearest cluster center.



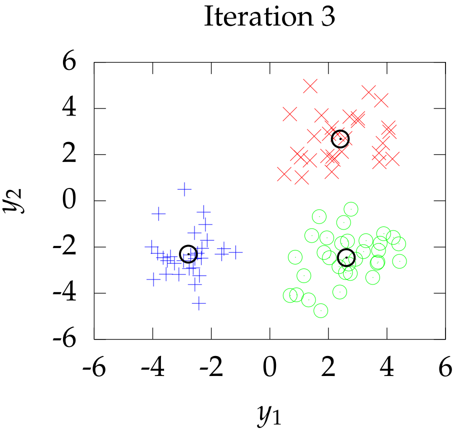
K-means Clustering

- *K*-means clustering.
  - Update each center by setting to the mean of the allocated points.



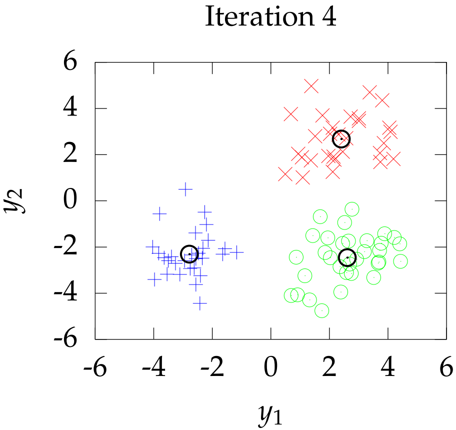
K-means Clustering

- *K*-means clustering.
  - Allocate each data point to the nearest cluster center.



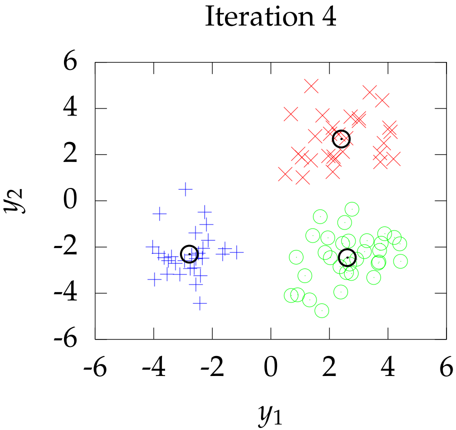
K-means Clustering

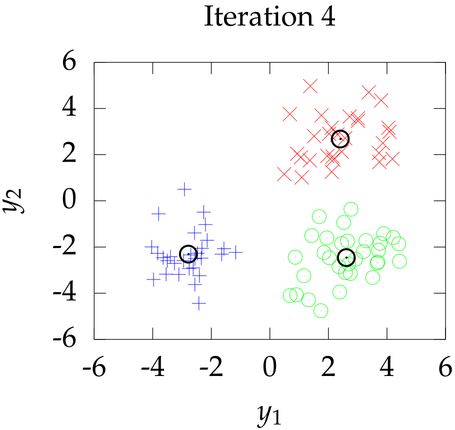
- *K*-means clustering.
  - Update each center by setting to the mean of the allocated points.



K-means Clustering

- *K*-means clustering.
  - Allocate each data point to the nearest cluster center.





- K-means clustering.
  - Allocation doesn't change so stop.

Other Clustering Approaches

- Spectral clustering (**??**).
  - Allows clusters which aren't convex hulls.
- Dirichlet processes
  - A probabilistic formulation for a clustering algorithm that is non-parameteric.

Mixture of Gaussians I

- Probabilistic clustering methods.
- Bayesian equivalent of K-means.
- Assume data is sampled from a Gaussian density:

$$p(\mathbf{y}_i|\mathbf{s}_i) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{y}_i|\boldsymbol{\mu}_k, \mathbf{C}_k\right)^{s_{i,k}}$$

- Where $\mathbf{s}_i$ is a binary vector encoding component with 1-of-n encoding.
- Multinomial prior over $\mathbf{s}_i$

$$p(\mathbf{s}_i) = \prod_{k=1}^{K} \pi_k^{s_{i,k}}$$

EM Algorithm

- Marginal likelihood

$$\log p(\mathbf{y}_i) = \log \sum_{\mathbf{s}_i} p(\mathbf{y}_i, \mathbf{s}_i)$$

EM Algorithm

- Marginal likelihood

$$\log p(\mathbf{y}_i) = \log \sum_{\mathbf{s}_i} p(\mathbf{y}_i, \mathbf{s}_i)$$

$$\log p(\mathbf{y}_i) = \log \sum_{\mathbf{s}_i} q(\mathbf{s}_i) \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{q(\mathbf{s}_i)}$$

EM Algorithm

- Marginal likelihood

$$\log p(\mathbf{y}_i) = \log \sum_{\mathbf{s}_i} q(\mathbf{s}_i) \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{q(\mathbf{s}_i)}$$

$$\log p(\mathbf{y}_i) \geq \sum_{\mathbf{s}_i} q(\mathbf{s}_i) \log \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{q(\mathbf{s}_i)}$$

- Jensen's inequality gives a bound.

EM Algorithm

- Marginal likelihood

$$\log p(\mathbf{y}_i) \geq \sum_{\mathbf{s}_i} q(\mathbf{s}_i) \log \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{q(\mathbf{s}_i)}$$

$$\log p(\mathbf{y}_i) = \sum_{\mathbf{s}_i} p(\mathbf{s}_i|\mathbf{y}_i) \log \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{p(\mathbf{s}_i|\mathbf{y}_i)}$$

- Jensen's inequality gives a bound.
- Bound becomes equality if $q(\mathbf{s}_i) = p(\mathbf{s}_i|\mathbf{y}_i)$

$$p(\mathbf{y}_i) = \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{p(\mathbf{s}_i|\mathbf{y}_i)}$$

EM Algorithm

- Iterate between
  1. **E Step** Set $q(\mathbf{s}_i) = p(\mathbf{s}_i|\mathbf{y}_i)$
  2. **M Step** Maximize $\sum_{\mathbf{s}_i} q(\mathbf{s}_i) \log p(\mathbf{y}_i, \mathbf{s}_i)$ with respect to parameters.

EM for Mixtures of Gaussians

- Iterate between
  1. **E Step** Set $q(\mathbf{s}_i) = \prod_{k=1}^{K} r_{i,k}^{s_{i,k}}$ where

$$r_{i,k} = \frac{\pi_k \mathcal{N}\left(\mathbf{y}_i | \boldsymbol{\mu}_k, \mathbf{C}_k\right)}{\sum_k \pi_k \mathcal{N}\left(\mathbf{y}_i | \boldsymbol{\mu}_k, \mathbf{C}_k\right)}$$

  2. **M Step** Maximize $\langle \log p(\mathbf{y}_i, \mathbf{s}_i) \rangle_{q(\mathbf{s}_i)}$ by setting

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} r_{i,k}, \quad \boldsymbol{\mu}_k = \frac{1}{\bar{n}_k} \sum_{i=1}^{n} r_{i,k} \mathbf{y}_i$$

$$\mathbf{C}_k = \frac{1}{\bar{n}_k} \sum_{i=1}^{n} r_{i,k} (\mathbf{y}_i - \boldsymbol{\mu}_k)(\mathbf{y}_i - \boldsymbol{\mu}_k)^\top$$

$$\bar{n}_k = \sum_{i=1}^{n} r_{i,k}$$

Netlab Demo

`demgmm1.m`

Variational Inference

- EM algorithm relies on computation of setting $q(\mathbf{s}_i)$ to $p(\mathbf{s}_i|y_i)$.
- In variational inference we use approximate posteriors for the $q(\cdot)$ distributions.
- This makes the algorithms tractable but non exact.

Conclusions

- Bayesian approach treats parameters as random variables.
- Learning proceeds through combination of prior and likelihood.
- Latent variable models and mixture of Gaussians are not Bayesian but use Bayes' rule.
- All these models sit in the wider family of probabilistic models.

References I