# Uncertainty and Probability

MLAI: Week 1

Neil D. Lawrence

Department of Computer Science
Sheffield University

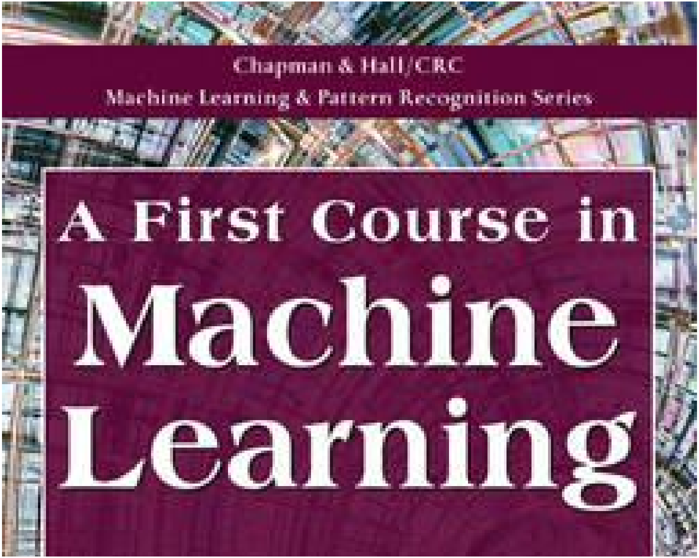
30th October 2014

# Outline

Course Text

Review: Basic Probability

# Rogers and Girolami



# Bishop


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP

# 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 =

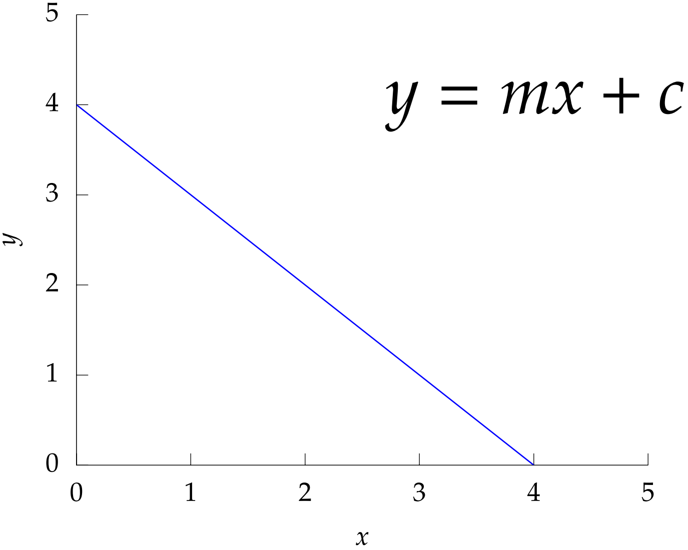
- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

# What is Machine Learning?

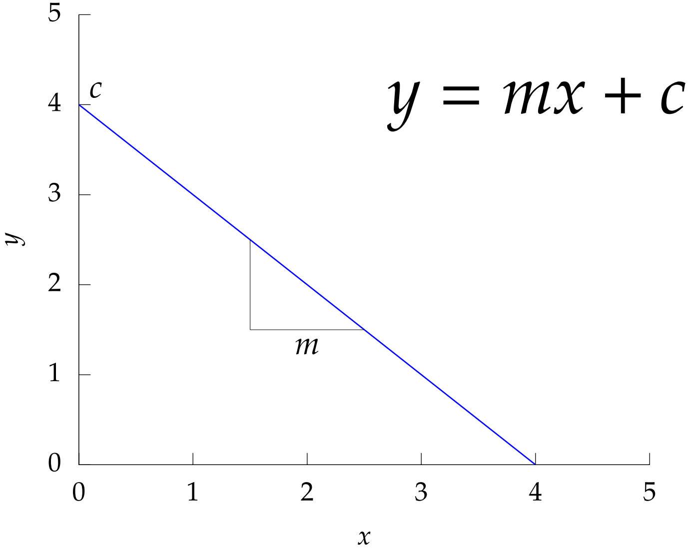
data + model = prediction

- 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.

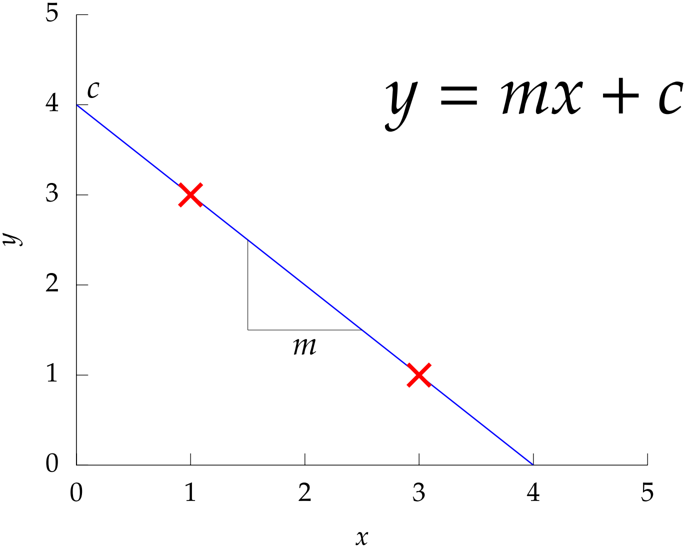
$$y = mx + c$$



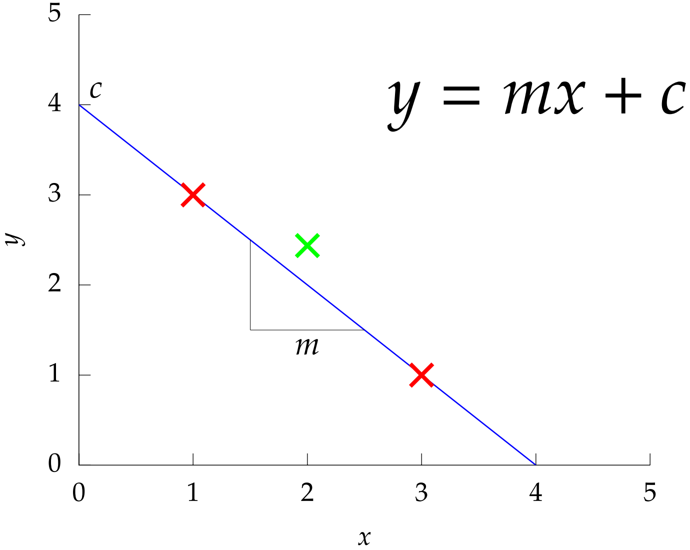
$y = mx + c$
0
1
2
3
4
5
0
1
2
3
4
5
$x$
$y$

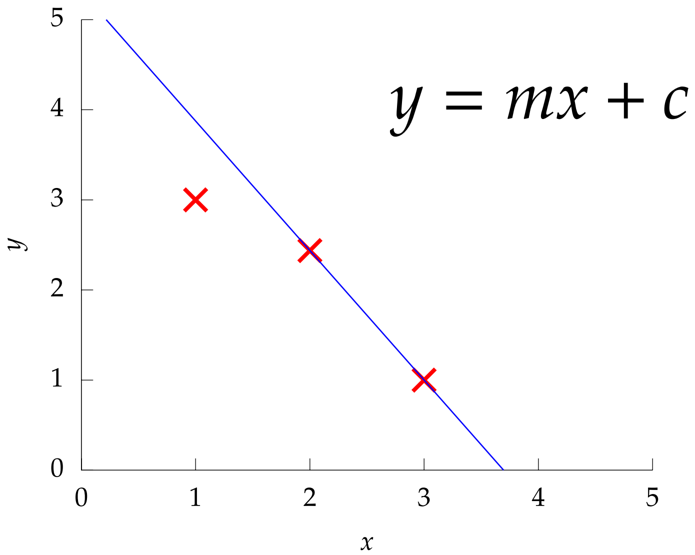


$y = mx + c$
$c$
$m$
0
1
2
3
4
5
$x$
$y$

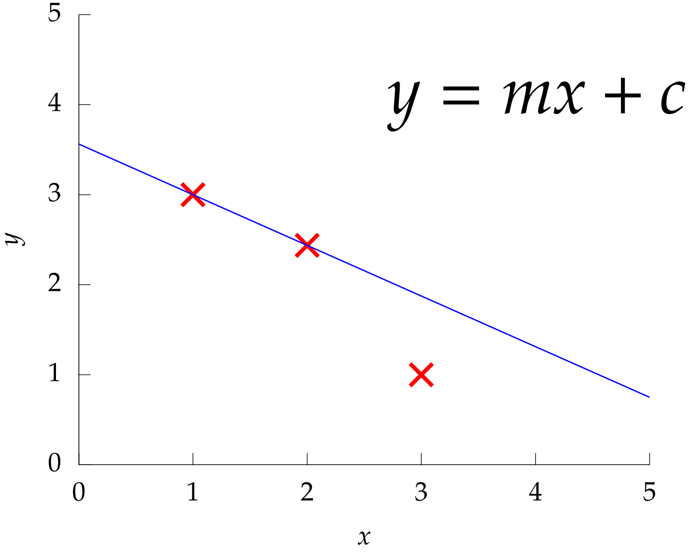


$y = mx + c$
c
m
y
x
0
1
2
3
4
5

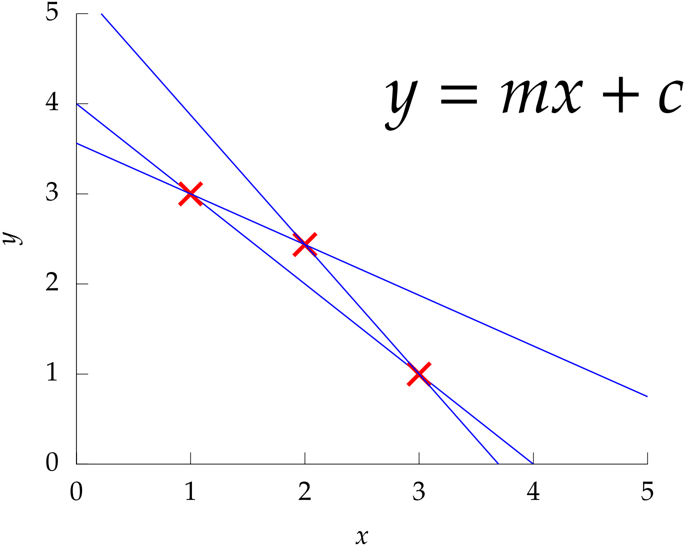

$y = mx + c$
$c$
$m$
$x$
$y$
0
1
2
3
4
5
0
1
2
3
4
5



$y = mx + c$
$y$
$x$
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
$y$
$x$



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

$y = mx + c$

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

$$3 = m + c$$

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

$$1 = 3m + c$$

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

$$2.5 = 2m + c$$

riens. L'opinion contraire est une illusion de l'esprit qui, perdant de vue les raisons fugitives du choix de la volonté dans les choses indifférentes, se persuade qu'elle s'est déterminée d'elle-même et sans motifs.

Nous devons donc envisager l'état présent de l'univers, comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui, pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvemens des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir comme le passé, serait présent à ses yeux. L'esprit humain offre, dans la perfection qu'il a su donner à l'Astronomie, une faible esquisse de cette intelligence. Ses découvertes en Mécanique et en *Géométrie*, *jointes* à celle de la pesanteur universelle, *l'ont mis à portée* de comprendre dans les *mêmes* expressions analytiques, les états passés et futurs du système du monde. En appliquant la même méthode à quelques autres objets de ses connaissances, il est parvenu à ramener à des lois générales, les phénomènes observés, et à prévoir ceux que des circonstances données doivent faire éclore. Tous ces efforts dans la recherche de la vérité, tendent à le rapprocher sans cesse de l'intelligence que nous venons de concevoir, mais dont il restera toujours infiniment éloigné. Cette tendance propre à l'espèce humaine, est ce qui la rend supérieure aux animaux; et ses progrès en ce genre, distinguent les nations et les siècles, et font leur véritable gloire.

Rappelons-nous qu'autrefois, et à une époque qui 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 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

$y = mx + c + \epsilon$

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

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

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

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

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

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

# Outline

Course Text

Review: Basic Probability

# Probability Review I

- We are interested in trials which result in two random variables, $X$ and $Y$, each of which has an 'outcome' denoted by $x$ or $y$.
- We summarise the notation and terminology for these distributions in the following table.

| Terminology | Notation | Description |
|---|---|---|
| Joint Probability | $P(X = x, Y = y)$ | 'The probability that $X = x$ *and* $Y = y$' |
| Marginal Probability | $P(X = x)$ | 'The probability that $X = x$ *regardless of* $Y$' |
| Conditional Probability | $P(X = x \vert Y = y)$ | 'The probability that $X = x$ *given that* $Y = y$' |

Table : The different basic probability distributions.

# A Pictorial Definition of Probability

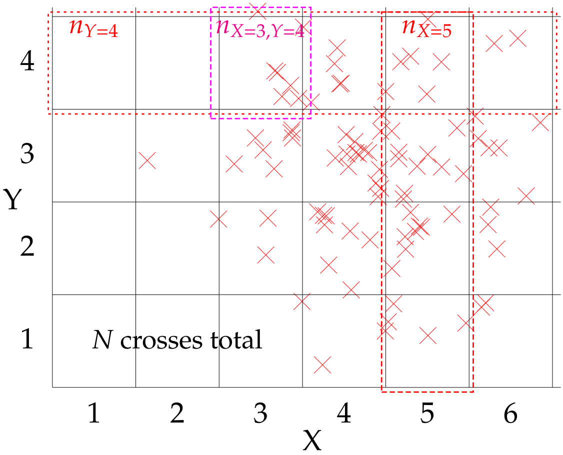



Figure : Representation of joint and conditional probabilities.

# Different Distributions

| Terminology | Definition | Notation |
|---|---|---|
| Joint Probability | $\lim_{N\to\infty} \frac{n_{X=3,Y=4}}{N}$ | $P(X=3, Y=4)$ |
| Marginal Probability | $\lim_{N\to\infty} \frac{n_{X=5}}{N}$ | $P(X=5)$ |
| Conditional Probability | $\lim_{N\to\infty} \frac{n_{X=3,Y=4}}{n_{Y=4}}$ | $P(X=3 \mid Y=4)$ |

Table : Definition of probability distributions.

# Notational Details

- Typically we should write out $P(X = x, Y = y)$.
- In practice, we often use $P(x, y)$.
- This looks very much like we might write a multivariate function, *e.g.* $f(x, y) = \frac{x}{y}$.
  - For a multivariate function though, $f(x, y) \neq f(y, x)$.
  - However $P(x, y) = P(y, x)$ because $P(X = x, Y = y) = P(Y = y, X = x)$.
- We now quickly review the 'rules of probability'.

# Normalization

*All* distributions are normalized. This is clear from the fact that $\sum_x n_x = N$, which gives

$$\sum_x P(x) = \frac{\sum_x n_x}{N} = \frac{N}{N} = 1.$$

A similar result can be derived for the marginal and conditional distributions.

# The Sum Rule

Ignoring the limit in our definitions:

- The marginal probability $P(y)$ is $\frac{n_y}{N}$ (ignoring the limit).
- The joint distribution $P(x, y)$ is $\frac{n_{x,y}}{N}$.
- $n_y = \sum_x n_{x,y}$ so

$$\frac{n_y}{N} = \sum_x \frac{n_{x,y}}{N},$$

  in other words

$$P(y) = \sum_x P(x, y).$$

  This is known as the sum rule of probability.

# The Product Rule

- $P(x|y)$ is
$$\frac{n_{x,y}}{n_y}.$$
- $P(x,y)$ is
$$\frac{n_{x,y}}{N} = \frac{n_{x,y}}{n_y}\frac{n_y}{N}$$
  or in other words
$$P(x,y) = P(x|y)\,P(y)\,.$$
  This is known as the product rule of probability.

# Bayes' Rule

- From the product rule,

$$P(y, x) = P(x, y) = P(x|y)\,P(y)\,,$$

so

$$P(y|x)\,P(x) = P(x|y)\,P(y)$$

which leads to Bayes' rule,

$$P(y|x) = \frac{P(x|y)\,P(y)}{P(x)}.$$

# Bayes' Theorem Example

- There are two barrels in front of you. Barrel One contains 20 apples and 4 oranges. Barrel Two other contains 4 apples and 8 oranges. You choose a barrel randomly and select a fruit. It is an apple. What is the probability that the barrel was Barrel One?

# Bayes' Theorem Example: Answer I

- We are given that:

$$
\begin{aligned}
P(\text{F} = \text{A}|\text{B} = 1) =& 20/24 \\
P(\text{F} = \text{A}|\text{B} = 2) =& 4/12 \\
P(\text{B} = 1) =& 0.5 \\
P(\text{B} = 2) =& 0.5
\end{aligned}
$$

# Bayes' Theorem Example: Answer II

- We use the sum rule to compute:

$$
\begin{aligned}
P(\mathrm{F}=\mathrm{A}) = & P(\mathrm{F}=\mathrm{A}|\mathrm{B}=1)P(\mathrm{B}=1) \\
& + P(\mathrm{F}=\mathrm{A}|\mathrm{B}=2)P(\mathrm{B}=2) \\
= & 20/24 \times 0.5 + 4/12 \times 0.5 = 7/12
\end{aligned}
$$

# Bayes' Theorem Example: Answer II

- We use the sum rule to compute:

$$
\begin{aligned}
P(\mathrm{F} = \mathrm{A}) =& P(\mathrm{F} = \mathrm{A} | \mathrm{B} = 1) P(\mathrm{B} = 1) \\
&+ P(\mathrm{F} = \mathrm{A} | \mathrm{B} = 2) P(\mathrm{B} = 2) \\
=& 20/24 \times 0.5 + 4/12 \times 0.5 = 7/12
\end{aligned}
$$

- And Bayes' theorem tells us that:

$$
\begin{aligned}
P(\mathrm{B} = 1 | \mathrm{F} = \mathrm{A}) =& \frac{P(\mathrm{F} = \mathrm{A} | \mathrm{B} = 1) P(\mathrm{B} = 1)}{P(\mathrm{F} = \mathrm{A})} \\
=& \frac{20/24 \times 0.5}{7/12} = 5/7
\end{aligned}
$$

# Reading & Exercises

Before Friday, review the example on Bayes Theorem!

- Read and *understand* Bishop on probability distributions: page 12–17 (Section 1.2).
- Complete Exercise 1.3 in Bishop.

# Distribution Representation

- We can represent probabilities as tables

| $y$ | 0 | 1 | 2 |
|---|---|---|---|
| $P(y)$ | 0.2 | 0.5 | 0.3 |

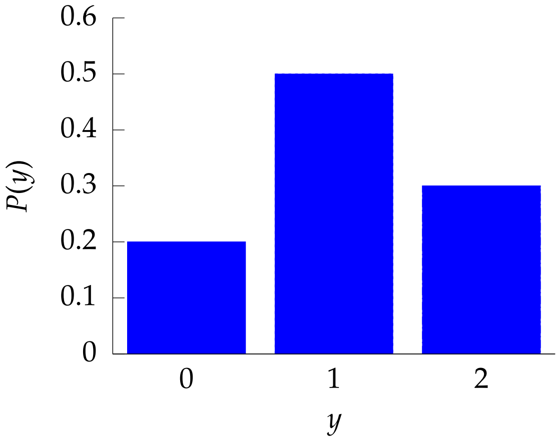



Figure : Histogram representation of the simple distribution.

# Expectations of Distributions

- Writing down the entire distribution is tedious.
- Can summarise through expectations.

$$\langle f(y) \rangle_{P(y)} = \sum_y f(y)p(y)$$

- Consider:

| $y$ | 0 | 1 | 2 |
|---|---|---|---|
| $P(y)$ | 0.2 | 0.5 | 0.3 |

- We have $\langle y \rangle_{P(y)} = 0.2 \times 0 + 0.5 \times 1 + 0.3 \times 2 = 1.1$
- This is the *first moment* or mean of the distribution.

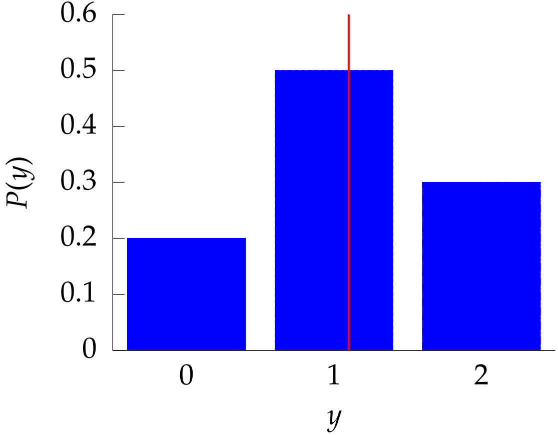



Figure : Histogram representation of the simple distribution including the expectation of $y$ (red line), the mean of the distribution.

# Variance and Standard Deviation

- Mean gives us the centre of the distribution.
- Consider:

| $y$ | 0 | 1 | 2 |
|---|---|---|---|
| $y^2$ | 0 | 1 | 4 |
| $P(y)$ | 0.2 | 0.5 | 0.3 |

- *Second moment* is $\left\langle y^2 \right\rangle_{P(y)} = 0.2 \times 0 + 0.5 \times 1 + 0.3 \times 4 = 1.7$
- Variance is $\left\langle y^2 \right\rangle - \langle y \rangle^2 = 1.7 - 1.1 \times 1.1 = 0.49$
- Standard deviation is square root of variance.
- Standard deviation gives us the “width” of the distribution.

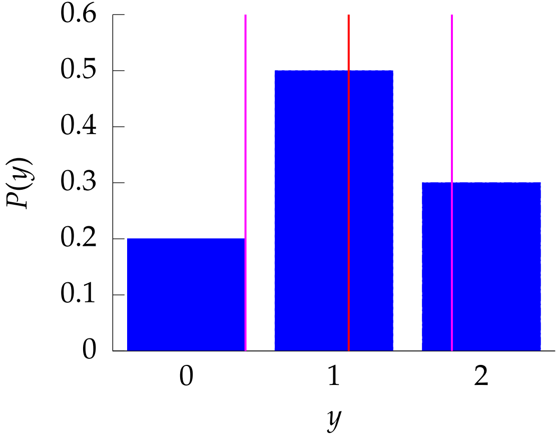



Figure : Histogram representation of the simple distribution including lines at one standard deviation from the mean of the distribution (magenta lines).

# Expectation Computation Example

- Consider the following distribution.

| $y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P(y)$ | 0.3 | 0.2 | 0.1 | 0.4 |

- What is the mean of the distribution?

# Expectation Computation Example

- Consider the following distribution.

| $y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P(y)$ | 0.3 | 0.2 | 0.1 | 0.4 |

- What is the mean of the distribution?
- What is the standard deviation of the distribution?

# Expectation Computation Example

- Consider the following distribution.

| $y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P(y)$ | 0.3 | 0.2 | 0.1 | 0.4 |

- What is the mean of the distribution?
- What is the standard deviation of the distribution?
- Are the mean and standard deviation representative of the distribution form?

# Expectation Computation Example

- Consider the following distribution.

| $y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P(y)$ | 0.3 | 0.2 | 0.1 | 0.4 |

- What is the mean of the distribution?
- What is the standard deviation of the distribution?
- Are the mean and standard deviation representative of the distribution form?
- What is the expected value of $-\log P(y)$?

# Expectations Example: Answer

- We are given that:

| $y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P(y)$ | 0.3 | 0.2 | 0.1 | 0.4 |
| $y^2$ | 1 | 4 | 9 | 16 |
| $-\log(P(y))$ | 1.204 | 1.609 | 2.302 | 0.916 |

- Mean: $1 \times 0.3 + 2 \times 0.2 + 3 \times 0.1 + 4 \times 0.4 = 2.6$
- Second moment: $1 \times 0.3 + 4 \times 0.2 + 9 \times 0.1 + 16 \times 0.4 = 8.4$
- Variance: $8.4 - 2.6 \times 2.6 = 1.64$
- Standard deviation: $\sqrt{1.64} = 1.2806$
- Expectation $-\log(P(y))$:
  $0.3 \times 1.204 + 0.2 \times 1.609 + 0.1 \times 2.302 + 0.4 \times 0.916 = 1.280$

# Sample Based Approximation Example

- You are given the following values samples of heights of students,

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $y_i$ | 1.76 | 1.73 | 1.79 | 1.81 | 1.85 | 1.80 |

- What is the sample mean?

# Sample Based Approximation Example

- You are given the following values samples of heights of students,

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $y_i$ | 1.76 | 1.73 | 1.79 | 1.81 | 1.85 | 1.80 |

- What is the sample mean?
- What is the sample variance?

# Sample Based Approximation Example

- You are given the following values samples of heights of students,

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $y_i$ | 1.76 | 1.73 | 1.79 | 1.81 | 1.85 | 1.80 |

- What is the sample mean?
- What is the sample variance?
- Can you compute sample approximation expected value of $-\log P(y)$?

# Sample Based Approximation Example

- You are given the following values samples of heights of students,

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $y_i$ | 1.76 | 1.73 | 1.79 | 1.81 | 1.85 | 1.80 |

- What is the sample mean?
- What is the sample variance?
- Can you compute sample approximation expected value of $-\log P(y)$?
- Actually these "data" were sampled from a Gaussian with mean 1.7 and standard deviation 0.15. Are your estimates close to the real values? If not why not?

# Sample Based Approximation Example: Answer

- We can compute:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $y_i$ | 1.76 | 1.73 | 1.79 | 1.81 | 1.85 | 1.80 |
| $y_i^2$ | 3.0976 | 2.9929 | 3.2041 | 3.2761 | 3.4225 | 3.2400 |

- Mean: $\frac{1.76+1.73+1.79+1.81+1.85+1.80}{6} = 1.79$
- Second moment: $\frac{3.0976+2.9929+3.2041+3.2761+3.4225+3.2400}{6} = 3.2055$
- Variance: $3.2055 - 1.79 \times 1.79 = 1.43 \times 10^{-3}$
- Standard deviation: 0.0379
- No, you can't compute it. You don't have access to $P(y)$ directly.

# Reading

- See probability review at end of slides for reminders.
- Read and *understand* **?** on:
  1. Section 2.2 (pg 41–53).
  2. Section 2.4 (pg 55–58).
  3. Section 2.5.1 (pg 58–60).
  4. Section 2.5.3 (pg 61–62).
- For other material in **?** read:
  1. Probability densities: Section 1.2.1 (Pages 17–19).
  2. Expectations and Covariances: Section 1.2.2 (Pages 19–20).
  3. The Gaussian density: Section 1.2.4 (Pages 24–28) (don't worry about material on bias).
  4. For material on information theory and KL divergence try Section 1.6 & 1.6.1 of **?** (pg 48 onwards).
- If you are unfamiliar with probabilities you should complete the following exercises:
  1. **?** Exercise 1.7
  2. **?** Exercise 1.8
  3. **?** Exercise 1.9

# References I


C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .

P. S. Laplace. *Essai philosophique sur les probabilités*. Courcier, Paris, 2nd edition, 1814. Sixth edition of 1840 translated and repreinted (1951) as *A Philosophical Essay on Probabilities*, New York: Dover; fifth edition of 1825 reprinted 1986 with notes by Bernard Bru, Paris: Christian Bourgois Éditeur, translated by Andrew Dale (1995) as *Philosophical Essay on Probabilities*, New York:Springer-Verlag.

S. Rogers and M. Girolami. *A First Course in Machine Learning*. CRC Press, 2011. [Google Books] .