Uncertainty and Probability

MLAI: Week 1

Neil D. Lawrence

Department of Computer Science
Sheffield University

1st October 2013

Outline

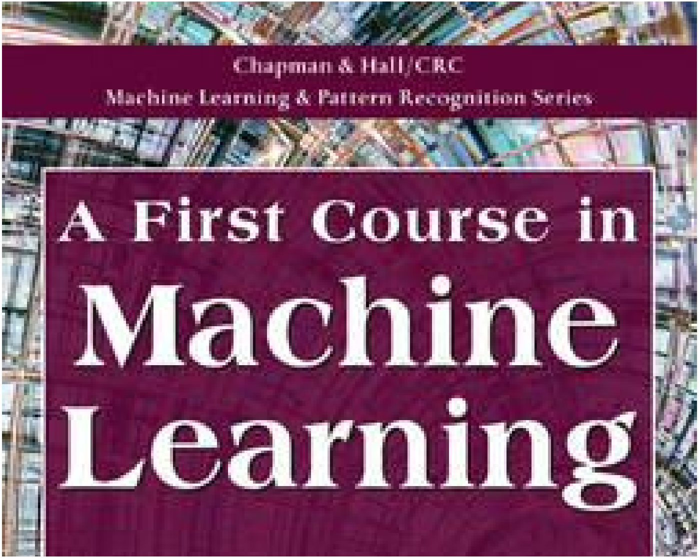
Course Text

Probability Density Functions

Sample Based Approximations

Maximum Likelihood

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 =

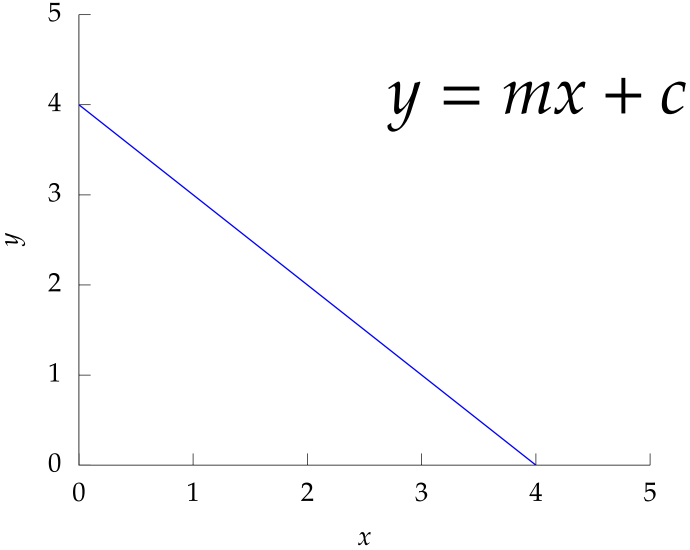
- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

What is Machine Learning?

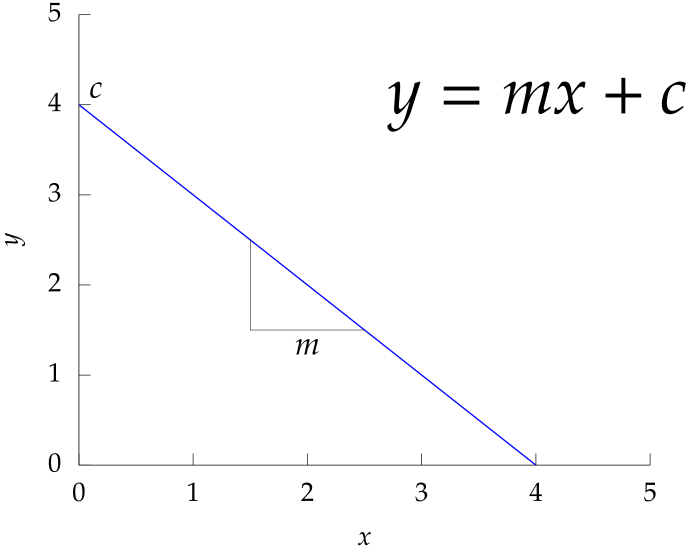
$$\text{data} \quad + \quad \text{model} \quad = \quad \text{prediction}$$

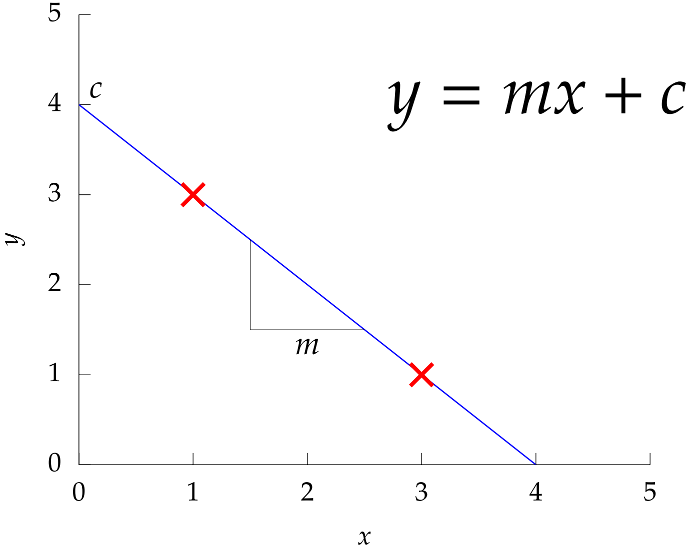
- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.
- prediction: an action to be taken or a categorization or a quality score.

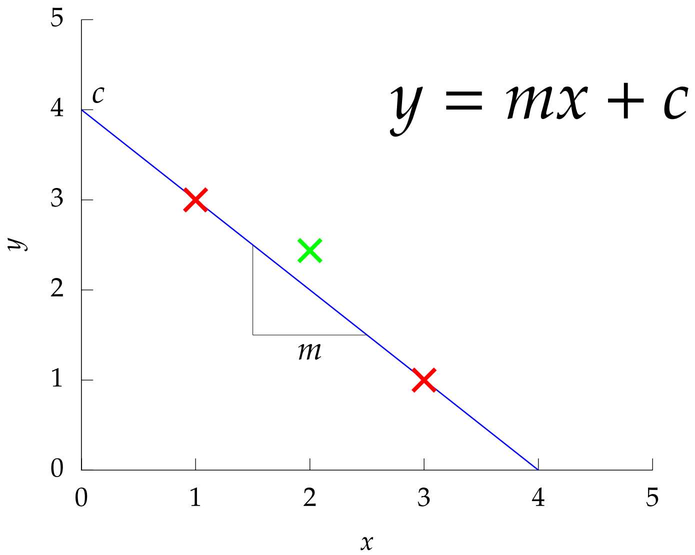
$$y = mx + c$$



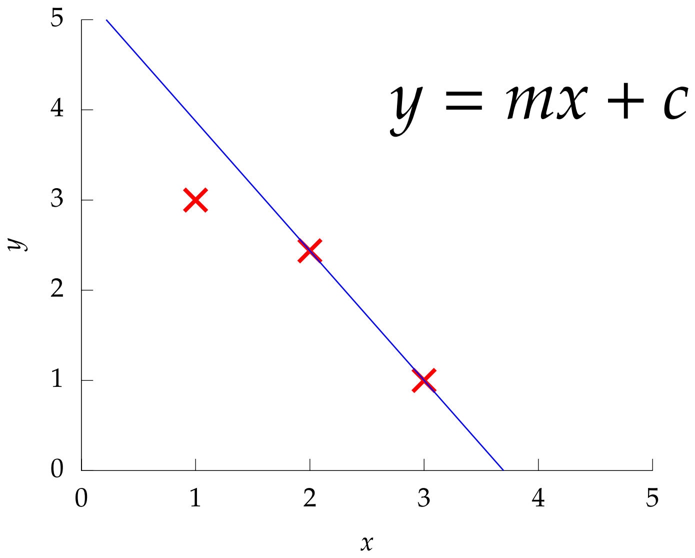
$y = mx + c$
5
4
3
2
1
0
y
0
1
2
3
4
5
x


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

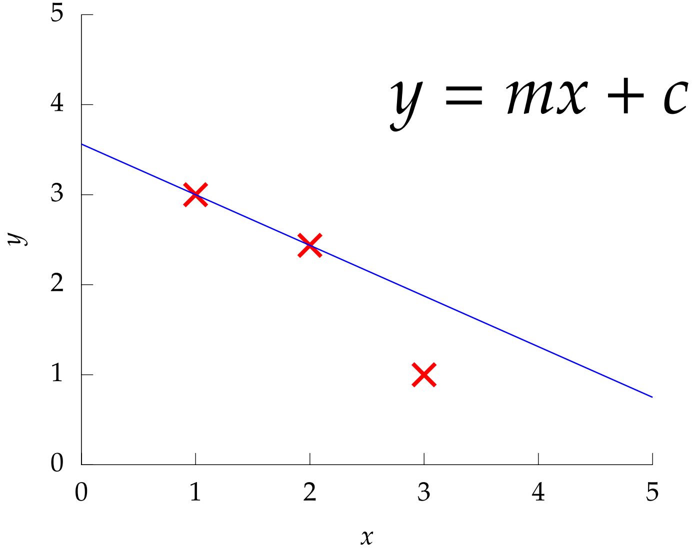

$y = mx + c$
c
m
0
1
2
3
4
5
0
1
2
3
4
5
x
y

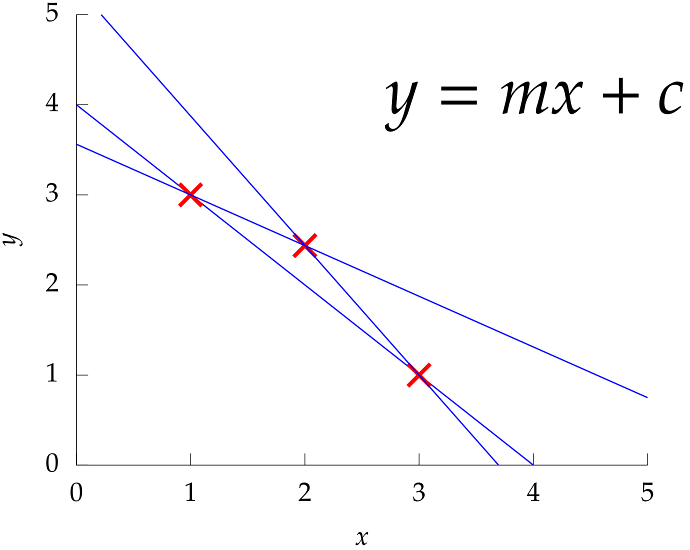


$y = mx + c$
c
m
x
y
0
1
2
3
4
5



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



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

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: $x = 2, y = 2.5$

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

Outline

Continuous Variables

- For continuous models we use the *probability density function* (PDF).
- Discrete case: defined probability distributions over a discrete number of states.
- How do we represent continuous as probability?
- Student heights:
  - Develop a representation which could answer *any* question we chose to ask about a student's height.
- PDF is a positive function, integral over the region of interest is one[1].

Manipulating PDFs

- Same rules for PDFs as distributions *e.g.*

$$p\left(y|x\right) = \frac{p\left(x|y\right)p\left(y\right)}{p\left(x\right)}$$

  where $p\left(x,y\right) = p\left(x|y\right)p\left(y\right)$ and for continuous variables $p\left(x\right) = \int p\left(x,y\right)\mathrm{d}y$.

- Expectations under a PDF

$$\left\langle f\left(x\right)\right\rangle_{p(x)} = \int f\left(x\right)p\left(x\right)\mathrm{d}x$$

  where the integral is over the region for which our PDF for x is defined.

The Gaussian Density

- Perhaps the most common probability density.

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$
$$= \mathcal{N}\left(y|\mu, \sigma^2\right)$$

- Also available in multivariate form.
- First proposed maybe by de Moivre but also used by Laplace.

Gaussian PDF I

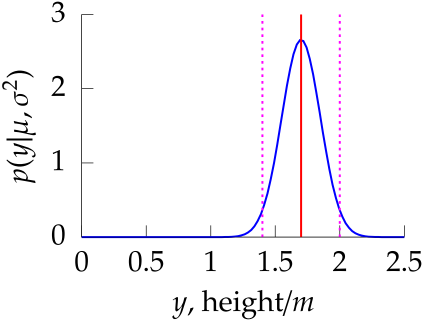


Figure: The Gaussian PDF with $\mu = 1.7$ and variance $\sigma^2 = 0.0225$. Mean shown as red line. Two standard deviations are shown as magenta. It could represent the heights of a population of students.

Cumulative Distribution Functions

- PDF doesn't represent probabilities directly
- One very common question is: what is the probability that $x < y$?
- The cumulative distribution function (CDF) represents the answer for $-\infty < x < \infty$ the CDF is given by

$$P\left(x > y\right) = \int_{-\infty}^{y} p\left(x\right) \mathrm{d}x,$$

for $0 \leq x < \infty$ then the CDF is given by

$$P\left(x > y\right) = \int_{0}^{y} p\left(x\right) \mathrm{d}x.$$

Gaussian PDF and CDF

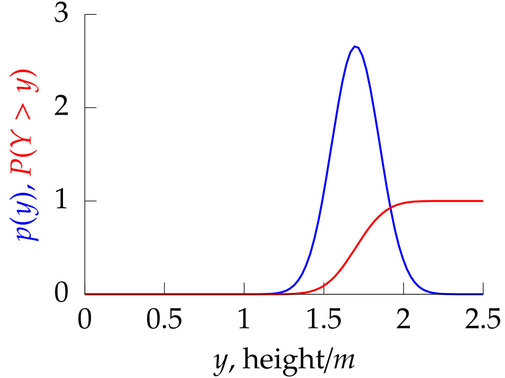


Figure: The cumulative distribution function (CDF) for the heights of computer science students. The thick curve gives the CDF and the thinner curve the associated PDF.

PDF from CDF

- The PDF can be recovered from the CDF through differentiation.

Outline

Sample Based Approximations I

- It is not always possible to compute expectations directly.
- Sample based approximation

$$\left\langle f\left(y\right)\right\rangle_{P(y)} \approx \frac{1}{N}\sum_{i=1}^{N} f\left(y_i\right).$$

- Special cases of this include the *sample mean*, often denoted by $\bar{y}$, and computed as

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

Sample Mean vs True Mean

- This is an approximation to the true distribution mean
$$\langle y \rangle \approx \bar{y}.$$
- The same approximations can used for continuous PDFs, so we have
$$\begin{aligned}\langle f(y) \rangle_{p(y)} &= \int f(y)\, p(y)\, \mathrm{d}y \\ &\approx \frac{1}{N} \sum_{i=1}^{N} f(y_i),\end{aligned}$$
where y_i are independently obtained samples from the density $p(y)$.
- Approximation gets better for increasing N and worse if the samples from $P(y)$ are *not* independent.

Outline

Entropy

- A particular expectation: $-\langle \log P(y) \rangle_{P(y)}$.
- This special expectation is known as the entropy of a distribution.
- It is a measure of how much “uncertainty” is in a distribution (learn it!).

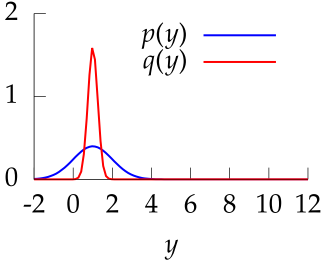
$$\mathcal{H}(y) = -\sum_{y} P(y) \log P(y)$$

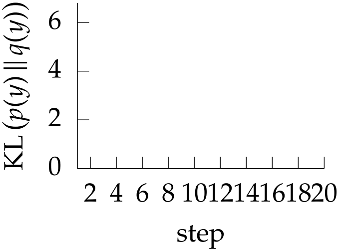
Kullback Leibler Divergence

- The Kullback Leibler divergence is another special expectation (learn it!).

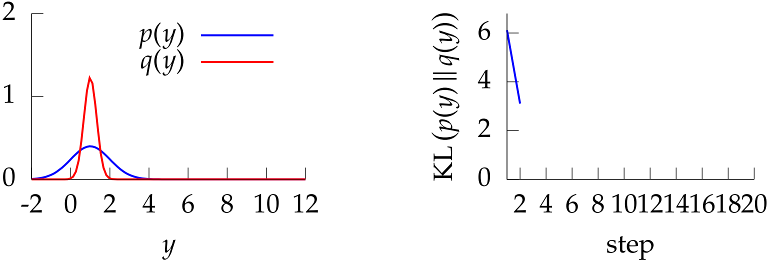
$$\mathrm{KL}\left(P(y)\,\|\,Q(y)\right) = \left\langle \log \frac{P(y)}{Q(y)} \right\rangle_{P(y)} = \left\langle \log P(y) \right\rangle_{P(y)} - \left\langle \log Q(y) \right\rangle_{P(y)}$$

- It is a measure of divergence between two distributions $Q(y)$ and $P(y)$.
- It is zero if they are identical (this is obviously true).
- It is positive if they are different (this is less obvious).

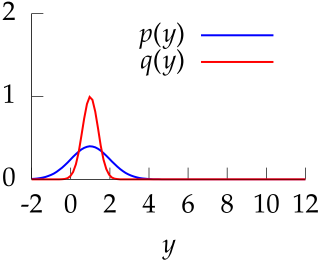


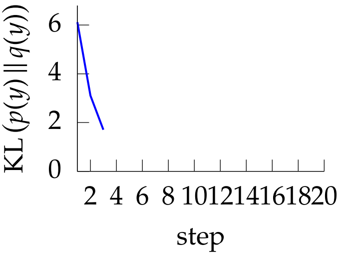


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

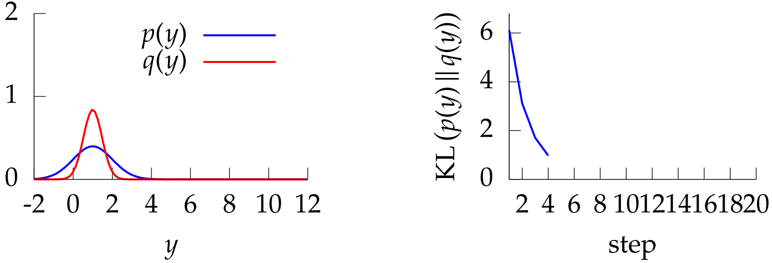


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

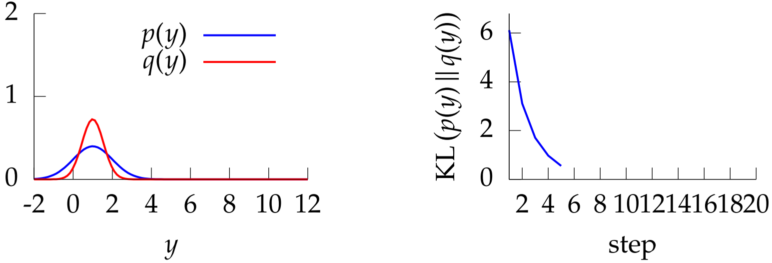




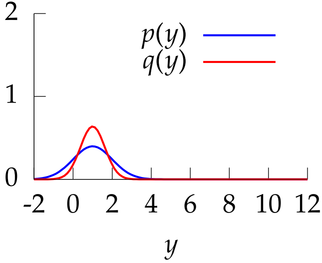
As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

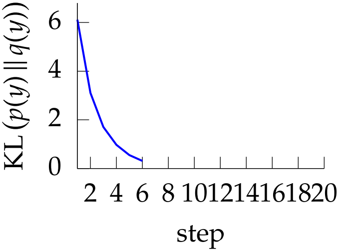


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

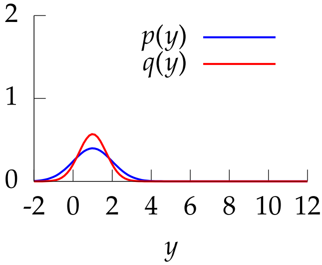


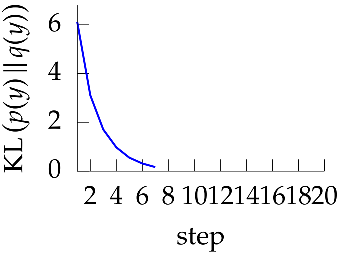
As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



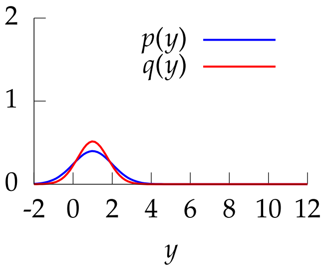


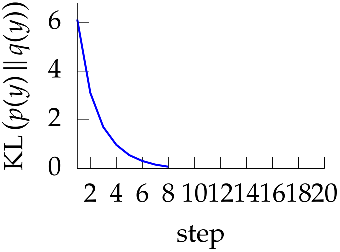
As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



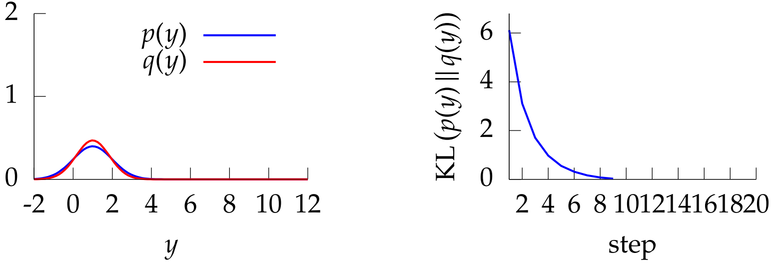


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

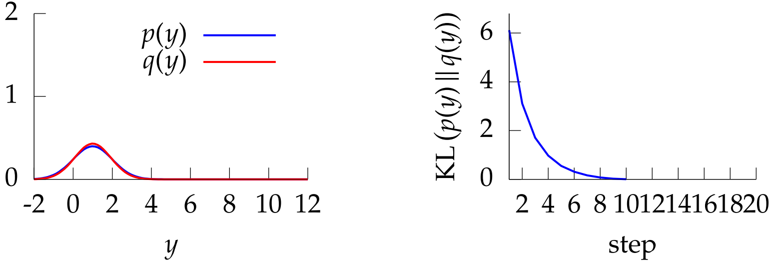




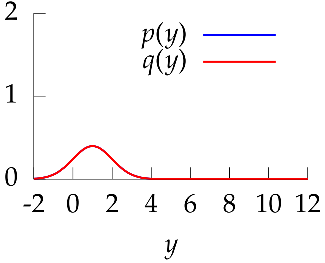
As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

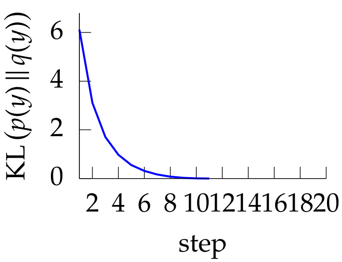


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

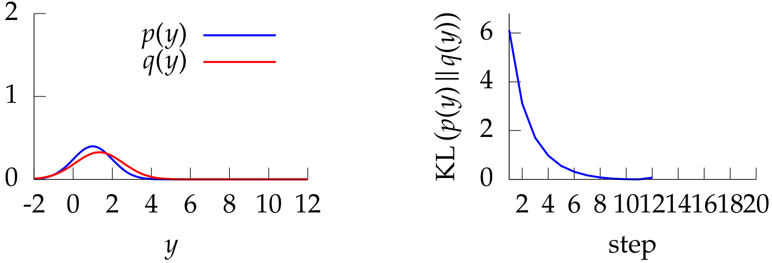


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

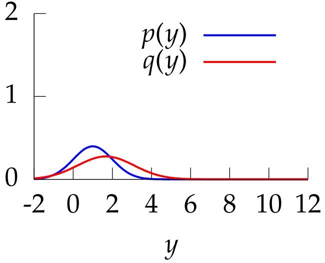


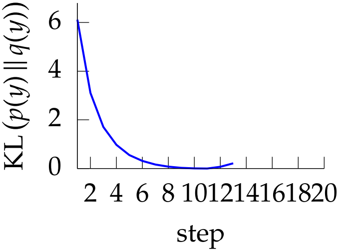


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

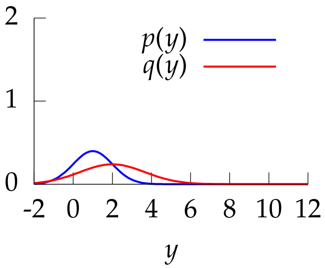


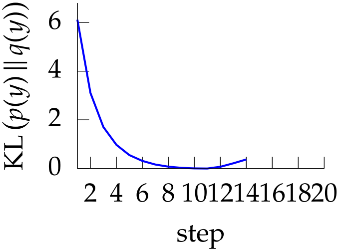
As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



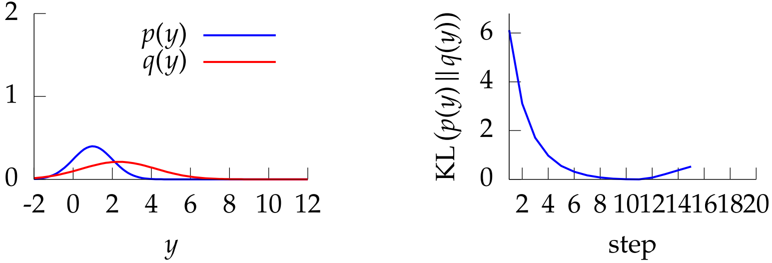


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

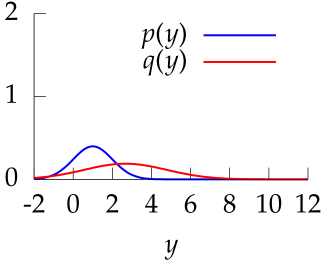


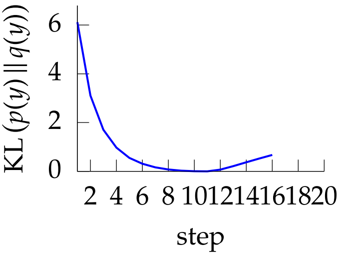


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

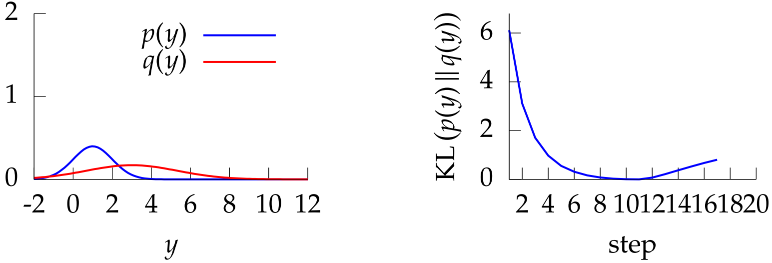


As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

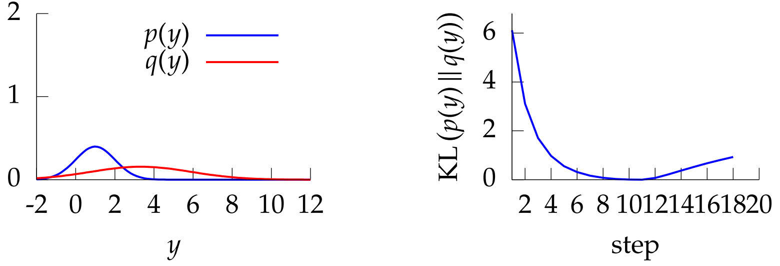




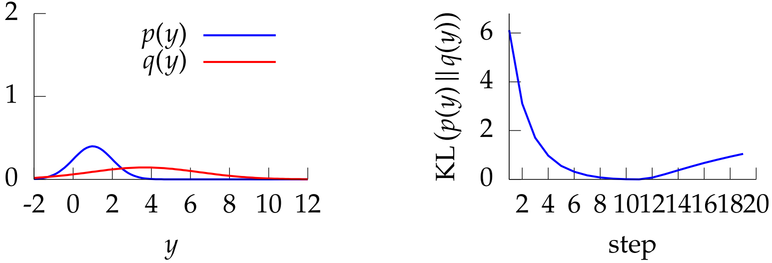
As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



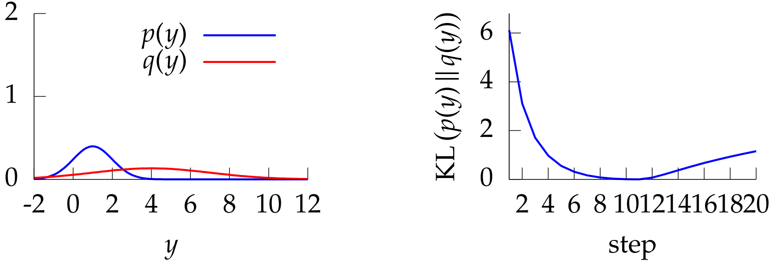
As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.



As the red Gaussian density ($q(y)$) approaches the blue Gaussian density ($p(y)$) the KL divergence approaches zero. As they move apart, KL divergence increases again.

Matching Two Distributions

- To match two distributions $P(y)$ and $Q(y)$ we can *minimize* the KL divergence.
- If we know the form of $Q(y)$ (our approximation) and it has parameters like a and b for the Gamma or mean and variance for Gaussian, we can change these parameters to find the best fit of $Q(y)$ to $P(y)$.
- If we have only got *samples* from $P(y)$ we use a sample based approximation.

Sample Based Approximation to the KL

$$\mathrm{KL}\left(P(y) \,\|\, Q(y)\right) \approx \frac{1}{N}\sum_{i=1}^{N} \log P(y_i) - \frac{1}{N}\sum_{i=1}^{N} \log Q(y_i)$$

- *Can't* compute the first term, but it *doesn't* depend on $Q(y)$ anyway.
- *Can* compute the second term. It is known as the negative log likelihood.

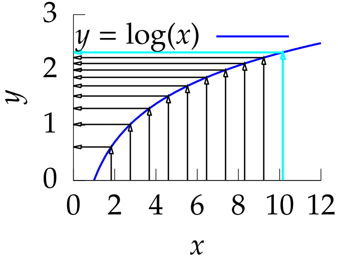
Maximum Likelihood

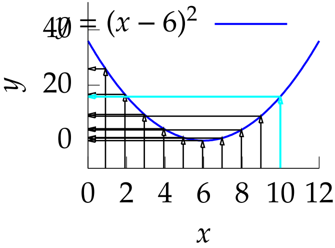
- Minimizing sample based KL divergence is equivalent to maximum likelihood (ML).
- The likelihood is defined as

$$P(\mathbf{y}|\boldsymbol{\theta})$$

  where $\mathbf{y}$ is a vector containing the data and $\boldsymbol{\theta}$ is a vector of parameters. i.e. this is the probability of the data given the parameters.
- Maximizing log likelihood is equivalent to maximizing likelihood because log is a *monotonic* function.

Monotonicity and Ordering





Monotonic functions preserve the ordering of input points, so the largest x is also the largest y. *Left*: gives an impression of this idea, cyan arrow is largest in x and correspondingly the largest in y. This transformation is log. *Right*: this quadratic function doesn't preserve the ordering and the largest x (again cyan arrow) is not the largest y value.

Sample Based Approximation implies i.i.d

- The log likelihood is
$$L(\boldsymbol{\theta}) = \log P(\mathbf{y}|\boldsymbol{\theta})$$
- If the likelihood is independent over the individual data points,
$$P(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} P(y_i|\boldsymbol{\theta})$$
- This is equivalent to the assumption that the data is *independent* and *identically* distributed. This is known as *i.i.d.*.
- Now the log likelihood is
$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(y_i|\boldsymbol{\theta})$$
which matches the sample based KL approximation up to a scaling by $-N$.

Maximum Likelihood Properties

Properties of ML arise due to the relationship with the KL divergence, and law of large numbers.

- As $N \to \infty$ *If* class of distributions considered for $Q(y)$ contains $P(y)$ then we will obtain $Q(y) = P(y)$.
- This is known as the consistency of maximum likelihood.
- In practice
  - We won't have infinite data.
  - We cannot prove that $Q(y)$ will include $P(y)$.

Maximum Likelihood, Minimum Error

- To maximize likelihood we use optimization techniques.
- In the optimization community *minimization* is the convention.
- Define the "error function" to be negative log likelihood.

$$E(\boldsymbol{\theta}) = -\log L(\boldsymbol{\theta})$$

- $E(\cdot)$ can also be thought of as an *energy* function. This is a physics interpretation.

Basic Optimization Overview

- To find a minimum, want to find a point where gradient is zero (this is a stationary point).
- If we can show that curvature is positive, this is a minimum.
- Procedure: differentiate the function, find parameters which set derivative to zero.
- This can sometimes be done by a fixed point equation, other times iterative optimization methods are required.

Example: Maximum Likelihood in the Gaussian

$$P(\mathbf{y}|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$$

1. Write down error function.

Example: Maximum Likelihood in the Gaussian

$$P(\mathbf{y}|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$$

1. Write down error function.
2. Differentiate error function.

Example: Maximum Likelihood in the Gaussian

$$P(\mathbf{y}|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right)$$

1. Write down error function.
2. Differentiate error function.
3. Solve such that the derivatives are zero.

Reading

- See probability review at end of slides for reminders.
- Read and *understand* Rogers and Girolami 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 Bishop 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 Bishop (pg 48 onwards).
- If you are unfamiliar with probabilities you should complete the following exercises:
  1. Bishop Exercise 1.7
  2. Bishop Exercise 1.8
  3. Bishop Exercise 1.9

References I

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

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

Outline

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	$P(X = x \mid Y = y)$	'The probability that $X = x$ *given that* $Y = y$'

Table: The different basic probability distributions.

A Pictorial Definition of Probability

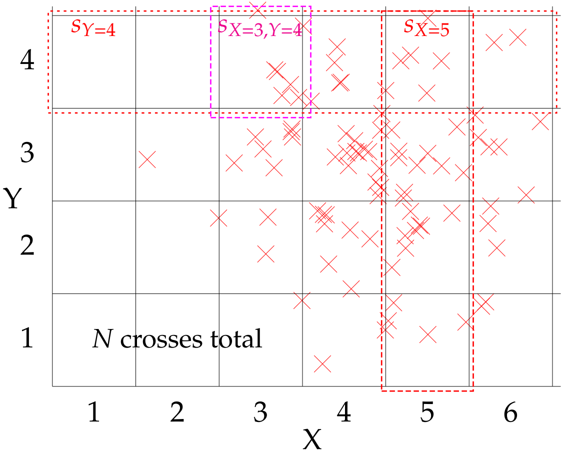


Figure: Representation of joint and conditional probabilities.

Different Distributions

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

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 $\sum_x s_x = N$, which gives

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

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

$$\frac{s_y}{N} = \sum_x \frac{s_{x,y}}{N},$$

  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
$$\frac{s_{x,y}}{s_y}.$$
- $P(x,y)$ is
$$\frac{s_{x,y}}{N} = \frac{s_{x,y}}{s_y}\frac{s_y}{N}$$
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:

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

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

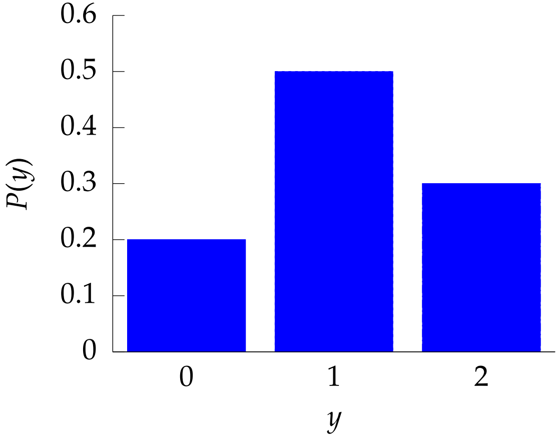


Figure: Histogram representation of the simple distribution.

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.

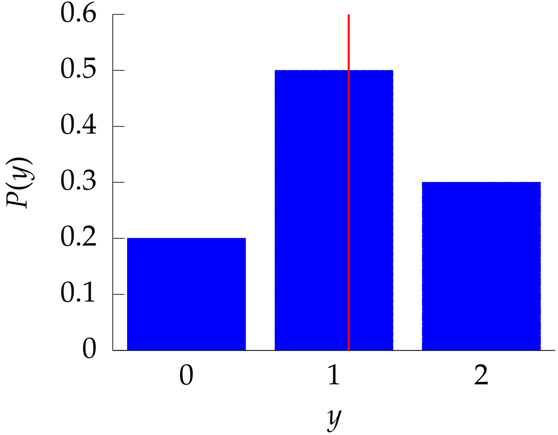


Figure: Histogram representation of the simple distribution including the expectation of y (red line), the mean of the distribution.

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.

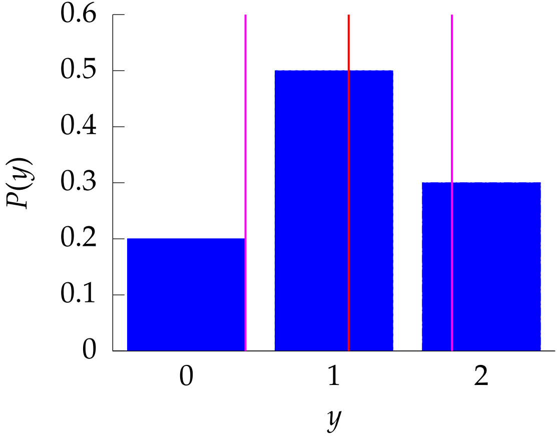


Figure: Histogram representation of the simple distribution including lines at one standard deviation from the mean of the distribution (magenta lines).

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.