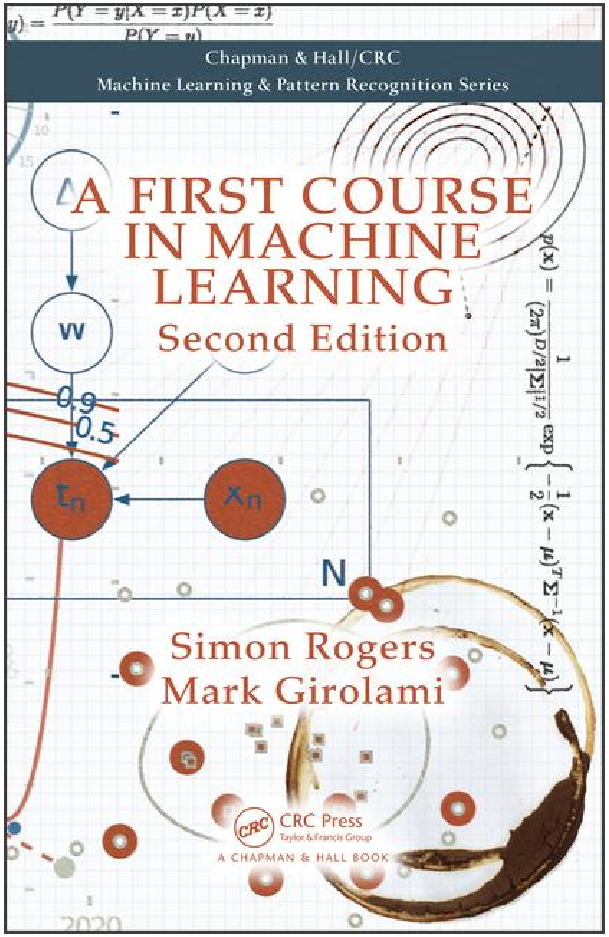
Introduction to Machine learning and Probability

Machine Learning and Adaptive Intelligence

Mauricio Álvarez

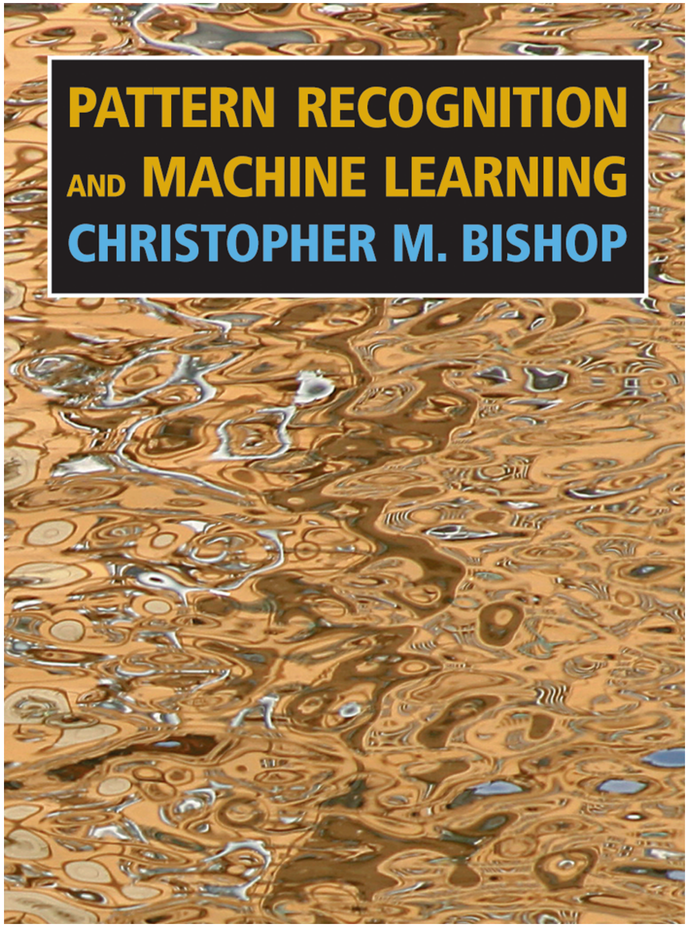
Based on slides by Neil D. Lawrence

Course Text



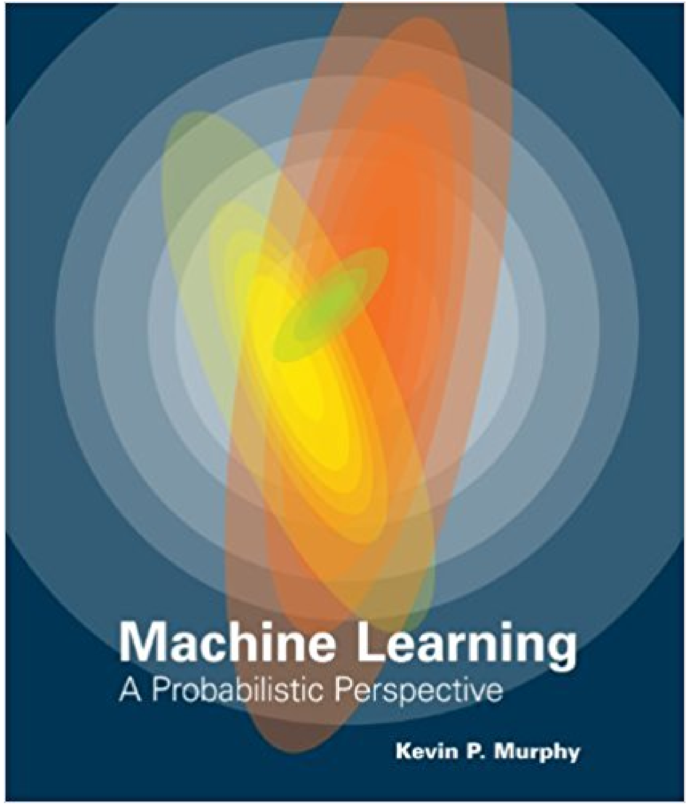
Simon Rogers and Mark Girolami, *A First Course in Machine Learning*, Chapman and Hall/CRC Press, 2nd Edition, 2016.

Course Text



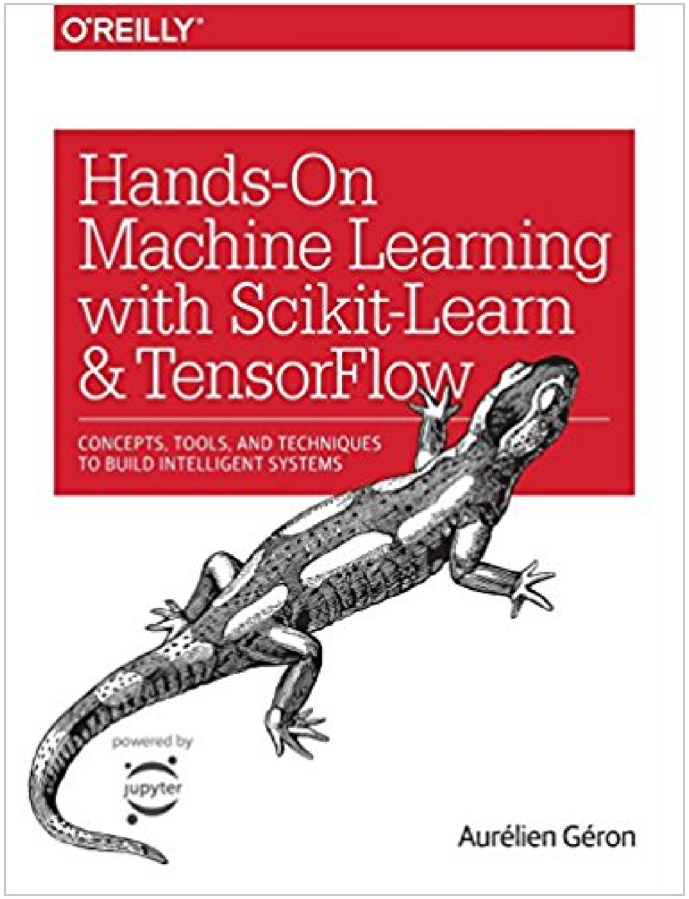
Christopher Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006.

Course Text



Kevin Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

Course Text



Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, O'Reilly, 2017.

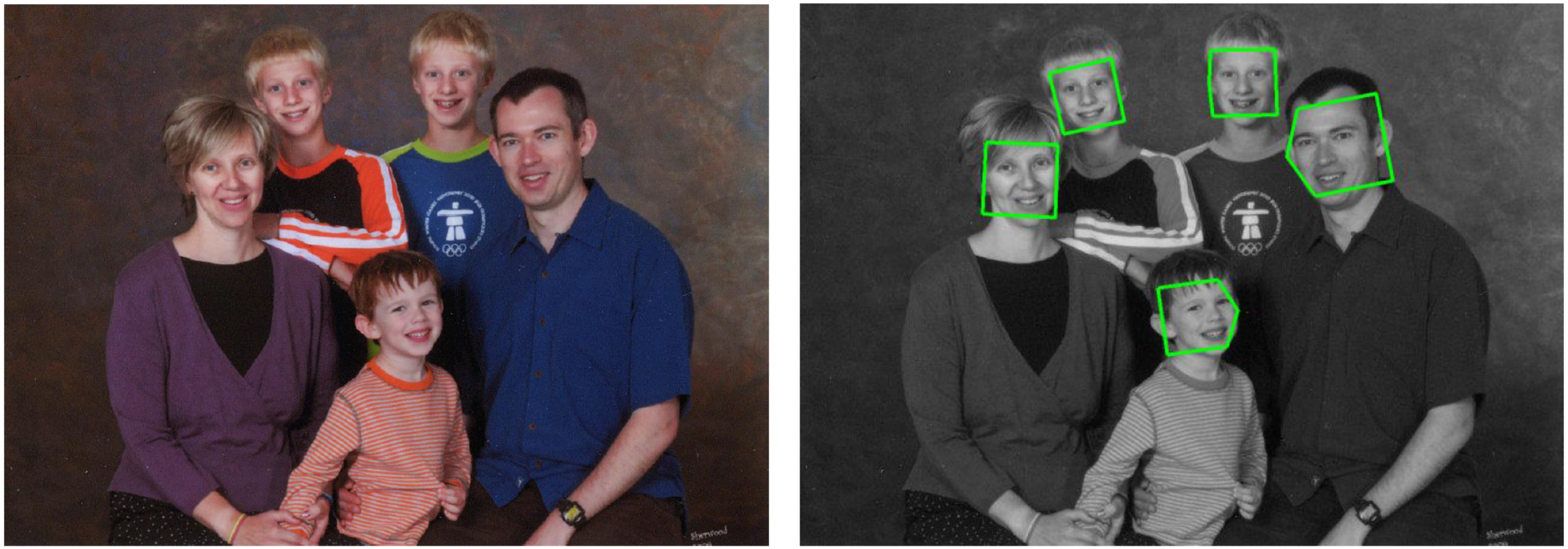
There is barely Deep Learning in this module

Machine Learning or Statistical Learning

- We would like to come up with a **model** that help us to solve a **prediction** problem.
- The model is built using a **dataset**.
- The ultimate goal is to extract knowlegde from data.

Handwritten digit recognition

Face detection and face recognition



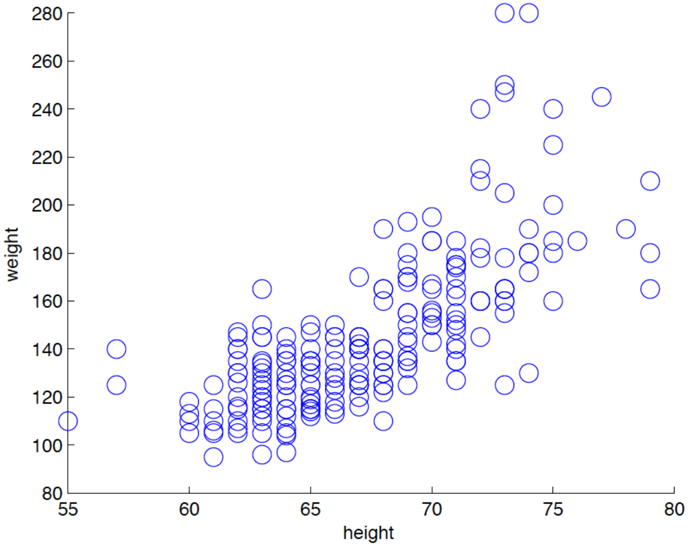
Taken from Murphy (2012).

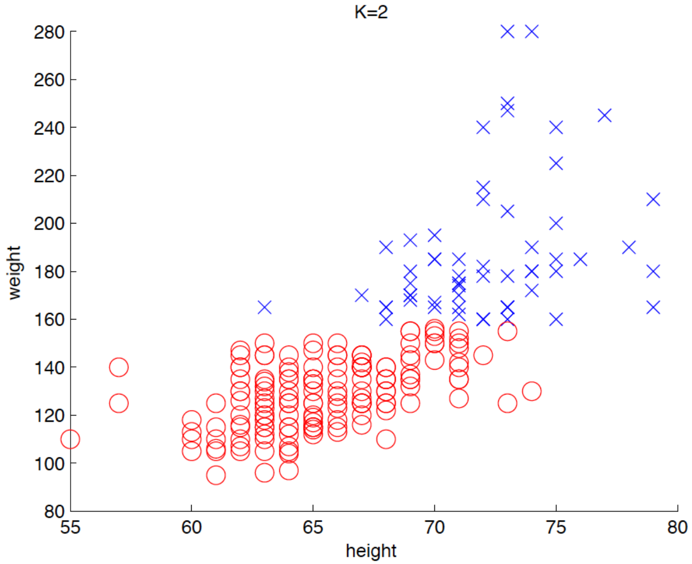
Predicting the age of a person

Predicting the age of a person looking at a particular YouTube video

Stock market

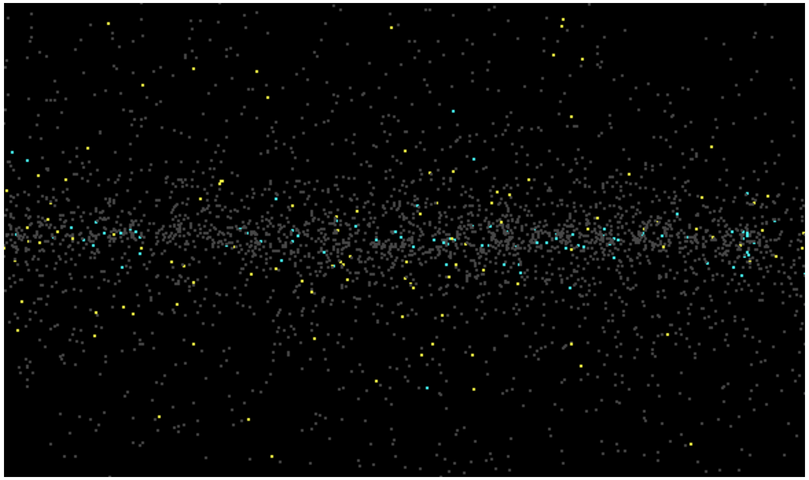
Clustering



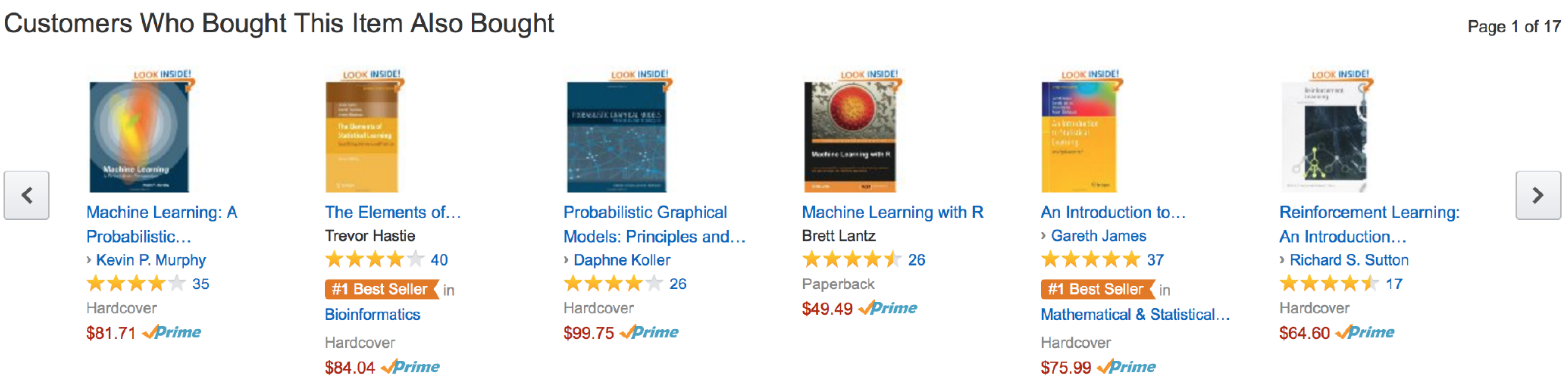


Taken from Murphy (2012).

Autoclass



Recommendation systems



Basic definitions

- Handwritten digit recognition

- Variability.

- Each image can be transformed into a vector $\mathbf{x}$ (feature extraction).

- An instance is made of a the pair $(\mathbf{x}, y)$, where y is the label of the image.

- Objective: find a function $f(\mathbf{x}, \mathbf{w})$ that allows predictions.

Basic definitions

- **Training set**: a set of N images and their labels, $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$, to fit the predictive model.
- **Estimation or training phase**: process of getting the values of $\mathbf{w}$ of the function $f(\mathbf{x}, \mathbf{w})$ that best fit the data.
- **Generalisation**: ability to correctly predict the label of new images $\mathbf{x}_*$.

Supervised and unsupervised learning

- Supervised learning:
  - Variable y is discrete: *classification*.
  - Variable y is continuous: *regression*.

- Unsupervised learning: from the set of instances $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ we only have access to $\mathbf{x}_1, \ldots, \mathbf{x}_N$.
  - Find similar groups: *clustering*.
  - Find a probability function for $\mathbf{x}$: *density estimation*.
  - Find a lower dimensionality representation for $\mathbf{x}$: *dimensionality reduction and feature selection*.

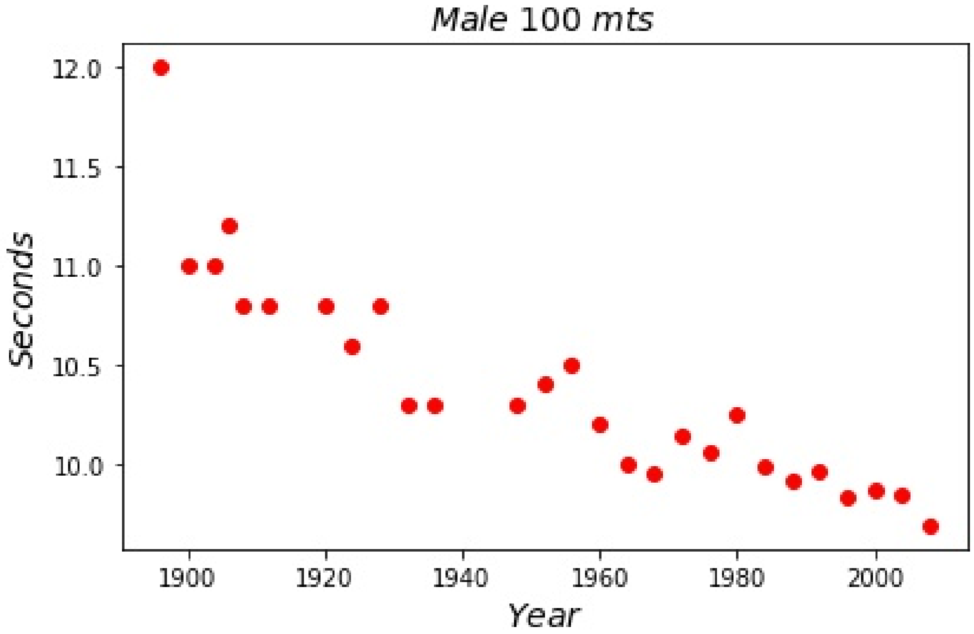
- Other types of learning: semi-supervised learning, active learning, multi-task learning.

Example: Olympic 100m Data

- Gold medal times for Olympic 100 m runners since 1896.

Image from Wikimedia Commons http://bit.ly/191adDC (http://bit.ly/191adDC)

Example: Olympic 100m Data



Model

- We will use a linear model $y = f(x)$, where y is the time in seconds and x the year of the competition.

- The linear model is given as

$$y = w_1 x + w_0,$$

where w_0 is the intercept and w_1 is the slope.

Objective function

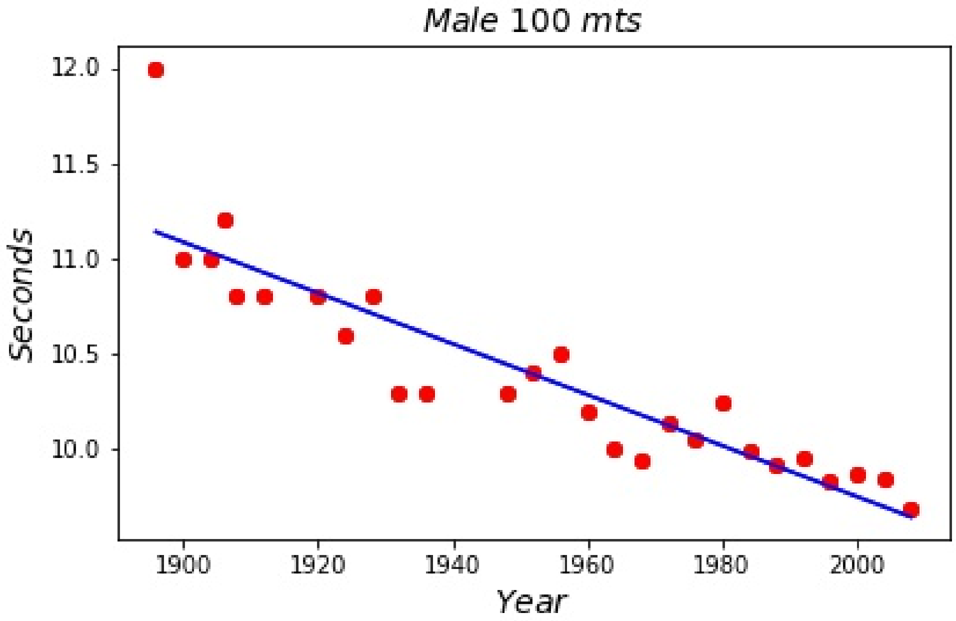
- We use an objetive function to estimate the parameters w_0 and w_1 that best fit the data.

- In this example, we use a a least squares objective function

$$E(w_0, w_1) = \sum_{\forall\ i} (y_i - f(x_i))^2 = \sum_{\forall\ i} [y_i - (w_1 x_i + w_0)]^2.$$

- Minimising the error we get the solution as $w_0 = 36.4$ and $w_1 = -1.34 \times 10^{-2}$.

Data and model



Predictions

- We can now use this model for making predictions.
- For example, what does the model predict for 2012?
- If we say $x = 2012$, then

$$y = f(x) = f(2012) = w_1 x + w_0 = (-1.34 \times 10^{-2}) \times 2012 + 36.4 = 9.59.$$

- The actual value was 9.63.

Probability Review

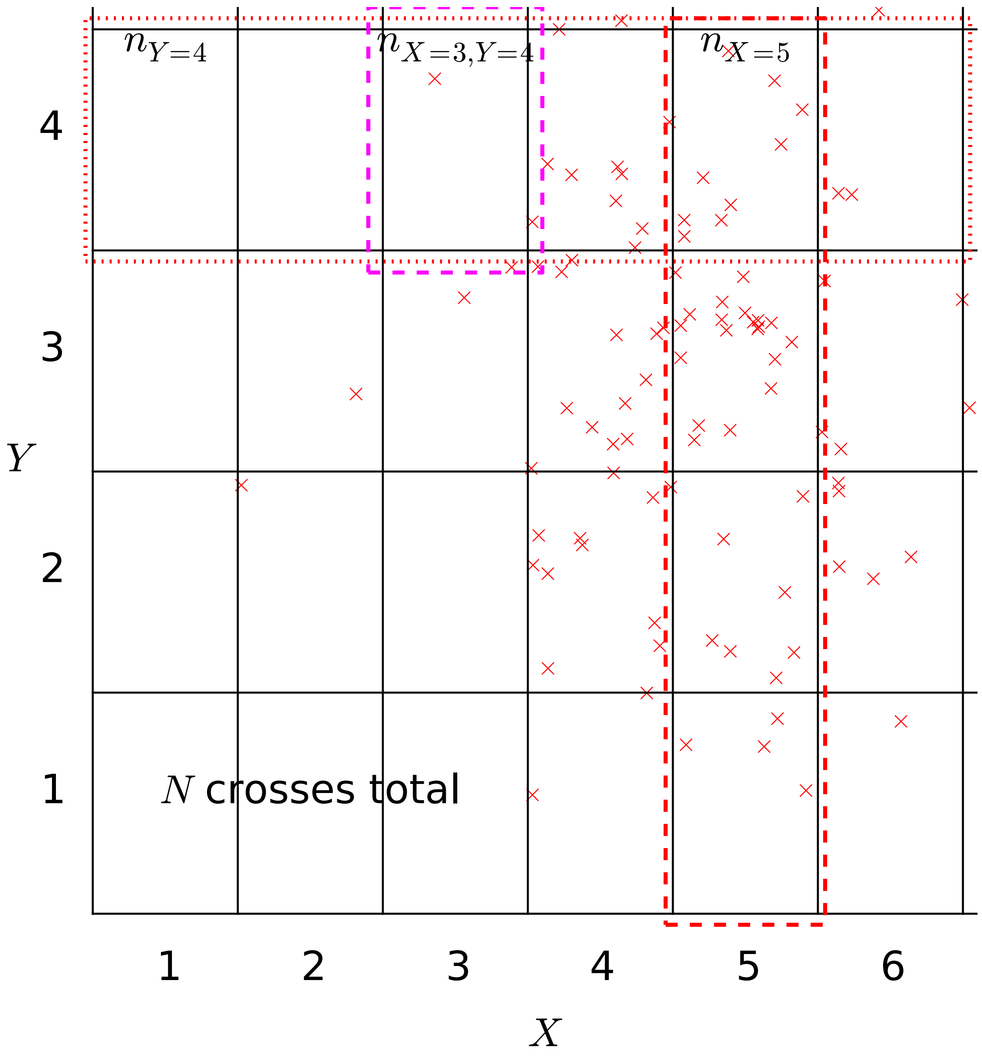
- We are interested in trials which result in two random variables, X and Y, each of which has an 'outcome'denoted by x or y.

- We summarise the notation and terminology for these distributions in the following table.

Terminology	**Mathematical notation**	**Description**
joint	$P(X = x, Y = y)$	probability that X=x *and* Y=y
marginal	$P(X = x)$	probability that X=x *regardless of* Y
conditional	$P(X = x \mid Y = y)$	probability that X=x *given that* Y=y

The different basic probability distributions.

A Pictorial Definition of Probability



Different Distributions

- Definition of probability distributions.

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

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

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

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 $\lim_{N\rightarrow\infty} \frac{n_y}{N}$.
- The joint distribution $P(x, y)$ is $\lim_{N\rightarrow\infty} \frac{n_{x,y}}{N}$.

- $n_y = \sum_x n_{x,y}$ so

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

  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
$$\lim_{N\rightarrow\infty} \frac{n_{x,y}}{n_y}.$$

- $P(x, y)$ is
$$\lim_{N\rightarrow\infty} \frac{n_{x,y}}{N} = \lim_{N\rightarrow\infty} \frac{n_{x,y}}{n_y} \frac{n_y}{N}$$

  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,

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

Notice that the expression above works as long as $P(x) \neq 0$.

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:

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

- 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 next week, 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.

Expectation Computation Example

- Consider the following distribution.

$\mathbf{y}$	**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: $\mu = \mathbb{E}[y] = \sum_y yP(y) = 1 \times 0.3 + 2 \times 0.2 + 3 \times 0.1 + 4 \times 0.4 = 2.6$

- Second moment: $\mathbb{E}[y^2] = \sum_y y^2 P(y) = 1 \times 0.3 + 4 \times 0.2 + 9 \times 0.1 + 16 \times 0.4 = 8.4$

- Variance: $\sigma_y^2 = \mathbb{E}[y^2] - \mu^2 = 8.4 - 2.6 \times 2.6 = 1.64$

- Standard deviation: $\sqrt{1.64} = 1.2806$

- Expectation of $-\log(P(y))$: $\mathbb{E}[-\log(P(y))] = \sum_y -\log(P(y))P(y) = 0.3 \times 1.204 + 0.2 \times 1.609 + 0.1 \times 2.302 + 0.4 \times 0.916 = 1.280$

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?
- 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: $\hat{\mu} = \frac{1}{N} \sum_{\forall i} y_i = \frac{1.76+1.73+1.79+1.81+1.85+1.80}{6} = 1.79$

- Second moment:
$\hat{\mathbf{E}}[y^2] = \frac{1}{N} \sum_{\forall i} y_i^2 = \frac{3.0976+2.9929+3.2041+3.2761+3.4225+3.2400}{6} = 3.2055$

- Variance: $\hat{\sigma}_y^2 = \hat{\mathbf{E}}[y^2] - \hat{\mu}^2 = 3.2055 - 1.79 \times 1.79 = 1.43 \times 10^{-3}$

- Standard deviation: $\sqrt{\hat{\sigma}_y^2} = 0.0379$

- 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* Rogers and Girolami (2016) on:

  1. Section 2.2 (Random Variables and Probability).
  2. Section 2.4 (Continuous Random Variables - Density functions).
  3. Section 2.5.1 (the Uniform density function).
  4. Section 2.5.3 (the Gaussian density function).

- For other material in Bishop (2006) 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 of Bishop (2006) (pg 48 onwards).

- If you are unfamiliar with probabilities you should complete exercises 1.7, 1.8, 1.9 in Bishop (2006).