

# Introduction to Machine Learning

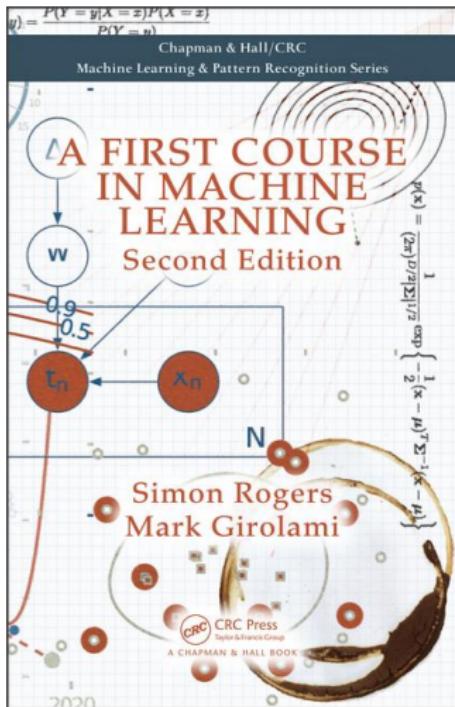
Mauricio A. Álvarez

Machine Learning and Adaptive Intelligence  
The University of Sheffield

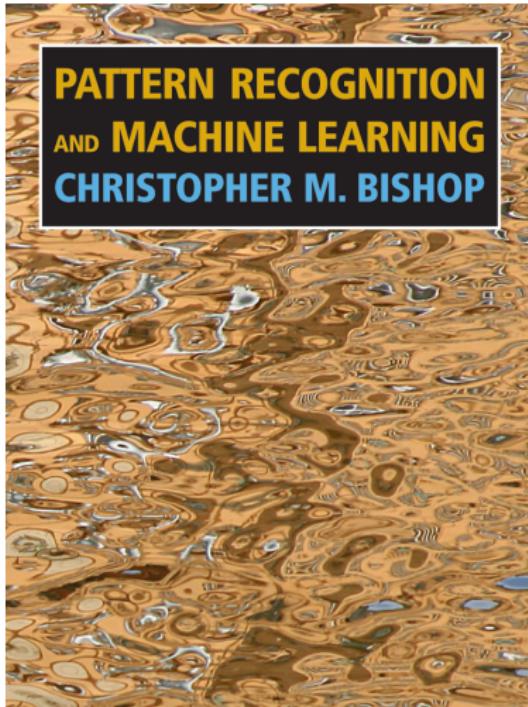


The  
University  
Of  
Sheffield.

# Textbooks

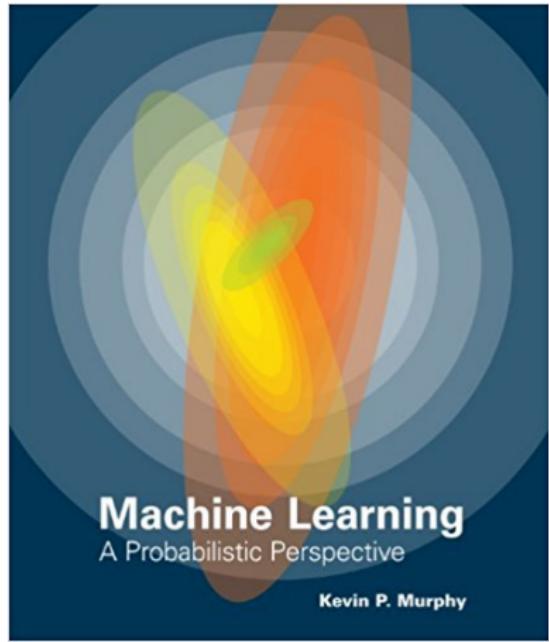


Rogers and Girolami, *A First Course in Machine Learning*, Chapman and Hall/CRC Press, 2nd Edition, 2016.

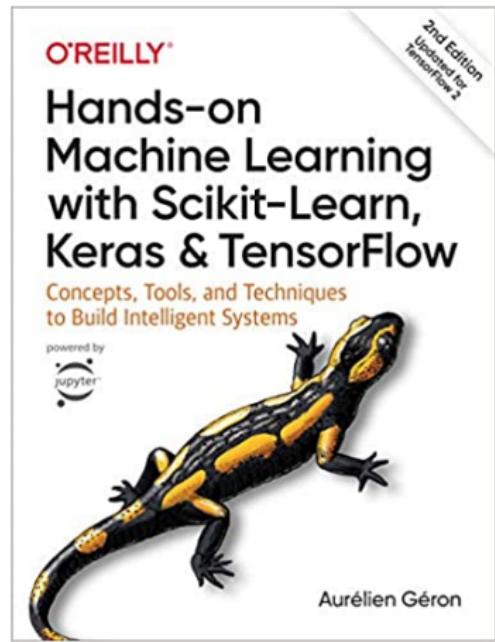


Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, 2006.

# Textbooks



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



Géron, *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow*, O'Reilly, 2nd Edition, 2019.

# Contents

Machine learning

Definitions

An example of a predictive model

Review of probability

Random variables

Discrete random variables

Continuous random variables

Additional comments

# Machine learning or Statistical Learning

- We would like to design an algorithm that help us to solve different prediction problems.
- The algorithm is designed based on a mathematical model or function, and a dataset.
- Extract knowlegde from data.

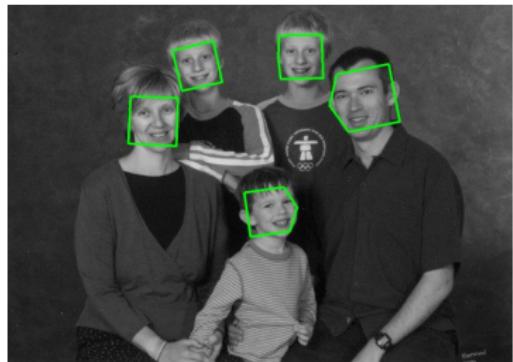
# Examples of prediction problems

Handwritten digit recognition



# Examples of prediction problems

Face detection and face recognition



From Murphy (2012).

## Examples of prediction problems

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



# Examples of prediction problems

Stock market



# Examples of prediction problems

Clustering: segmenting customers in e-commerce



# Examples of prediction problems

## Recommendation systems

Customers Who Bought This Item Also Bought

Page 1 of 17

The screenshot shows a row of six book covers from an Amazon search results page. Each book has a 'Look Inside' button at the top right. The books are:

- Machine Learning: A Probabilistic... by Kevin P. Murphy (4.5 stars, 35 reviews, \$81.71 Prime)
- The Elements of... by Trevor Hastie (4.5 stars, 40 reviews, #1 Best Seller in Bioinformatics)
- Probabilistic Graphical Models: Principles and... by Daphne Koller (4.5 stars, 26 reviews, Hardcover, \$99.75 Prime)
- Machine Learning with R by Brett Lantz (4.5 stars, 26 reviews, Paperback, \$49.49 Prime)
- An Introduction to... by Gareth James (4.5 stars, 37 reviews, #1 Best Seller in Mathematical & Statistical...)
- Reinforcement Learning: An Introduction... by Richard S. Sutton (4.5 stars, 17 reviews, Hardcover, \$64.60 Prime)

Navigation arrows are visible on the left and right sides of the row.



# Contents

Machine learning

Definitions

An example of a predictive model

Review of probability

Random variables

Discrete random variables

Continuous random variables

Additional comments

## Basic definitions

- Handwritten digit recognition



- Variability
- Each image can be transformed into a vector  $\mathbf{x}$  (feature extraction).
- An instance is made of the pair  $(\mathbf{x}, y)$ , where  $y$  is the label of the image.
- Objective: find a function  $f(\mathbf{x}, \mathbf{w})$ .

# Basic definitions

- **Training set:** a set of  $N$  images and their labels  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$ , 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  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$ , we **only** have access to  $\mathbf{x}_1, \dots, \mathbf{x}_N$ 
  - Find similar groups: *clustering*.
  - Find a probability function for  $\mathbf{x}$ : *density estimation*.
  - Find a lower dimensionality representation for  $\mathbf{x}$ : *dimensionality reduction and visualisation*.
- **Other types of learning:** reinforcement learning, semi-supervised learning, active learning, multi-task learning.

# Contents

Machine learning

Definitions

An example of a predictive model

Review of probability

Random variables

Discrete random variables

Continuous random variables

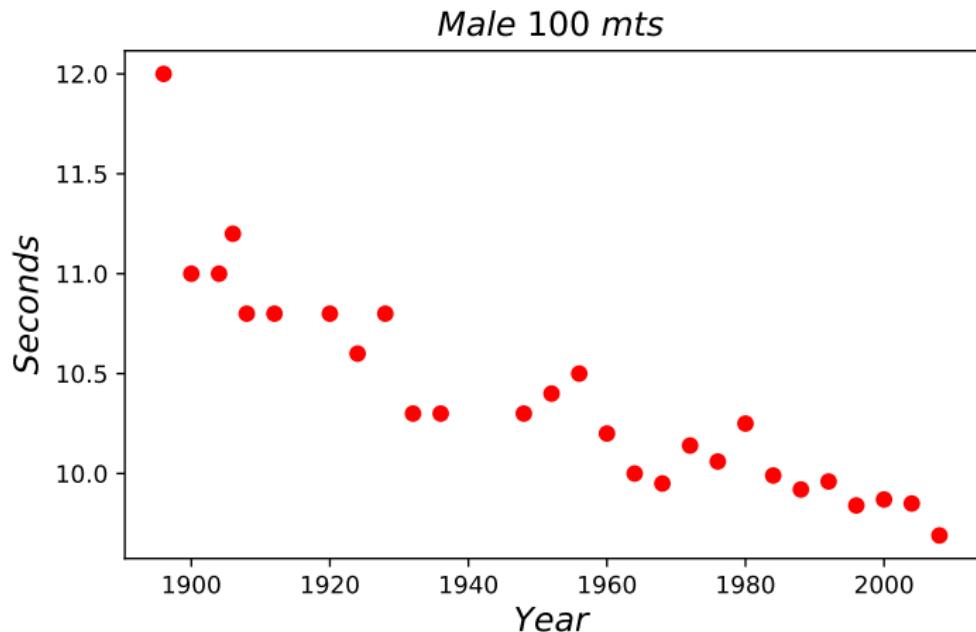
Additional comments

# Olympic 100m Data



Image from Wikimedia Commons <http://bit.ly/191adDC>.

# Dataset



# Model

- We will use a linear model  $y = f(x, \mathbf{w})$ , 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.

- We use  $\mathbf{w}$  to refer both to  $w_0$  and  $w_1$ .

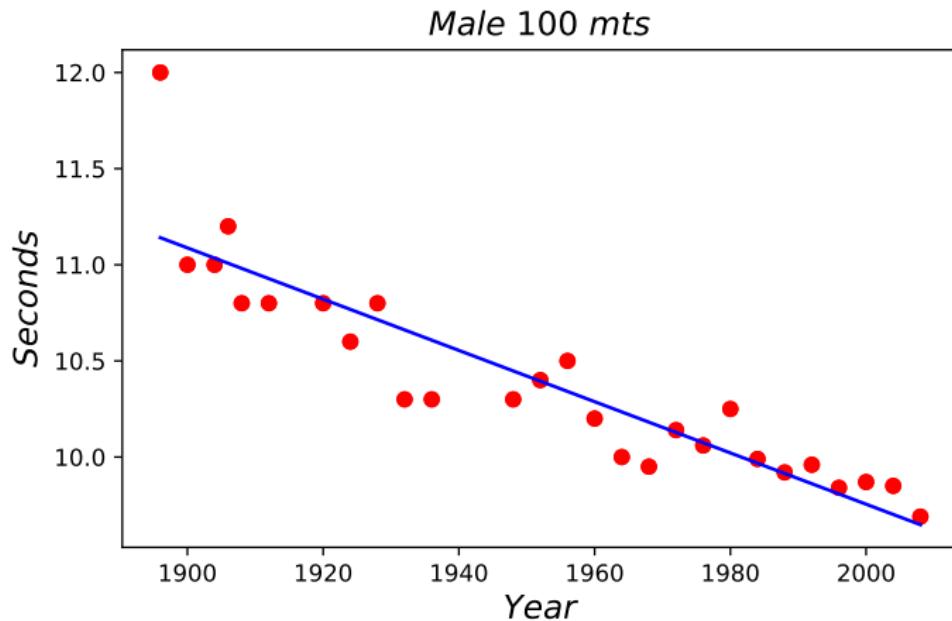
# Objective function

- We use an objective function to estimate the parameters  $w_0$  and  $w_1$  that best fit the data.
- In this example, we use a least squares objective function

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

- By minimising the error with respect to  $\mathbf{w}$ , 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

$$\begin{aligned}y &= f(x, \mathbf{w}) = f(x = 2012, \mathbf{w}) \\&= w_1 x + w_0 = (-1.34 \times 10^{-2}) \times 2012 + 36.4 = 9.59.\end{aligned}$$

- The actual value was 9.63.

# Main challenges of machine learning

- Insufficient quantity of training data.
- Nonrepresentative training data.
- Poor-quality data.
- Irrelevant features.
- Overfitting the training data.
- Underfitting the training data.

# Contents

Machine learning

Definitions

An example of a predictive model

Review of probability

Random variables

Discrete random variables

Continuous random variables

Additional comments

# Contents

Machine learning

Definitions

An example of a predictive model

**Review of probability**

**Random variables**

Discrete random variables

Continuous random variables

Additional comments

## Definition

- A *random variable* (RV) is a *function* that assigns a number to the outcome of a random experiment.
- For example, we toss a coin (random experiment).
- We assign the number 0 to the outcome “tails” and the number “1” to the outcome “heads”.

# Discrete and continuous random variables

- A random variable can either be discrete or continuous.
- A **discrete RV** can take a value only from a countable number of distinct values, like 1, 2, 3, ....
- For example, the number of phone calls received in a call-center from 9:00 to 10:00, the number of COVID patients in the Royal Hallamshire Hospital on May 30, 2020.
- A **continuous RV** can take any value from an infinite possible values within one or more intervals.
- Examples include the time that a cyclist takes to finish the Tour de France; the exchange rate between the british pound and the US dollar on June 30, 2020.

# Notation

- We use capital letters to denote random variables, e.g.  $X, Y, Z, \dots$
- We use lowercase letters to denote the values that the random variable takes,  $x, y, z, \dots$

# Contents

Machine learning

Definitions

An example of a predictive model

**Review of probability**

Random variables

**Discrete random variables**

Continuous random variables

Additional comments

# Probability mass function

- A discrete RV  $X$  is completely defined by a set of values it can take,  $x_1, x_2, \dots, x_n$ , and their corresponding probabilities.
- The probability that  $X = x_i$  is denoted as  $P(X = x_i)$  for  $i = 1, \dots, n$ , and it is known as the *probability mass function* (pmf).
- Properties
  1.  $P(X = x_i) \geq 0, \quad i = 1, \dots, n.$
  2.  $\sum_{i=1}^n P(X = x_i) = 1.$

## Two discrete RVs

- In machine learning, we are usually interested in more than one random variable.
- Consider two RVs  $X$  and  $Y$  taking values  $x_1, x_2, \dots, x_n$  and  $y_1, y_2, \dots, y_m$ , respectively.
- These two random variables can be fully described with a **joint probability mass function**  $P(X = x_i, Y = y_j)$  specifying the probability of  $X = x_i$  and  $Y = y_j$ .
- Properties
  1.  $P(X = x_i, Y = y_j) \geq 0, \quad i = 1, \dots, n, j = 1, \dots, m.$
  2.  $\sum_{i=1}^n \sum_{j=1}^m P(X = x_i, Y = y_j) = 1.$

# Rules of probability

- **Marginal**

$$P(X = x_i) = \sum_{j=1}^m P(X = x_i, Y = y_j).$$

We obtain the probability of  $P(X = x_i)$  regardless of the value of  $Y$ .  
This is also known as the **sum rule of probability**.

- **Conditional**

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}, \quad P(Y = y_j) \neq 0.$$

- From the above expression, we can also write

$$P(X = x_i, Y = y_j) = P(X = x_i | Y = y_j)P(Y = y_j).$$

This expression is also known as the **product rule of probability**.

## How do we compute $P(X = x_i)$ from data?

- A way to compute the probability  $P(X = x_i)$  is to repeat an experiment several times, say  $N$ , see how many outcomes we get for which  $X = x_i$ , say  $n_{X=x_i}$  and then approximate the probability as

$$P(X = x_i) \approx \frac{n_{X=x_i}}{N}.$$

- We expect the approximation to become an equality when  $N \rightarrow \infty$ ,

$$P(X = x_i) = \lim_{N \rightarrow \infty} \frac{n_{X=x_i}}{N}.$$

## What about $P(X = x_i, Y = y_j)$ and $P(X = x_i | Y = y_j)$ ?

- We can follow a similar procedure to compute  $P(X = x_i, Y = y_j)$ ,

$$P(X = x_i, Y = y_j) = \lim_{N \rightarrow \infty} \frac{n_{X=x_i, Y=y_j}}{N},$$

where  $n_{X=x_i, Y=y_j}$  is the number of times we observe a simultaneous occurrence of  $X = x_i$  and  $Y = y_j$ .

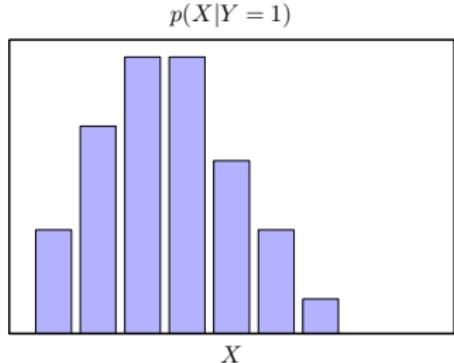
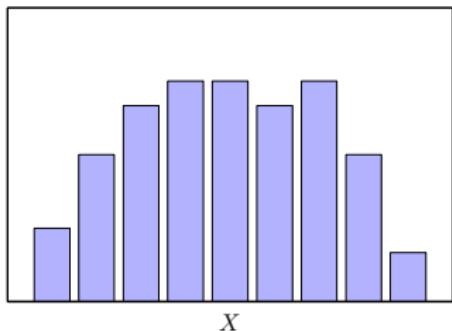
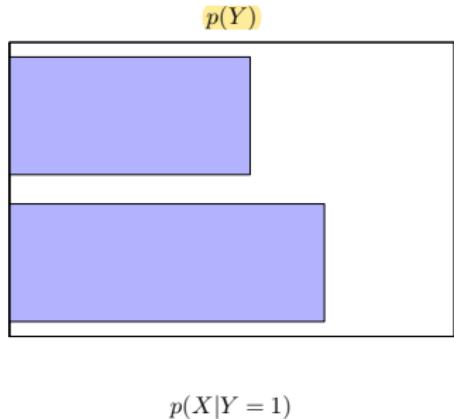
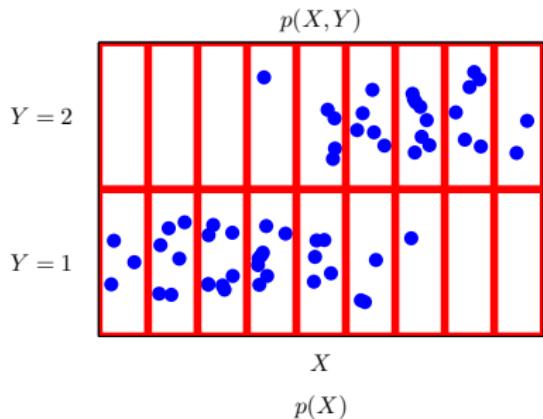
- To compute  $P(X = x_i | Y = y_j)$ , we can use the definition of the conditional probability

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} \approx \frac{\frac{n_{X=x_i, Y=y_j}}{N}}{\frac{n_{Y=y_j}}{N}} = \frac{n_{X=x_i, Y=y_j}}{n_{Y=y_j}}.$$

- In the limit  $N \rightarrow \infty$ ,

$$P(X = x_i | Y = y_j) = \lim_{N \rightarrow \infty} \frac{n_{X=x_i, Y=y_j}}{n_{Y=y_j}}.$$

# Examples of the different pmf



From Bishop (2006).

# Statistical independence

Two discrete RVs are *statistically independent* if

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j), \quad i = 1, \dots, n \quad j = 1, \dots, m.$$

# Bayes theorem

- Bayes theorem allows to go from  $P(X = x)$  to  $P(X = x|Y = y)$ .
- According to Bayes theorem

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}$$

## Example: Bayes theorem

There are two barrels in front of you. Barrel One contains 20 apples and 4 oranges. Barrel Two 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?

## Answer (I)

- There are two random variables involved.
- Let  $B$  be the random variable associated to picking one of the barrels. So  $B$  can either be “One” or “Two”.
- Let  $F$  be the random variables associated to picking a fruit. So  $F$  can either be “Apple” (A) or “Orange” (O).
- The probability we want to compute is the conditional probability  $P(B = \text{One}|F = A)$ .

## Answer (II)

- The statement says “You choose a barrel randomly” which means that  $P(B = \text{One}) = P(B = \text{Two}) = \frac{1}{2}$ .
- Since we want to go from  $P(B = \text{One})$  to  $P(B = \text{One}|F = A)$ , we can use Bayes theorem,

$$P(B = \text{One}|F = A) = \frac{P(F = A|B = \text{One})P(B = \text{One})}{P(F = A)}.$$

- We need to compute  $P(F = A|B = \text{One})$  and  $P(F = A)$ .

## Answer (III)

- Using the sum rule of probability and the product rule of probability

$$\begin{aligned} P(F = A) &= \sum P(F = A, B) = \sum P(F = A|B)P(B) \\ &= P(F = A|B = \text{One})P(B = \text{One}) \\ &\quad + P(F = A|B = \text{Two})P(B = \text{Two}). \end{aligned}$$

- From the statement,

$$P(F = A|B = \text{One}) = \frac{20}{24}$$

$$P(F = A|B = \text{Two}) = \frac{4}{12}$$

- We then have  $P(F = A) = \frac{20}{24} \frac{1}{2} + \frac{4}{12} \frac{1}{2}$ .

## Answer (IV)

We can finally compute  $P(B = \text{One}|F = A)$

$$\begin{aligned} P(B = \text{One}|F = A) &= \frac{P(F = A|B = \text{One})P(B = \text{One})}{P(F = A)} \\ &= \frac{\frac{20}{24} \frac{1}{2}}{\frac{20}{24} \frac{1}{2} + \frac{4}{12} \frac{1}{2}} \\ &= \frac{\frac{20}{24}}{\frac{20}{24} + \frac{4}{12}} = \frac{5}{7} \approx 0.71 \end{aligned}$$

## Expected value and statistical moments

- The expected value of a function of a discrete RV,  $g(X)$  is defined as

$$E\{g(X)\} = \sum_{i=1}^n g(x_i)P(X = x_i).$$

- Two expected values or *statistical moments* of the discrete RV  $X$ , used frequently are the *mean*  $\mu_X$  and the *variance*  $\sigma_X^2$ ,

$$\mu_X = E\{X\} = \sum_{i=1}^n x_i P(X = x_i),$$

$$\begin{aligned}\sigma_X^2 &= \text{var}\{X\} = E\{(X - \mu_X)^2\} = \sum_{i=1}^n (x_i - \mu_X)^2 P(X = x_i) \\ &= E\{X^2\} - \mu_X^2\end{aligned}$$

- The squared root of the variance,  $\sigma_X$ , is known as the *standard deviation*.

## Example: Expected values

- Consider the following discrete RV  $X$  and its pmf. Compute  $\mu_X$  and  $\sigma_X^2$ .

$X$	1	2	3	4
$P(X)$	0.3	0.2	0.1	0.4

- For the mean  $\mu_X$ , we have

$$\mu_X = \sum_{i=1}^n x_i P(X = x_i) = (1)(0.3) + (2)(0.2) + (3)(0.1) + (4)(0.4) = 2.6.$$

- For the variance, we first compute  $E\{X^2\}$  and then use  $\sigma_X^2 = E\{X^2\} - \mu_X^2$ .
- To compute  $E\{X^2\}$ , we can use  $E\{g(X)\}$ , where  $g(X) : X \rightarrow X^2$ ,

$$E\{X^2\} = \sum_{i=1}^n x_i^2 P(X = x_i) = (1)^2(0.3) + (2)^2(0.2) + (3)^2(0.1) + (4)^2(0.4) = 8.4.$$

- We finally get  $\sigma_X^2 = E\{X^2\} - \mu_X^2 = 8.4 - (2.6)^2 \approx 1.64$ .

# Notation

- When referring to the probability  $P(X = x)$ , we usually simply write  $P(x)$ .
- Likewise, instead of writing  $P(X = x, Y = y)$ , we simply write  $P(x, y)$ .

# Contents

Machine learning

Definitions

An example of a predictive model

## Review of probability

Random variables

Discrete random variables

**Continuous random variables**

Additional comments

# Probability density function

- A continuous RV  $X$  takes values within one or more intervals of the real line.
- We use **probability density functions** (pdf),  $p_X(x)$ , to describe a continuous RV  $X$ .
- Properties of a pdf
  1.  $p_X(x) \geq 0$ .
  2.  $\int_{-\infty}^{\infty} p_X(x)dx = 1$ .
  3.  $P(X \leq a) = \int_{-\infty}^a p_X(x)dx$ .
  4.  $P(a \leq X \leq b) = \int_a^b p_X(x)dx$ .

## Two continuous RVs

- As it was the case for discrete RVs, in ML, we are interested in analysing more than one continuous RV.
- We can use a **joint probability density function**,  $p_{X,Y}(x,y)$  to fully characterise two continuous random variables  $X$  and  $Y$ .
- Properties of a joint pdf
  1.  $p_{X,Y}(x,y) \geq 0$ .
  2.  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{X,Y}(x,y) dx dy = 1$ .
  3.  $P(X \leq a, Y \leq c) = \int_{-\infty}^a \int_{-\infty}^c p_{X,Y}(x,y) dx dy$ .
  4.  $P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d p_{X,Y}(x,y) dx dy$ .

## Rules of probability (continuous RVs)

- **Sum rule of probability.** In the case of continuous RVs, we replace the sums we had before with an integral

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x,y)dy,$$

where  $p_X(x)$  is known as the **marginal pdf**.

- **Product rule of probability.** The conditional pdf can be obtained as

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)},$$

which can also be written as

$$p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y).$$

The conditional pdf in this last form is known as the product rule of probability for two continuous RVs.

# Bayes theorem and statistical independence

- For the case of continuous RVs, Bayes theorem follows as

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)}.$$

- We say that two continuous RVs  $X$  and  $Y$  are statistically independent if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y).$$

# Expected values and statistical moments

For continuous RVs, expected values are computed as

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x)p_X(x)dx,$$

$$\mu_X = E\{X\} = \int_{-\infty}^{\infty} xp_X(x)dx,$$

$$\begin{aligned}\sigma_X^2 &= \text{var}\{X\} = E\{(X - \mu_X)^2\} = \int_{-\infty}^{\infty} (x - \mu_X)^2 p_X(x)dx. \\ &= E\{X^2\} - \mu_X^2.\end{aligned}$$

# Notation

We have been using  $p_X(x)$  or  $p_{X,Y}(x,y)$  to refer to pdfs. We will normally drop the subindex for the RVs and simply use  $p(x)$  or  $p(x,y)$  to refer to the pdfs.

# Contents

Machine learning

Definitions

An example of a predictive model

## Review of probability

Random variables

Discrete random variables

Continuous random variables

Additional comments

## What if we don't have the pmf or pdf?

- ❑ Discrete RVs. In practice, we can use data to compute the probabilities  $P(X = x_i)$  or  $P(X = x_i, Y = y_j)$  by applying the definitions we saw before.
- ❑ Notice that those definitions are valid in the limit  $N \rightarrow \infty$ .
- ❑ Continuous RVs. In practice, we assume a particular model for the pdf, eg. a Gaussian pdf, and estimate the parameters of that pdf, e.g. the mean and variance for the Gaussian pdf.
- ❑ There are advanced methods to model both pmf and pdfs but we will not study those in this module.

## What about moments?

- We can estimate  $\mu_X$  and  $\sigma_X^2$  when we have access to observations of the random variable  $X$ ,  $x_1, x_2, \dots, x_N$ , but no access to the pmf or the pdf.
- In statistics, these are called “estimators” for  $\mu_X$  and  $\sigma_X^2$ , denoted as  $\hat{\mu}_X$  and  $\hat{\sigma}^2_X$ .
- An estimator for  $\mu_X$  is given as

$$\hat{\mu}_X = \frac{1}{N} \sum_{k=1}^N x_k.$$

- An estimator for  $\sigma_X^2$  is given as

$$\hat{\sigma}^2_X = \frac{1}{N-1} \sum_{k=1}^N (x_k - \hat{\mu}_X)^2.$$

## What if we have more than two RVs?

- In ML, we are usually faced with problems where we have more than two RVs.
  - In fact, there are applications of ML in Natural Language Processing, speech processing, computer vision, computational biology, etc. where we can have hundreds of thousands or even millions of RVs.
  - The ideas that we saw before can be extended to these cases and we will see some examples in the following lectures.