# Data Analytics and Data Driven Decision

## Predictive Analytics



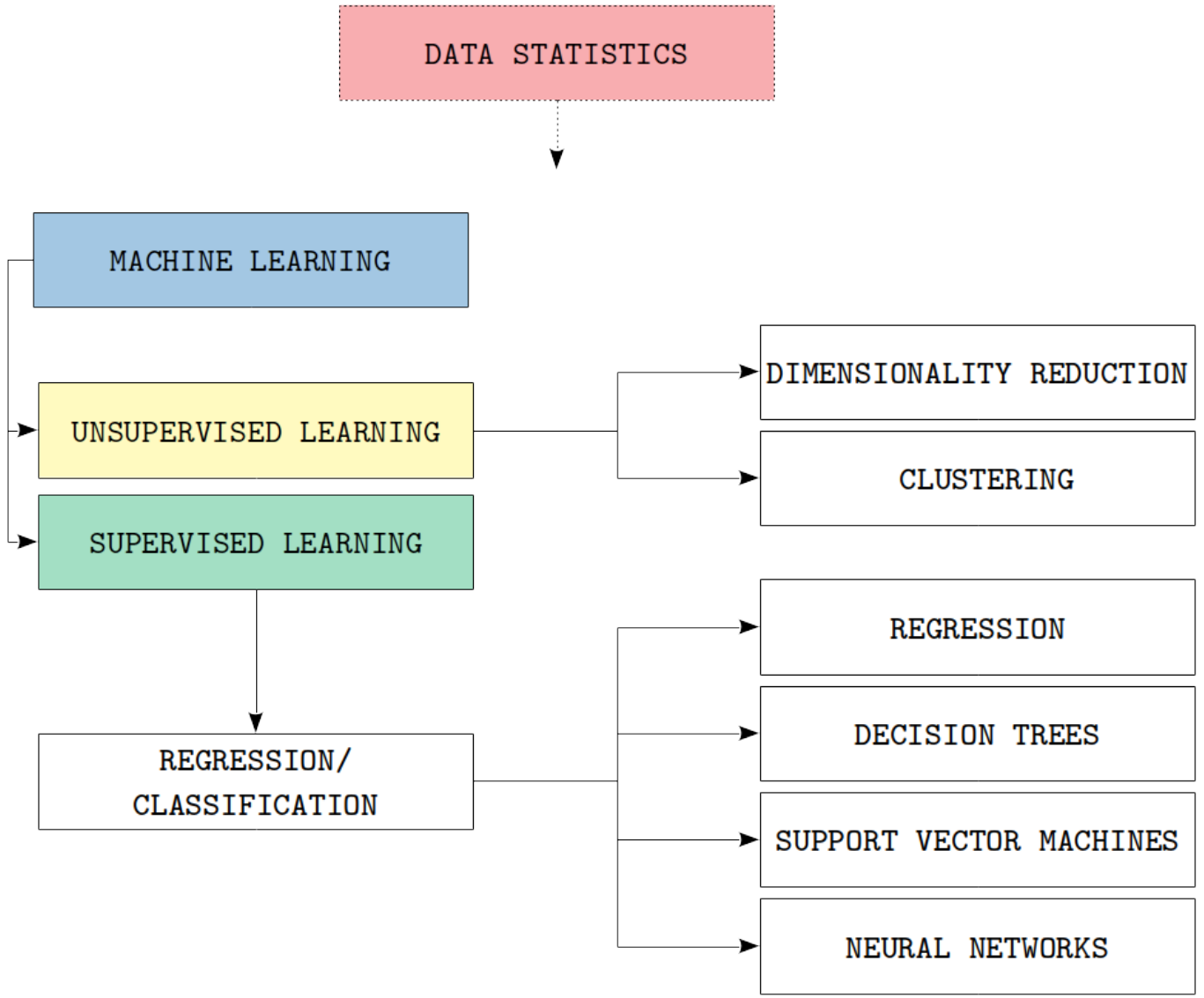
Andrea Manno
andrea.manno@univaq.it

Department of Information Engineering, Computer Science and Mathematics,
University of L'Aquila

# Overview

# Overview

# Course in a glance



Machine Learning is about learning from data (predictive models)

- data collected, manipulated and cleaned (**Descriptive Analytics**)
- good machine learning algorithms are used
- not enough...
  - data are partial (sampled) and noisy (errors)
  - big data: properly sampled? unbiased?

statistics needed with data!

introductory part of the course on **basic statistics**

# General course information

- material used during lectures (slides, exercises) is sufficient
- reference books
  - 'OpenIntro Statistics' www.leanpub.com/openintro-statistics
  - 'An Introduction to Statistical Learning' www.statlearning.com
  - 'The Elements of Statistical Learning' www.web.stanford.edu/~hastie/ElemStatLearn
- slides and exercises published on the Teams course channel
- exercises are Python notebooks (can be replicated with any software/tool you prefer (Matlab,Excel,R,...))

UNIVERSITÀ DEGLI STUDI DELL'AQUILA

# Overview



# Basic statistics concepts

arguments of this part

- DISTRIBUTIONS: SPREAD AND SIZE
- RANDOM VARIABLES
- SAMPLING
- INFERENCE
- RESAMPLING

# Distributions: spread and size

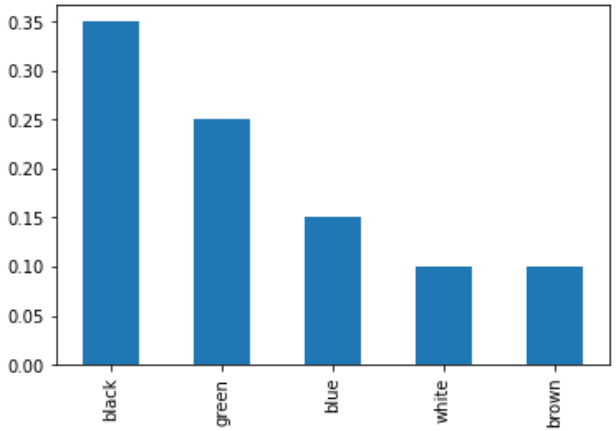
how a variable is distributed in a population

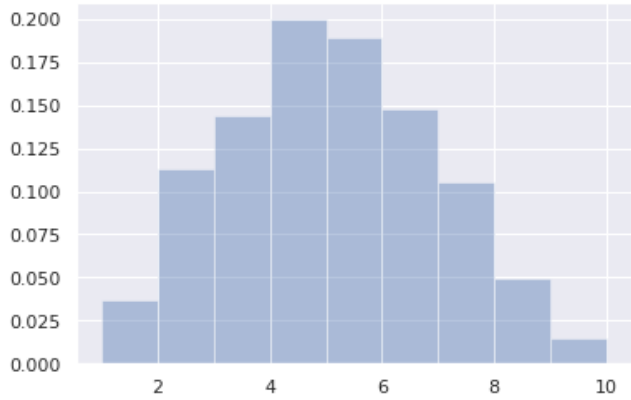
statistic is about synthesis, how to synthesize distributions?

different representations according to

- type of variables: continuous, categorical, ordered,...
- variable values, frequencies,...
- ...

| idx | value |
|---|---|
| 1 | 6.86 |
| 2 | 0.62 |
| 3 | 4.53 |
| 4 | 9.95 |
| 5 | 8.88 |
| 6 | 9.58 |
| 7 | 3.62 |
| 8 | 7.06 |
| 9 | 5.29 |
| 10 | 4.65 |
| 11 | 9.16 |
| 12 | 1.32 |
| 13 | 4.97 |
| 14 | 5.72 |
| 15 | 4.47 |
| 16 | 8.56 |
| 17 | 3.58 |
| 18 | 9.69 |
| 19 | 6.87 |
| 20 | 9.31 |

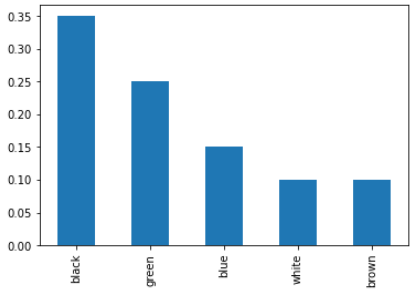


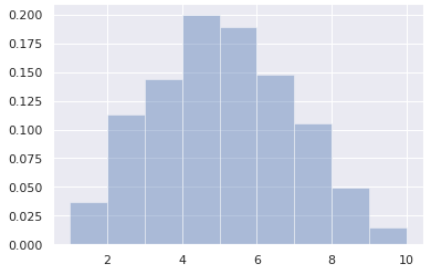




a distribution is a way of representing how a variable is distributed in a population

which measures are important in a distribution?

| idx | value |
|---|---|
| 1 | 6.86 |
| 2 | 0.62 |
| 3 | 4.53 |
| 4 | 9.95 |
| 5 | 8.88 |
| 6 | 9.58 |
| 7 | 3.62 |
| 8 | 7.06 |
| 9 | 5.29 |
| 10 | 4.65 |
| 11 | 9.16 |
| 12 | 1.32 |
| 13 | 4.97 |
| 14 | 5.72 |
| 15 | 4.47 |
| 16 | 8.56 |
| 17 | 3.58 |
| 18 | 9.69 |
| 19 | 6.87 |
| 20 | 9.31 |




spread (variability): how much the values of the variable vary in the population?

given a population of $n$ individuals (units,occurrences,events,...) with (observed) values $x_1, x_2, x_3, \ldots, x_n$, how to measure variability?

- total variability $\sum_{i=1,n} \sum_{j=2,n,j>i} |x_i - x_j|$
- average variability $\frac{1}{\frac{n(n-1)}{2}} \sum_{i=1} \sum_{j=2,n,j>i} |x_i - x_j|$

spread based on differences among values

another way to measure spread: distances with a reference value '$a$'

$$\sqrt[d]{\frac{1}{n}\sum_{i=1,n}|x_i - a|^d} \quad \text{or} \quad \frac{1}{n}\sum_{i=1,n}|x_i - a|^d \tag{1}$$

- (1) grows with the differences between values $x_i$ and $a$
- (1) represents an overall distance measure between the values and the reference

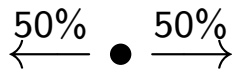
once chosen a distance how to determine the reference value (center) w.r.t. that distance?

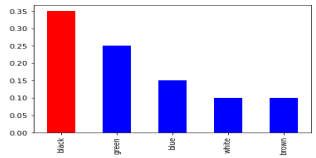
- center is something intermediate between values...
- center is not too far from any of the values...

$$c_d = \arg\min_{a\in\Re} \sqrt[d]{\frac{1}{n}\sum_{i=1,n}|x_i - a|^d} \quad \text{or} \quad \arg\min_{a\in\Re} \frac{1}{n}\sum_{i=1,n}|x_i - a|^d \tag{2}$$

if we replace each value $x_i$ with $c_d$ the loss of information is minimized!

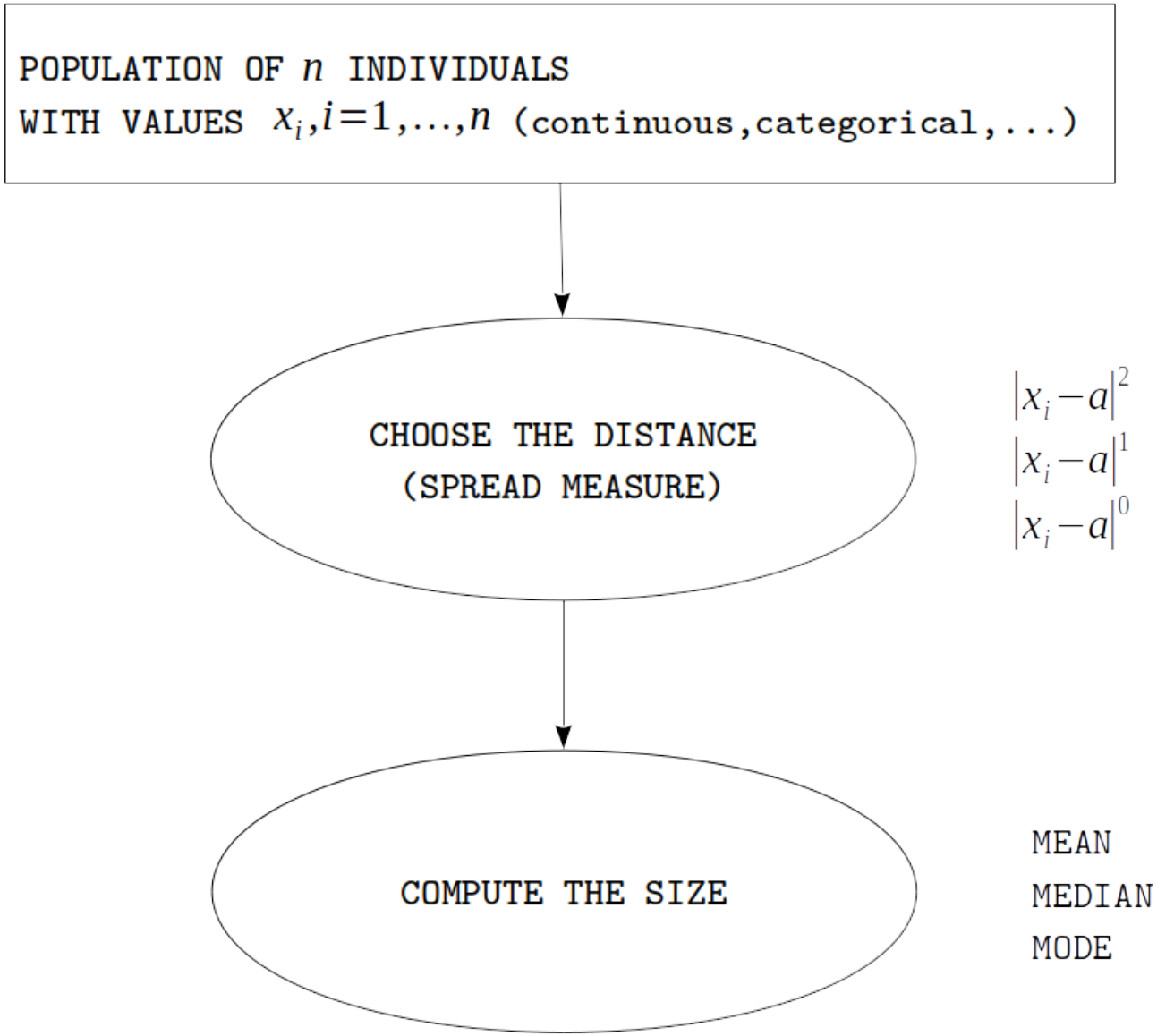
each distance has its center: the size (centrum,dimension) of the distribution

- $\arg\min_{a\in\Re} \frac{1}{n}\sum_{i=1,n} |x_i - a|^2$ MEAN $\mu = \frac{1}{n}\sum_{i=1,n} x_i$
- $\arg\min_{a\in\Re} \frac{1}{n}\sum_{i=1,n} |x_i - a|^1$ MEDIAN 

- $\arg\min_{a\in\Re} \frac{1}{n}\sum_{i=1,n} |x_i - a|^0$ MODE 


spread + size = good synthesis of the distribution!

# summing up...


POPULATION OF $n$ INDIVIDUALS
WITH VALUES $x_i, i=1,\ldots,n$ (continuous,categorical,...)
CHOOSE THE DISTANCE
(SPREAD MEASURE)
$|x_i - a|^2$
$|x_i - a|^1$
$|x_i - a|^0$
COMPUTE THE SIZE
MEAN
MEDIAN
MODE

# EXAMPLE: 2 different populations...

| idx | age |
|---|---|
| 1 | 35 |
| 2 | 36 |
| 3 | 34 |
| 4 | 35 |
| 5 | 33 |
| 6 | 37 |
| 7 | 36 |
| 8 | 31 |
| 9 | 38 |
| 10 | 35 |

Figure: population 1: mean=35

| idx | age |
|---|---|
| 1 | 68 |
| 2 | 70 |
| 3 | 65 |
| 4 | 71 |
| 5 | 67 |
| 6 | 1 |
| 7 | 2 |
| 8 | 2 |
| 9 | 3 |
| 10 | 1 |

Figure: population 2: mean=35

the spread values are 2 and 35.0365!

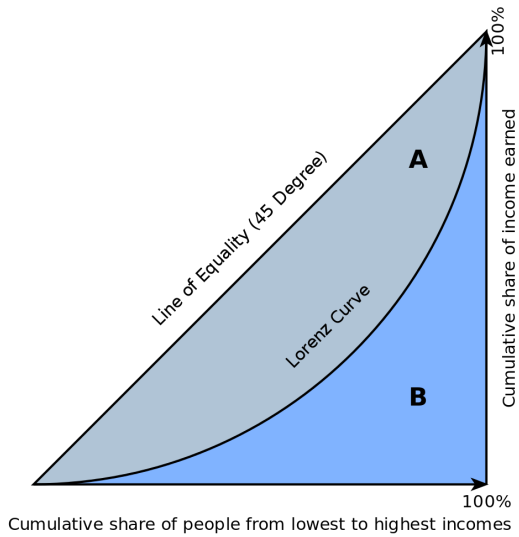
most used measures for continuous variables

- SIZE $\mu = \frac{1}{n} \sum_{i=1,n} x_i$
- SPREAD
  - VARIANCE $\sigma^2 = \frac{1}{n} \sum_{i=1,n} (x_i - \mu)^2$
  - STANDARD DEVIATION $\sigma = \sqrt[2]{\frac{1}{n} \sum_{i=1,n} (x_i - \mu)^2}$

for categorical, ordinal... ?

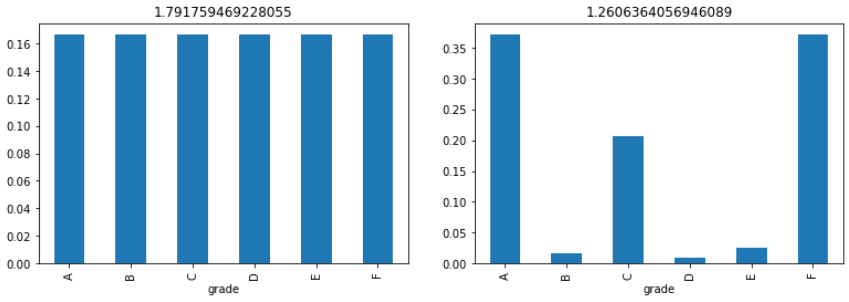
# other SPREAD measures

- INTERQUARTILE RANGE (IQR): $Q3 - Q1$


- GINI COEFFICIENT: $G = \frac{\sum_{i=1,n} \sum_{j=1,n} |x_i - x_j|}{2n \sum_{i=1,n} x_i}$

- ENTROPY: $H = -\sum_{i=1,n} f_i \log(f_i)$




exercise 1.1

VARIANCE ($\sigma^2 = \frac{1}{n} \sum_{i=1,n} (x_i - \mu)^2$) has an alternative formula

- $\sigma^2 = \frac{1}{n} \sum_{i=1,n} x_i^2 - (\frac{1}{n} \sum_{i=1,n} x_i)^2 = E\left[x^2\right] - E\left[x\right]^2$
  where $E\left[\cdot\right]$ is the expected value (average)
- $\sigma^2 = \frac{1}{n} \sum_{i=1,n} (x_i - \mu)^2 = \frac{1}{n} \sum_{i=1,n} (x_i^2 + \mu^2 - 2x_i\mu) =$

$$\frac{1}{n} \sum_{i=1,n} x_i^2 + \frac{1}{n} \sum_{i=1,n} \mu^2 - \frac{2\mu}{n} \sum_{i=1,n} x_i = \frac{1}{n} \sum_{i=1,n} x_i^2 + \frac{1}{n} \sum_{i=1,n} \mu^2 - 2\mu^2 =$$

$$\frac{1}{n} \sum_{i=1,n} x_i^2 - \mu^2 - \mu^2 + \frac{1}{n} \sum_{i=1,n} \mu^2 = E\left[x^2\right] - \mu^2 - \mu^2 + \frac{1}{n} n\mu^2 =$$

$$E\left[x^2\right] - \mu^2 = E\left[x^2\right] - E\left[x\right]^2 = \sigma^2$$

alternative formula convenient for incremental computations!

exercise 1.2

motivating game:

- I observe a variable $n$ times and hide the results to you, then compute mean and variance and let you know only these 2 values
- then a new value of the same variable is observed.

Based on the information you have

1. what do you expect this new value to be ?
2. should you bet on it, how much money would you bet ?

# Random Variables

what is a random variable?

It is a variable that can assume different values, for which it is known a probability distribution over such values (states)

probability distribution is how the random variable is distributed over its states when we observe it many time

probability distribution of a random variable measures the "likelihood" of its states

example of a random variable:

$$\text{TOSS A COIN...} P(H) = P(T) = 0.5 = p$$

# BINOMIAL DISTRIBUTION

- Bernoulli process: a sequence of observations of a binary random variable where each observation has probability of success $p$ and of failure $1 - p$ with $p \in (0, 1)$
- Binomial distribution:
  discrete probability distribution of the number of successes $k$ of $n$ independent observations of a Bernoulli process

$$\binom{n}{k} p^k (1-p)^{n-k}$$

the probability of getting $k$ heads (or tails) by tossing a coin $n$ times

- the probability of $n$ independent events is the product of the probabilities of the single events

$$\prod_{i=1}^{k} p \prod_{i=1}^{n-k} (1-p) = p^k (1-p)^{n-k}$$

- how many possible way of getting $k$ successes on $n$ observations (trials)?

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

example: $k = 2, n = 4 \rightarrow \binom{4}{2} = 6$:

| | | | | |
|---|---|---|---|---|
| 1 | H | H | T | T |
| 2 | T | H | H | T |
| 3 | T | T | H | H |
| 4 | H | T | T | H |
| 5 | H | T | H | T |
| 6 | T | H | T | H |

exercise 1.3

given a set $\Omega$ of all possible values $x_i$ of a discrete random variable $X$, and a probability $p_i$ associated to each $x_i$ according to a probability distribution

- MEAN,EXPECTED VALUE,PREDICTION

$$m = \mathbb{P}(X) = \sum_{i \in \Omega} p_i x_i$$

- VARIANCE

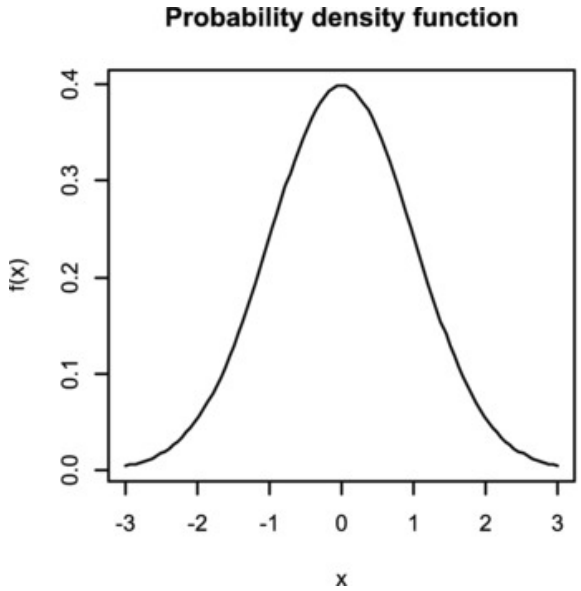
$$var(X) = \sum_{i \in \Omega} p_i (x_i - m)^2$$

## continuous probability distributions

- density function $f(x)$:

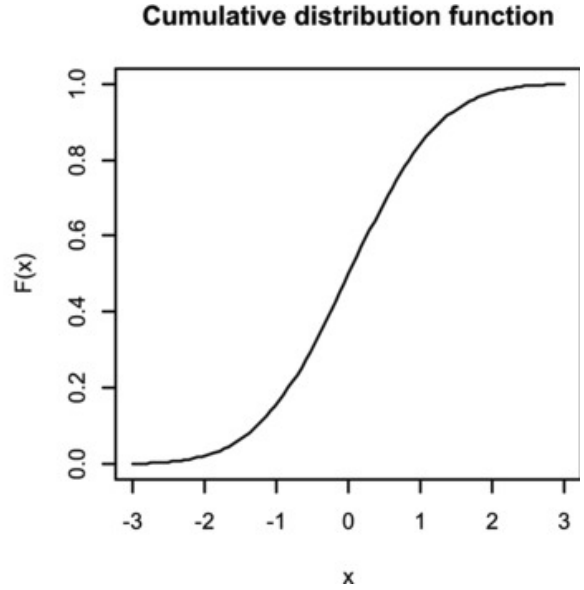
  $f(x) \geq 0, f(x) \leq 1,$

  $$\int_{-\infty}^{\infty} f(x)dx = 1$$




- cumulative function $F(x)$ :

  $$F(x) = \int_{-\infty}^{x} f(y)dy$$




what is the probability that...

... $x = a$?

... $x \leq b$?

... $a \leq x \leq b$?

given a continuous random variable $X$ with values $x$ defined on $[-\infty, \infty]$ associated to a continuous probability density function $f(x)$

- MEAN,EXPECTED VALUE,PREDICTION

$$m = \mathbb{P}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

- VARIANCE

$$var(X) = \int_{-\infty}^{\infty} (x - m)^2 f(x)dx$$

# UNIFORM DISTRIBUTION $U_{a,b}(x)$

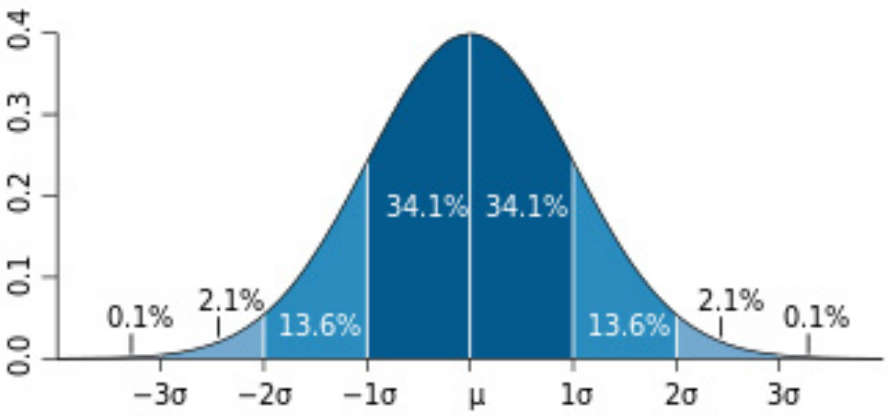
- defined on a interval $[a, b]$ with $a < b$
- the probability of any value in the interval is the same (outside is 0)

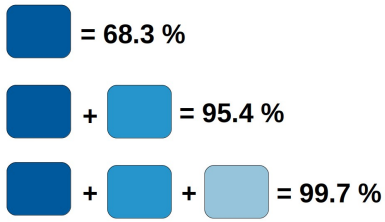
$$f(x) = U_{a,b}(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

- probability of an interval of lenght $h$: $\frac{h}{b-a}$

exercise 1.4

# NORMAL DISTRIBUTION $N_{\mu,\sigma}(x)$




- defined on the interval $[-\infty, \infty]$

density:

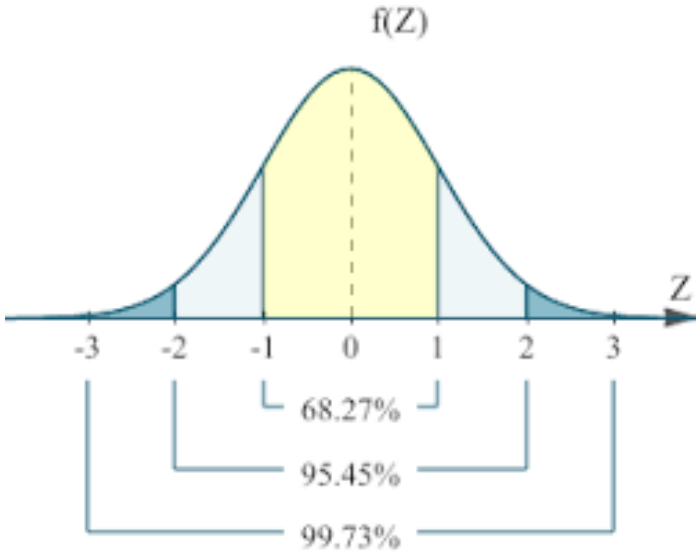
$$f(x) = N_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

cumulative:

$$F(x) = \Phi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy$$

- symmetric
- fully described by parameters $\mu$ (mean) and $\sigma$ (standard deviation)
- central value = mean, median, mode

# STANDARD NORMAL DISTRIBUTION $N(x)$ (often denoted as $N(Z)$)




- $\mu = 0$ and $\sigma = 1$

density:

$$N(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$$

cumulative:

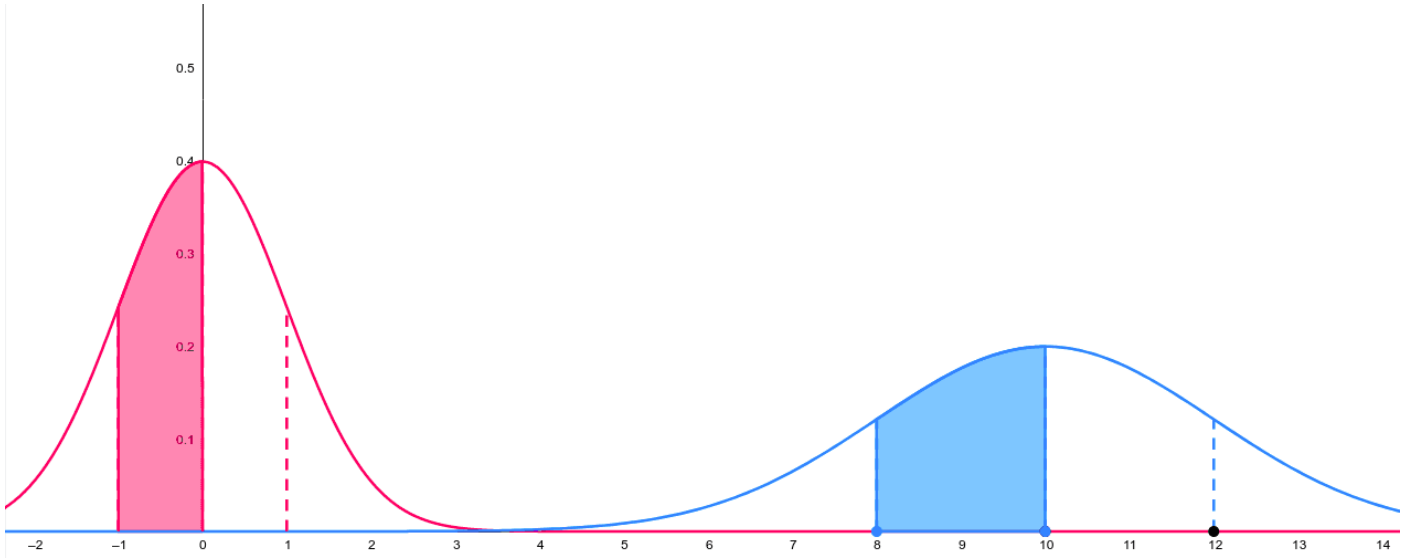
$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{1}{2}y^2} dy$$

- for $\Phi(Z)$ available a table of values

$$\Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} e^{-\frac{1}{2}y^2} dy$$

| z | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| +0 | .50000 | .50399 | .50798 | .51197 | .51595 | .51994 | .52392 | .52790 | .53188 | .53586 |
| +0.1 | .53983 | .54380 | .54776 | .55172 | .55567 | .55966 | .56360 | .56749 | .57142 | .57535 |
| +0.2 | .57926 | .58317 | .58706 | .59095 | .59483 | .59871 | .60257 | .60642 | .61026 | .61409 |
| +0.3 | .61791 | .62172 | .62552 | .62930 | .63307 | .63683 | .64058 | .64431 | .64803 | .65173 |
| +0.4 | .65542 | .65910 | .66276 | .66640 | .67003 | .67364 | .67724 | .68082 | .68439 | .68793 |
| +0.5 | .69146 | .69497 | .69847 | .70194 | .70540 | .70884 | .71226 | .71566 | .71904 | .72240 |
| +0.6 | .72575 | .72907 | .73237 | .73565 | .73891 | .74215 | .74537 | .74857 | .75175 | .75490 |
| +0.7 | .75804 | .76115 | .76424 | .76730 | .77035 | .77337 | .77637 | .77935 | .78230 | .78524 |
| +0.8 | .78814 | .79103 | .79389 | .79673 | .79955 | .80234 | .80511 | .80785 | .81057 | .81327 |
| +0.9 | .81594 | .81859 | .82121 | .82381 | .82639 | .82894 | .83147 | .83398 | .83646 | .83891 |
| +1 | .84134 | .84375 | .84614 | .84849 | .85083 | .85314 | .85543 | .85769 | .85993 | .86214 |
| +1.1 | .86433 | .86650 | .86864 | .87076 | .87286 | .87493 | .87698 | .87900 | .88100 | .88298 |
| +1.2 | .88493 | .88686 | .88877 | .89065 | .89251 | .89435 | .89617 | .89796 | .89973 | .90147 |
| +1.3 | .90320 | .90490 | .90658 | .90824 | .90988 | .91149 | .91308 | .91466 | .91621 | .91774 |
| +1.4 | .91924 | .92073 | .92220 | .92364 | .92507 | .92647 | .92785 | .92922 | .93056 | .93189 |
| +1.5 | .93319 | .93448 | .93574 | .93699 | .93822 | .93943 | .94062 | .94179 | .94295 | .94408 |
| +1.6 | .94520 | .94630 | .94738 | .94845 | .94950 | .95053 | .95154 | .95254 | .95352 | .95449 |
| +1.7 | .95543 | .95637 | .95728 | .95818 | .95907 | .95994 | .96080 | .96164 | .96246 | .96327 |
| +1.8 | .96407 | .96485 | .96562 | .96638 | .96712 | .96784 | .96856 | .96926 | .96995 | .97062 |
| +1.9 | .97128 | .97193 | .97257 | .97320 | .97381 | .97441 | .97500 | .97558 | .97615 | .97670 |
| +2 | .97725 | .97778 | .97831 | .97882 | .97932 | .97982 | .98030 | .98077 | .98124 | .98169 |
| +2.1 | .98214 | .98257 | .98300 | .98341 | .98382 | .98422 | .98461 | .98500 | .98537 | .98574 |
| +2.2 | .98610 | .98645 | .98679 | .98713 | .98745 | .98778 | .98809 | .98840 | .98870 | .98899 |
| +2.3 | .98928 | .98956 | .98983 | .99010 | .99036 | .99061 | .99086 | .99111 | .99134 | .99158 |
| +2.4 | .99180 | .99202 | .99224 | .99245 | .99266 | .99286 | .99305 | .99324 | .99343 | .99361 |
| +2.5 | .99379 | .99396 | .99413 | .99430 | .99446 | .99461 | .99477 | .99492 | .99506 | .99520 |
| +2.6 | .99534 | .99547 | .99560 | .99573 | .99585 | .99598 | .99609 | .99621 | .99632 | .99643 |
| +2.7 | .99653 | .99664 | .99674 | .99683 | .99693 | .99702 | .99711 | .99720 | .99728 | .99736 |
| +2.8 | .99744 | .99752 | .99760 | .99767 | .99774 | .99781 | .99788 | .99795 | .99801 | .99807 |
| +2.9 | .99813 | .99819 | .99825 | .99831 | .99836 | .99841 | .99846 | .99851 | .99856 | .99861 |
| +3 | .99865 | .99869 | .99874 | .99878 | .99882 | .99886 | .99889 | .99893 | .99896 | .99900 |

## standardization of a normal variable




$x \approx N_{\mu,\sigma}(x)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

standardization

$$\Rightarrow Z = \frac{x-\mu}{\sigma} \Rightarrow$$

$Z \approx N(Z)$:

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$$

it can be shown that

$$N_{\mu,\sigma}(x) = \frac{1}{\sigma} N\left(\frac{x-\mu}{\sigma}\right) \quad \text{and} \quad \Phi_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

# CENTRAL LIMIT THEOREM (simplified)

given $n$ random variables independent and identically distributed $X_1, X_2, \ldots, X_n$, their arithmetic mean is normally distributed

$$S_n = \frac{X_1 + X_2 +, \cdots + X_n}{n} \approx N(\cdot)$$

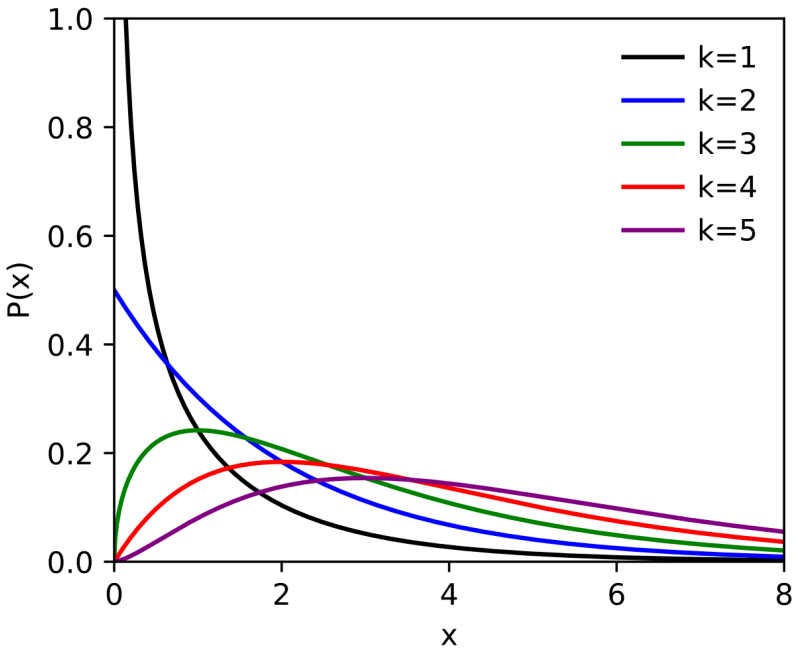
exercise 1.5

# $\chi^2_k$ DISTRIBUTION

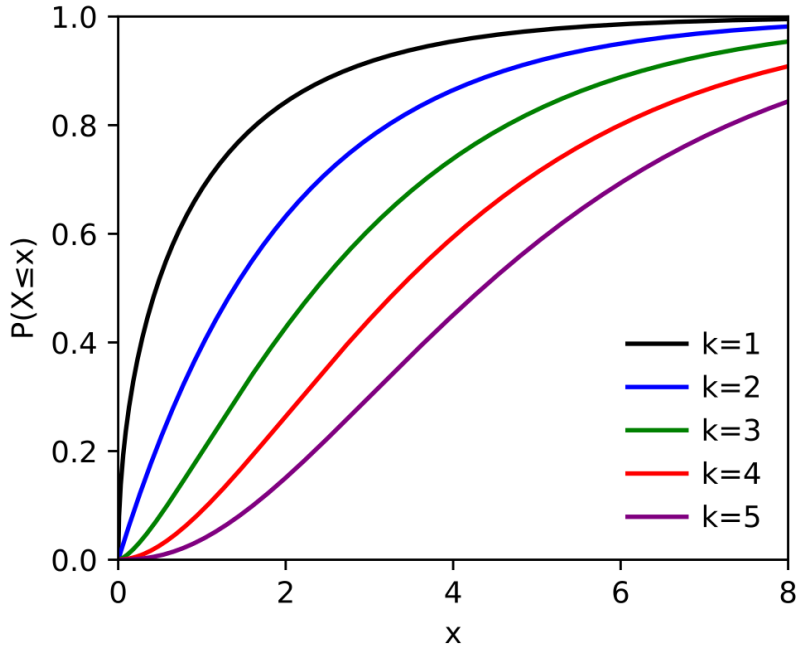
- $\chi^2_k = \sum_{i=1,k} x_i^2 = x_1^2 + x_2^2 + \cdots + x_k^2$
  with $x_i \approx N_{0,1}(x)$

- $k$=degrees of freedom

density:




cumulative:



# Sampling

inference based on population and samples

- population: large (infinite) collection of units on which a variable can be measured
- sample (with or without repetition): finite subset of units of the population
- random sample: a sample extracted randomly from the population

example:

- the mean of the age of a population of 1000 units is unknown
- a random sample of 100 units is extracted and its sample mean is computed

the sample mean is a random variable!

$\Rightarrow$ it must have a probability distribution (e.g. a mean and a variance)...

# Inference

what is inference? two settings...

- interval estimate:
  - I know a random variable distribution on a population
  - I observe the random variable on a sample of the population

  can I estimate an interval in which a certain parameter (e.g. $\mu$) of the distribution of the population may lie with high probability?

- hypothesis rejection:
  - I know a random variable distribution on a population
  - I make an hypothesis on the value of a certain parameter of the distribution
  - I observe the random variable on a sample of the population

  do the observations on the sample confim my hypothesis?

## confidence interval for the mean

- a random variable $X$ in a population is distributed as $N_{\mu,\sigma}(x)$ with $\sigma$ known and $\mu$ unknown
- a random sample $S := \{x_1, x_2, \ldots, x_n\}$ is observed
- sample mean $\bar{x} = \frac{1}{n} \sum_{i=1,n} x_i$ is a random variable

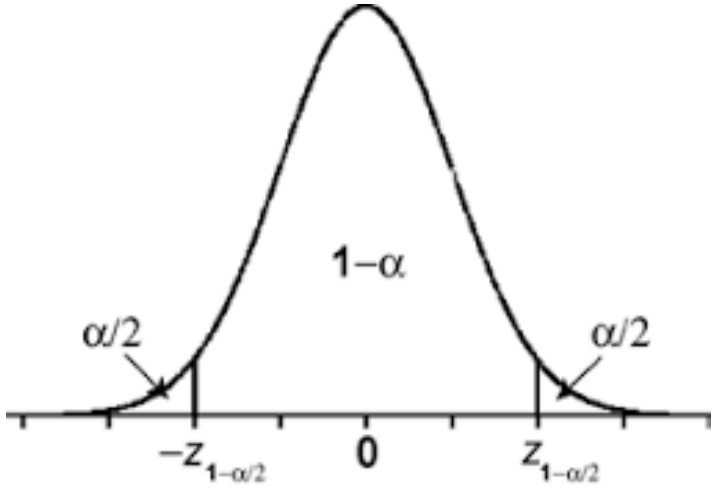
it can be shown that $\bar{x} \approx N_{\mu, \frac{\sigma}{\sqrt{n}}}(\bar{x})$, i.e. $E(\bar{x}) = \mu$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$$\Rightarrow N_{\mu, \frac{\sigma}{\sqrt{n}}}\left(\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \approx N(Z)$$

we want determine an interval of values in which the unknown $\mu$ falls with probability $1 - \alpha$ (e.g. 95% if $\alpha = 0.05$)...

exploit the fact that $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is a standard normal variable!

# confidence interval for the mean




$$P\{Z_{\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{1-\frac{\alpha}{2}}\} = 1-\alpha$$

$$P\{-\bar{x}+\frac{\sigma}{\sqrt{n}}Z_{\frac{\alpha}{2}} \leq -\mu \leq -\bar{x}+\frac{\sigma}{\sqrt{n}}Z_{1-\frac{\alpha}{2}}\} = 1-\alpha$$

$$P\{\bar{x}-\frac{\sigma}{\sqrt{n}}Z_{1-\frac{\alpha}{2}} \leq \mu \leq \bar{x}-\frac{\sigma}{\sqrt{n}}Z_{\frac{\alpha}{2}}\} = 1-\alpha$$

$$P\{\bar{x}-\frac{\sigma}{\sqrt{n}}Z_{1-\frac{\alpha}{2}} \leq \mu \leq \bar{x}+\frac{\sigma}{\sqrt{n}}Z_{1-\frac{\alpha}{2}}\} = 1-\alpha$$

example:
the viral load of asymptomatic covid patients in L'Aquila is distibuted as a normal random variable with standard deviation 0.7 and unknown mean. The viral load is observed on a sample of 25 asymptomatic patients and the sample mean is 5.23 copies/ml (in Log10 scale). Define an interval of values $[a, b]$ in which the mean of the whole population lies with a 95% probability.

$$\text{DATA:} \begin{cases} \sigma = 0.7 \\ \bar{x} = 5.23 \\ n = 25 \\ 1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \end{cases} \qquad \text{UNKNOWNS:} \begin{cases} a \\ b \end{cases}$$

$$P\{5.23 - \frac{0.7}{\sqrt{25}} Z_{0.975} \leq \mu \leq 5.23 + \frac{0.7}{\sqrt{25}} Z_{0.975}\} = 0.95$$

| z | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 |
|---|---|---|---|---|---|---|---|---|
| +0 | .50000 | .50399 | .50798 | .51197 | .51595 | .51994 | .52392 | .52790 |
| +0.1 | .53983 | .54380 | .54776 | .55172 | .55567 | .55966 | .56360 | .56749 |
| +0.2 | .57926 | .58317 | .58706 | .59095 | .59483 | .59871 | .60257 | .60642 |
| +0.3 | .61791 | .62172 | .62552 | .62930 | .63307 | .63683 | .64058 | .64431 |
| +0.4 | .65542 | .65910 | .66276 | .66640 | .67003 | .67364 | .67724 | .68082 |
| +0.5 | .69146 | .69497 | .69847 | .70194 | .70540 | .70884 | .71226 | .71566 |
| +0.6 | .72575 | .72907 | .73237 | .73565 | .73891 | .74215 | .74537 | .74857 |
| +0.7 | .75804 | .76115 | .76424 | .76730 | .77035 | .77337 | .77637 | .77935 |
| +0.8 | .78814 | .79103 | .79389 | .79673 | .79955 | .80234 | .80511 | .80785 |
| +0.9 | .81594 | .81859 | .82121 | .82381 | .82639 | .82894 | .83147 | .83398 |
| +1 | .84134 | .84375 | .84614 | .84849 | .85083 | .85314 | .85543 | .85769 |
| +1.1 | .86433 | .86650 | .86864 | .87076 | .87286 | .87493 | .87698 | .87900 |
| +1.2 | .88493 | .88686 | .88877 | .89065 | .89251 | .89435 | .89617 | .89796 |
| +1.3 | .90320 | .90490 | .90658 | .90824 | .90988 | .91149 | .91308 | .91466 |
| +1.4 | .91924 | .92073 | .92220 | .92364 | .92507 | .92647 | .92785 | .92922 |
| +1.5 | .93319 | .93448 | .93574 | .93699 | .93822 | .93943 | .94062 | .94179 |
| +1.6 | .94520 | .94630 | .94738 | .94845 | .94950 | .95053 | .95154 | .95254 |
| +1.7 | .95543 | .95637 | .95728 | .95818 | .95907 | .95994 | .96080 | .96164 |
| +1.8 | .96407 | .96485 | .96562 | .96638 | .96712 | .96784 | .96856 | .96926 |
| +1.9 | .97128 | .97193 | .97257 | .97320 | .97381 | .97441 | .97500 | .97558 |
| +2 | .97725 | .97778 | .97831 | .97882 | .97932 | .97982 | .98030 | .98077 |

$$Z_{0.975} = 1.96, \qquad \frac{0.7}{\sqrt{25}} = 0.14$$

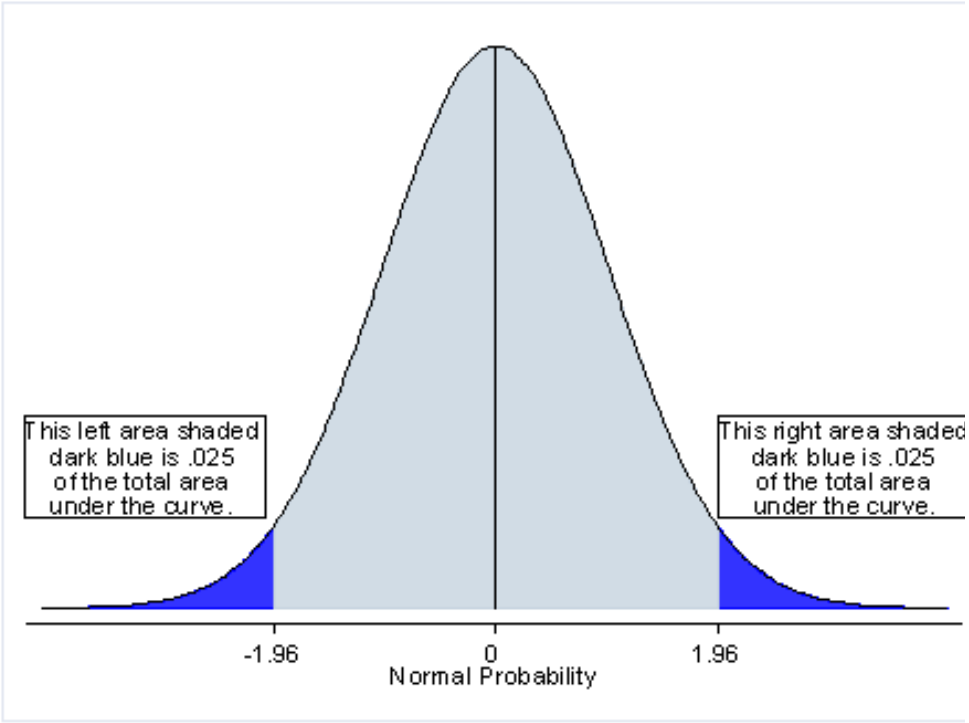
$$a = 5.23 - 1.96 * 0.14 = 4.9556$$

$$b = 5.23 + 1.96 * 0.14 = 5.5044$$

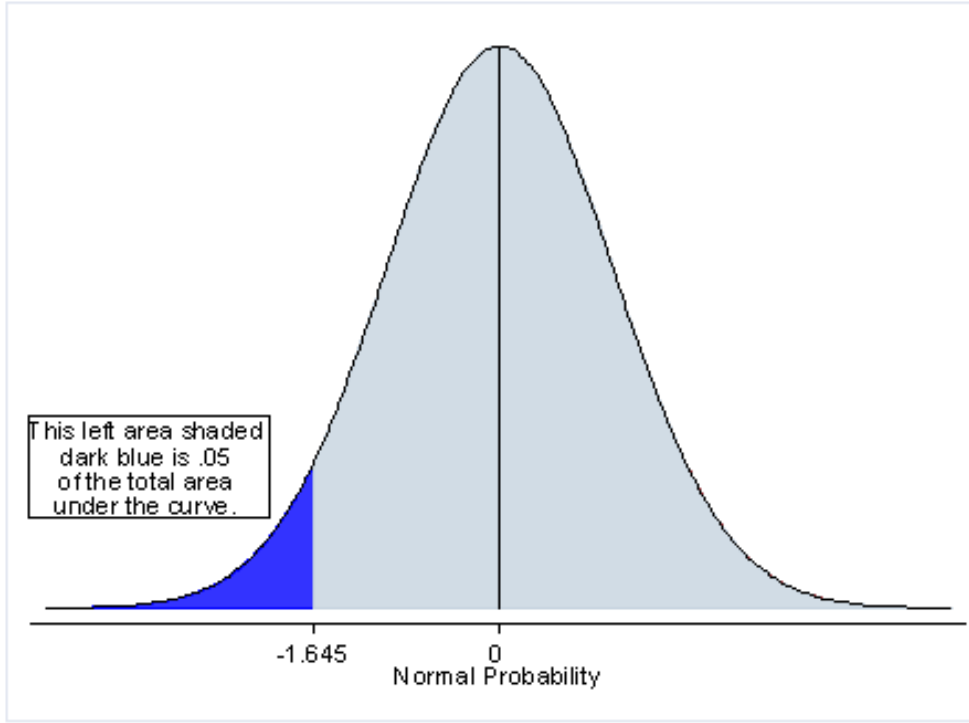
$$P\{4.9556 \leq \mu \leq 5.5044\} = 0.95$$

UNIVERSITÀ DEGLI STUDI DELL'AQUILA

two-tailed confidence interval




one-tailed confidence interval




what is the value $a$ such the mean of the viral load is greater than or equal to $a$ ($a \leq \mu$) with probability 95%?

$$P\{5.23 - \frac{0.7}{\sqrt{25}} Z_{0.95} \leq \mu\} = 0.95$$

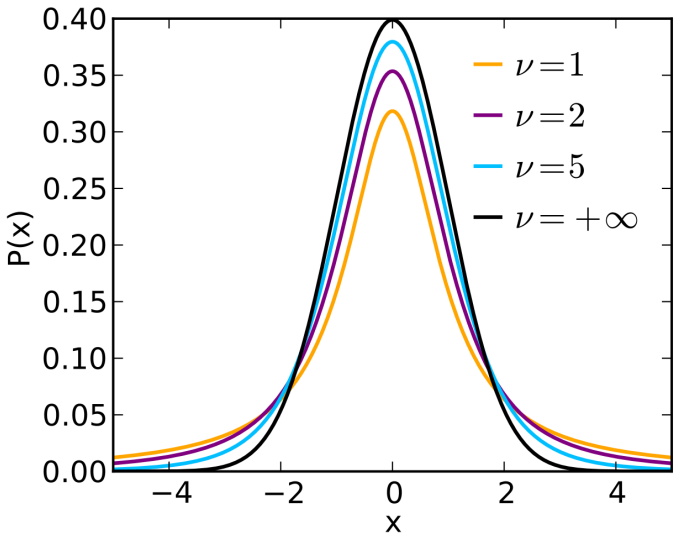
$$Z_{0.95} = 1.645, \frac{0.7}{\sqrt{25}} = 0.14, \qquad a = 5.23 - 1.645 * 0.14 = 4.9997$$

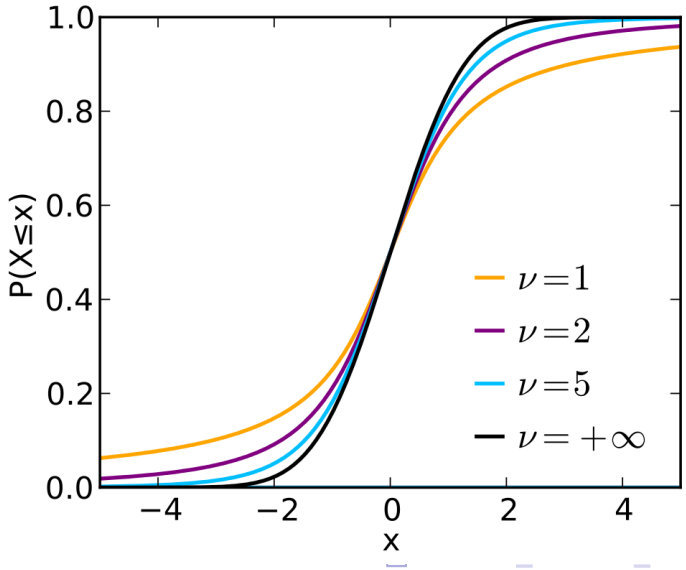
$$P\{4.9997 \leq \mu\} = 0.95$$

## confidence interval for the mean

what if $\sigma$ is unknown? $\Rightarrow$ STUDENT'S-t DISTRIBUTION

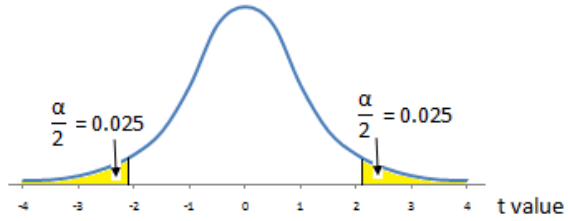
- $t_n = \frac{Z}{\frac{V}{\sqrt{n}}}$ with $\begin{cases} Z \approx N(x) \\ V^2 \approx \chi^2_n \end{cases}$ , and $\nu = n$ are the degrees of freedom
- symmmetric around the mean 0 like $N(x)$
- "shorter" and "fatter" than $N(x)$
- tends to $N(x)$ when $\nu$ grows
- table of values available for the cumulative distribution



# Student's t Distribution Table

For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** $\alpha = 0.05$).




| | 90% | 95% | 97.5% | 99% | 99.5% | 99.95% | 1-Tail Confidence Level |
|---|---|---|---|---|---|---|---|
| | 80% | 90% | 95% | 98% | 99% | 99.9% | 2-Tail Confidence Level |
| | 0.100 | 0.050 | 0.025 | 0.010 | 0.005 | 0.0005 | 1-Tail Alpha |
| *df* | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 | 0.001 | 2-Tail Alpha |
| 1 | 3.0777 | 6.3138 | 12.7062 | 31.8205 | 63.6567 | 636.6192 | |
| 2 | 1.8856 | 2.9200 | 4.3027 | 6.9646 | 9.9248 | 31.5991 | |
| 3 | 1.6377 | 2.3534 | 3.1824 | 4.5407 | 5.8409 | 12.9240 | |
| 4 | 1.5332 | 2.1318 | 2.7764 | 3.7469 | 4.6041 | 8.6103 | |
| 5 | 1.4759 | 2.0150 | 2.5706 | 3.3649 | 4.0321 | 6.8688 | |
| 6 | 1.4398 | 1.9432 | 2.4469 | 3.1427 | 3.7074 | 5.9588 | |
| 7 | 1.4149 | 1.8946 | 2.3646 | 2.9980 | 3.4995 | 5.4079 | |
| 8 | 1.3968 | 1.8595 | 2.3060 | 2.8965 | 3.3554 | 5.0413 | |
| 9 | 1.3830 | 1.8331 | 2.2622 | 2.8214 | 3.2498 | 4.7809 | |
| 10 | 1.3722 | 1.8125 | 2.2281 | 2.7638 | 3.1693 | 4.5869 | |
| 11 | 1.3634 | 1.7959 | 2.2010 | 2.7181 | 3.1058 | 4.4370 | |
| 12 | 1.3562 | 1.7823 | 2.1788 | 2.6810 | 3.0545 | 4.3178 | |
| 13 | 1.3502 | 1.7709 | 2.1604 | 2.6503 | 3.0123 | 4.2208 | |
| 14 | 1.3450 | 1.7613 | 2.1448 | 2.6245 | 2.9768 | 4.1405 | |
| 15 | 1.3406 | 1.7531 | 2.1314 | 2.6025 | 2.9467 | 4.0728 | |
| 16 | 1.3368 | 1.7459 | 2.1199 | 2.5835 | 2.9208 | 4.0150 | |
| 17 | 1.3334 | 1.7396 | 2.1098 | 2.5669 | 2.8982 | 3.9651 | |
| 18 | 1.3304 | 1.7341 | 2.1009 | 2.5524 | 2.8784 | 3.9216 | |
| 19 | 1.3277 | 1.7291 | 2.0930 | 2.5395 | 2.8609 | 3.8834 | |
| 20 | 1.3253 | 1.7247 | 2.0860 | 2.5280 | 2.8453 | 3.8495 | |
| 21 | 1.3232 | 1.7207 | 2.0796 | 2.5176 | 2.8314 | 3.8193 | |
| 22 | 1.3212 | 1.7171 | 2.0739 | 2.5083 | 2.8188 | 3.7921 | |
| 23 | 1.3195 | 1.7139 | 2.0687 | 2.4999 | 2.8073 | 3.7676 | |
| 24 | 1.3178 | 1.7109 | 2.0639 | 2.4922 | 2.7969 | 3.7454 | |
| 25 | 1.3163 | 1.7081 | 2.0595 | 2.4851 | 2.7874 | 3.7251 | |

## confidence interval for the mean

- a random variable $X$ in a population is distributed as $N_{\mu,\sigma}(x)$ with $\mu$ and $\sigma$ unknown
- a random sample $S := \{x_1, x_2, \ldots, x_n\}$ is observed
- sample mean $\bar{x} = \frac{1}{n} \sum_{i=1,n} x_i$, sample standard deviation $\sigma_S = \sqrt{\frac{\sum_{i=1,n} (x_i - \bar{x})^2}{n-1}}$

it can be shown that $\frac{\bar{x}-\mu}{\frac{\sigma_S}{\sqrt{n}}} \approx t_{n-1}$, i.e. a Student's-t distribution with $\nu = n - 1$ degrees of freedom

$$P\{t_{n-1}^{\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{\frac{\sigma_S}{\sqrt{n}}} \leq t_{n-1}^{1-\frac{\alpha}{2}}\} = 1 - \alpha$$

$$P\{\bar{x} - \frac{\sigma_S}{\sqrt{n}} t_{n-1}^{1-\frac{\alpha}{2}} \leq \mu \leq \bar{x} + \frac{\sigma_S}{\sqrt{n}} t_{n-1}^{1-\frac{\alpha}{2}}\} = 1 - \alpha$$

if $n > 30$ $P\{\bar{x} - \frac{\sigma_S}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} \leq \mu \leq \bar{x} + \frac{\sigma_S}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}\} = 1 - \alpha$

example 2.0:
consider the viral load of asymptomatic covid patients in L'Aquila is distibuted as a normal random variable with unknown mean and standard deviation. The viral load is observed on a sample of 25 asymptomatic patients and the sample mean is 5.23 copies/ml while the sample standard deviation is 0.7. Define an interval of values $[a, b]$ in which the mean of the whole population lies with a 95% probability.

DATA: $\begin{cases} \sigma_S = 0.7 \\ \bar{x} = 5.23 \\ n = 25 \\ 1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \end{cases}$ UNKNOWNS: $\begin{cases} a \\ b \end{cases}$

$$P\{5.23 - \frac{0.7}{\sqrt{25}} t_{24}^{0.975} \leq \mu \leq 5.23 + \frac{0.7}{\sqrt{25}} t_{24}^{0.975}\} = 0.95$$

| | | | |
|---|---|---|---|
| 16 | 1.3368 | 1.7459 | 2.1199 |
| 17 | 1.3334 | 1.7396 | 2.1098 |
| 18 | 1.3304 | 1.7341 | 2.1009 |
| 19 | 1.3277 | 1.7291 | 2.0930 |
| 20 | 1.3253 | 1.7247 | 2.0860 |
| 21 | 1.3232 | 1.7207 | 2.0796 |
| 22 | 1.3212 | 1.7171 | 2.0739 |
| 23 | 1.3195 | 1.7139 | 2.0687 |
| 24 | 1.3178 | 1.7109 | 2.0639 |
| 25 | 1.3163 | 1.7081 | 2.0595 |

$$t_{24}^{0.975} = 2.0639, \quad \frac{0.7}{\sqrt{25}} = 0.14$$

$$a = 5.23 - 2.0639 * 0.14 = 4.941$$

$$b = 5.23 + 2.0639 * 0.14 = 5.519$$

$$P\{4.941 \leq \mu \leq 5.519\} = 0.95$$

UNIVERSITÀ DEGLI STUDI DELL'AQUILA

## test of hypothesis

two alternative hypotheses

- $H_0$ (null hypothesis): a random variable follows a certain distribution with certain parameters
- $H_1$: $H_0$ is false (different alternatives are possible)

by observing a sample of the random variable is it possible to say if the random variable follows the distribution of $H_0$ with a certain confidence level $1 - \alpha$?

1. not possible to reject $H_0$ (different from accepting $H_0$)
2. $H_0$ is rejected ($H_0$ is false and $H_1$ is true)

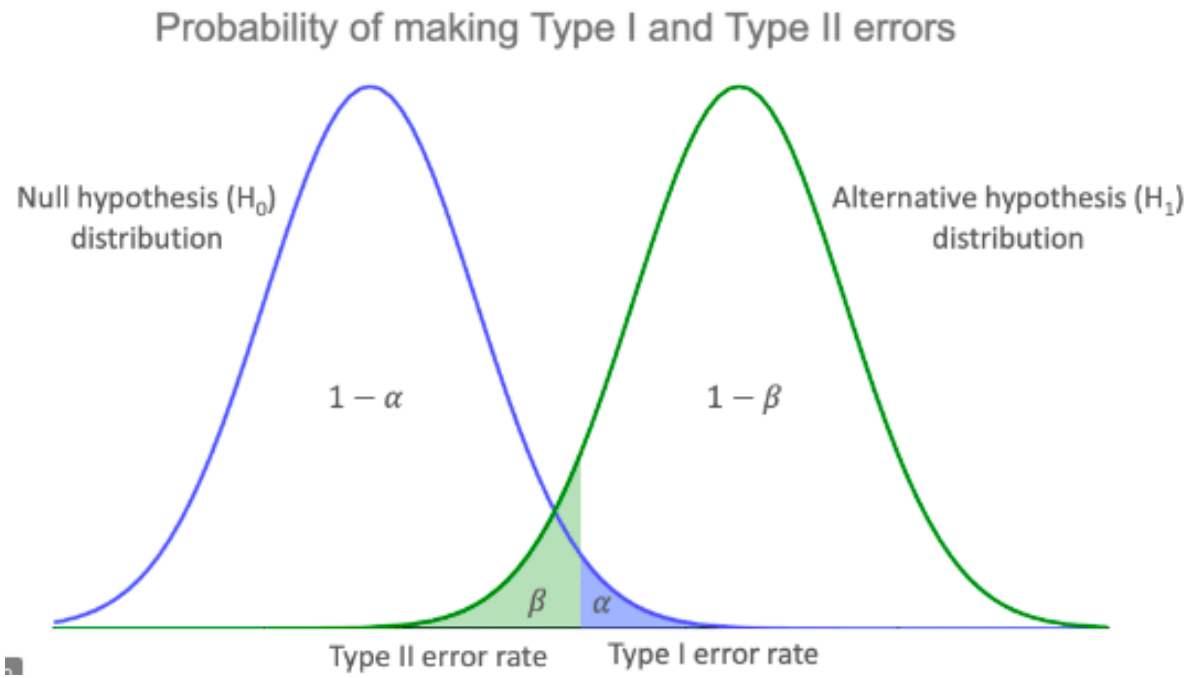
two settings

- one sample test
- two samples test

## test of hypothesis

two possible errors

- type 1 error: rejecting $H_0$ when $H_0$ is true (probability $\alpha$)
- type 2 error: not rejecting $H_0$ when $H_0$ is false (probability $\beta$)




$1-\beta$ power of the test: probability of rejecting $H_0$ when $H_0$ is false

## one sample hypothesis test

- a random variable $X$ in a population is normally distributed with known standard deviation $\sigma$ and an hypotesis test is made on the value of its mean $\mu$ with confidence level $1-\alpha$
  - $H_0$: $\mu = \mu_0$
  - $H_1$: $\mu > \mu_0$ (one sided alternative)
- a random sample $S := \{x_1, x_2, \ldots, x_n\}$ is observed with sample mean $\bar{x} = \frac{1}{n} \sum_{i=1,n} x_i$

if $H_0$ is "true" $\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \approx N(x)$ then

$$P\{\frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \leq Z_{1-\alpha}\} = 1-\alpha \quad \Rightarrow \quad P\{\bar{x} \leq \mu_0 + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha}\} = 1-\alpha$$

$$\text{if } \begin{cases} \bar{x} \leq \mu_0 + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha}, & H_0 \text{ not rejected (error type 2 with probability } \beta) \\ \bar{x} > \mu_0 + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha}, & H_0 \text{ rejected (error type 1 with probability } \alpha) \end{cases}$$

example:
a normal random variable has standard deviation $\sigma = 3.6$ with unknown mean. Consider the following hypotheses

- $H_0$: $\mu = 12$
- $H_1$: $\mu > 12$

by observing a sample of 16 realizations of the random variable with sample mean 13.7, can we accept $H_0$ with confidence level 95% ($\alpha = 0.05$)?

| z | 0 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 |
|---|---|---|---|---|---|---|---|---|
| +0 | .50000 | .50399 | .50798 | .51197 | .51595 | .51994 | .52392 | .52790 |
| +0.1 | .53983 | .54380 | .54776 | .55172 | .55567 | .55966 | .56360 | .56749 |
| +0.2 | .57926 | .58317 | .58706 | .59095 | .59483 | .59871 | .60257 | .60642 |
| +0.3 | .61791 | .62172 | .62552 | .62930 | .63307 | .63683 | .64058 | .64431 |
| +0.4 | .65542 | .65910 | .66276 | .66640 | .67003 | .67364 | .67724 | .68082 |
| +0.5 | .69146 | .69497 | .69847 | .70194 | .70540 | .70884 | .71226 | .71566 |
| +0.6 | .72575 | .72907 | .73237 | .73565 | .73891 | .74215 | .74537 | .74857 |
| +0.7 | .75804 | .76115 | .76424 | .76730 | .77035 | .77337 | .77637 | .77935 |
| +0.8 | .78814 | .79103 | .79389 | .79673 | .79955 | .80234 | .80511 | .80785 |
| +0.9 | .81594 | .81859 | .82121 | .82381 | .82639 | .82894 | .83147 | .83398 |
| +1 | .84134 | .84375 | .84614 | .84849 | .85083 | .85314 | .85543 | .85769 |
| +1.1 | .86433 | .86650 | .86864 | .87076 | .87286 | .87493 | .87698 | .87900 |
| +1.2 | .88493 | .88686 | .88877 | .89065 | .89251 | .89435 | .89617 | .89796 |
| +1.3 | .90320 | .90490 | .90658 | .90824 | .90988 | .91149 | .91308 | .91466 |
| +1.4 | .91924 | .92073 | .92220 | .92364 | .92507 | .92647 | .92785 | .92922 |
| +1.5 | .93319 | .93448 | .93574 | .93699 | .93822 | .93943 | .94062 | .94179 |
| +1.6 | .94520 | .94630 | .94738 | .94845 | .94950 | .95053 | .95154 | .95254 |
| +1.7 | .95543 | .95637 | .95728 | .95818 | .95907 | .95994 | .96080 | .96164 |
| +1.8 | .96407 | .96485 | .96562 | .96638 | .96712 | .96784 | .96856 | .96926 |
| +1.9 | .97128 | .97193 | .97257 | .97320 | .97381 | .97441 | .97500 | .97558 |
| +2 | .97725 | .97778 | .97831 | .97882 | .97932 | .97982 | .98030 | .98077 |

$$12 + \frac{3.6}{\sqrt{16}} Z_{0.95} = 12 + 0.9 * 1.645 =$$

$$= 12 + 1.4805 = 13.4805$$

$13.7 > 13.4805 \quad \Rightarrow \quad H_0$ rejected

## two samples hypothesis test

- two random variables $X$ and $Y$ are normally distributed with the same variance and an hypotesis test is made on their means $\mu_X, \mu_Y$ with confidence level $1-\alpha$
  - $H_0$: $\mu_X = \mu_Y$
  - $H_1$: $\mu_X \neq \mu_Y$ (two sided alternative)
- random samples $S_X := \{x_1, x_2, \ldots, x_{n_X}\}, S_Y := \{y_1, y_2, \ldots, y_{n_Y}\}$ are observed with sample means $\bar{x}, \bar{y}$ and sample variances $\sigma_X, \sigma_Y$

it is possible to show that if $H_0$ is "true" ( $\mu_X = \mu_Y$) then

$$\Psi = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2(n_X-1)+\sigma_Y^2(n_Y-1)}{n_X+n_Y-2}\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}} \approx t_{n_X+n_Y-2}$$

$$P\{t^{\frac{\alpha}{2}}_{n_X+n_Y-2} \leq \Psi \leq t^{1-\frac{\alpha}{2}}_{n_X+n_Y-2}\} = 1-\alpha$$

$$\text{if } \begin{cases} t^{\frac{\alpha}{2}}_{n_X+n_Y-2} \leq \Psi \leq t^{1-\frac{\alpha}{2}}_{n_X+n_Y-2}, & H_0 \text{ not rejected} \\ \Psi < t^{\frac{\alpha}{2}}_{n_X+n_Y-2} \text{ or } \Psi > t^{1-\frac{\alpha}{2}}_{n_X+n_Y-2}, & H_0 \text{ rejected} \end{cases}$$

example:

two samples $X$ and $Y$ normally distributed with the same variance are observed...

$$\text{DATA: } \begin{cases} \bar{x} = 26.8 \\ \sigma_X = 1.7 \\ n_X = 30 \\ \bar{y} = 27.2 \\ \sigma_Y = 0.8 \\ n_Y = 30 \end{cases} \qquad \text{HYPOTHESES: } \begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \\ 1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \end{cases}$$

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2(n_X-1)+\sigma_Y^2(n_Y-1)}{n_X+n_Y-2}\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}} = \frac{26.8 - 27.2}{\sqrt{\frac{1.7^2(29)+0.8^2(29)}{58}\left(\frac{1}{30}+\frac{1}{30}\right)}} = -1.166$$

$t_{58}^{0.025}$...

$t_{58}^{0.975}$...

example:

two samples $X$ and $Y$ normally distributed with the same variance are observed...

$$\text{DATA: } \begin{cases} \bar{x} = 26.8 \\ \sigma_X = 1.7 \\ n_X = 30 \\ \bar{y} = 27.2 \\ \sigma_Y = 0.8 \\ n_Y = 30 \end{cases} \qquad \text{HYPOTHESES: } \begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \\ 1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \end{cases}$$

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2(n_X-1)+\sigma_Y^2(n_Y-1)}{n_X+n_Y-2}\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}} = \frac{26.8 - 27.2}{\sqrt{\frac{1.7^2(29)+0.8^2(29)}{58}\left(\frac{1}{30}+\frac{1}{30}\right)}} = -1.166$$

$$t_{58}^{0.025} \approx Z_{0.025} = -1.96$$

$$t_{58}^{0.975} \approx Z_{0.975} = 1.96$$

$$-1.96 < -1.166 < 1.96 \Rightarrow H_0 \text{ not rejected}$$

$\chi^2$ test is used for distribution of frequencies (categorical variables)

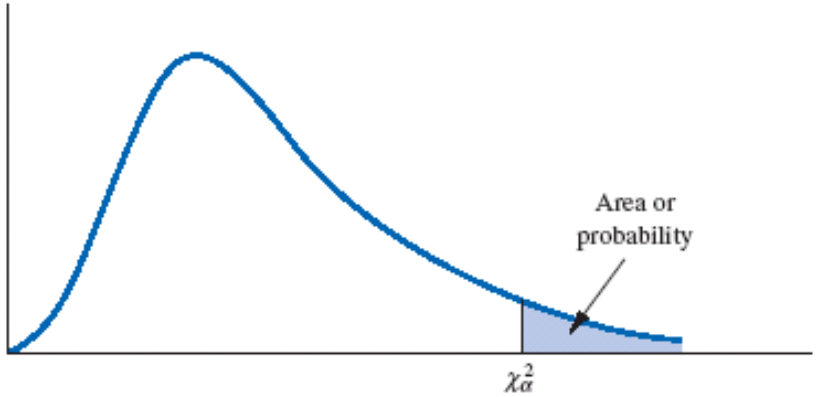
- a random variable $E$ can assume a finite set of $k$ values $E_1, E_2, \ldots, E_k$
- according to its probability distribution the expected frequencies of the values are $f_1, f_2, \ldots, f_k$ with $f_i \in (0, 1)$
- frequencies $o_1, o_2, \ldots, o_k$ are observed on a random sample of a random variable $X$
- $H_0$: $X = E$
  $H_1$: $X \neq E$

it is possible to show that if $H_0$ is true, then

$$\sum_{i=1,k} \frac{(o_i - f_i)^2}{f_i} \approx \chi^2_{k-1}$$

$\Rightarrow \chi^2$ table can be used to verify $H_0$

**TABLE 11.1** SELECTED VALUES FROM THE CHI-SQUARE DISTRIBUTION TABLE*




| Degrees of Freedom | Area in Upper Tail | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | .99 | .975 | .95 | .90 | .10 | .05 | .025 | .01 |
| 1 | .000 | .001 | .004 | .016 | 2.706 | 3.841 | 5.024 | 6.635 |
| 2 | .020 | .051 | .103 | .211 | 4.605 | 5.991 | 7.378 | 9.210 |
| 3 | .115 | .216 | .352 | .584 | 6.251 | 7.815 | 9.348 | 11.345 |
| 4 | .297 | .484 | .711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 |
| 5 | .554 | .831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.832 | 15.086 |
| 6 | .872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 |
| 7 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 |
| 8 | 1.647 | 2.180 | 2.733 | 3.490 | 13.362 | 15.507 | 17.535 | 20.090 |
| 9 | 2.088 | 2.700 | 3.325 | 4.168 | 14.684 | 16.919 | 19.023 | 21.666 |
| 10 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 |
| 11 | 3.053 | 3.816 | 4.575 | 5.578 | 17.275 | 19.675 | 21.920 | 24.725 |
| 12 | 3.571 | 4.404 | 5.226 | 6.304 | 18.549 | 21.026 | 23.337 | 26.217 |
| 13 | 4.107 | 5.009 | 5.892 | 7.041 | 19.812 | 22.362 | 24.736 | 27.688 |
| 14 | 4.660 | 5.629 | 6.571 | 7.790 | 21.064 | 23.685 | 26.119 | 29.141 |
| 15 | 5.229 | 6.262 | 7.261 | 8.547 | 22.307 | 24.996 | 27.488 | 30.578 |
| 16 | 5.812 | 6.908 | 7.962 | 9.312 | 23.542 | 26.296 | 28.845 | 32.000 |
| 17 | 6.408 | 7.564 | 8.672 | 10.085 | 24.769 | 27.587 | 30.191 | 33.409 |
| 18 | 7.015 | 8.231 | 9.390 | 10.865 | 25.989 | 28.869 | 31.526 | 34.805 |
| 19 | 7.633 | 8.907 | 10.117 | 11.651 | 27.204 | 30.144 | 32.852 | 36.191 |
| 20 | 8.260 | 9.591 | 10.851 | 12.443 | 28.412 | 31.410 | 34.170 | 37.566 |
| 21 | 8.897 | 10.283 | 11.591 | 13.240 | 29.615 | 32.671 | 35.479 | 38.932 |
| 22 | 9.542 | 10.982 | 12.338 | 14.041 | 30.813 | 33.924 | 36.781 | 40.289 |
| 23 | 10.196 | 11.689 | 13.091 | 14.848 | 32.007 | 35.172 | 38.076 | 41.638 |
| 24 | 10.856 | 12.401 | 13.848 | 15.659 | 33.196 | 36.415 | 39.364 | 42.980 |

example:
verify if a 6-sided dice is unbiased having observed the outcomes of 600 throwns with confidence level 99%.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| *observed* | 120 | 87 | 103 | 100 | 104 | 86 |
| *expected* | 100 | 100 | 100 | 100 | 100 | 100 |

HYPOTHESES:

$$\begin{cases} H_0 : \text{dice is unbiased} \\ H_1 : \text{dice is biased} \\ 1-\alpha = 0.99 \Rightarrow \alpha = 0.01 \end{cases}$$

$$\sum_{i=1,k} \frac{(o_i - f_i)^2}{f_i} = \frac{(\frac{120-100}{600})^2}{100/600} + \frac{(\frac{87-100}{600})^2}{100/600} + \frac{(\frac{103-100}{600})^2}{100/600} + \frac{(\frac{100-100}{600})^2}{100/600} +$$

$$+\frac{(\frac{104-100}{600})^2}{100/600} + \frac{(\frac{86-100}{600})^2}{100/600} = 0.013167$$

$\chi_5^{2\,0.01} = 15.086 > 0.013167 \Rightarrow H_0$ not rejected

# Resampling

practical limitations...

- often we don't know the distribution of the random variable
- often samples are not drawn at random

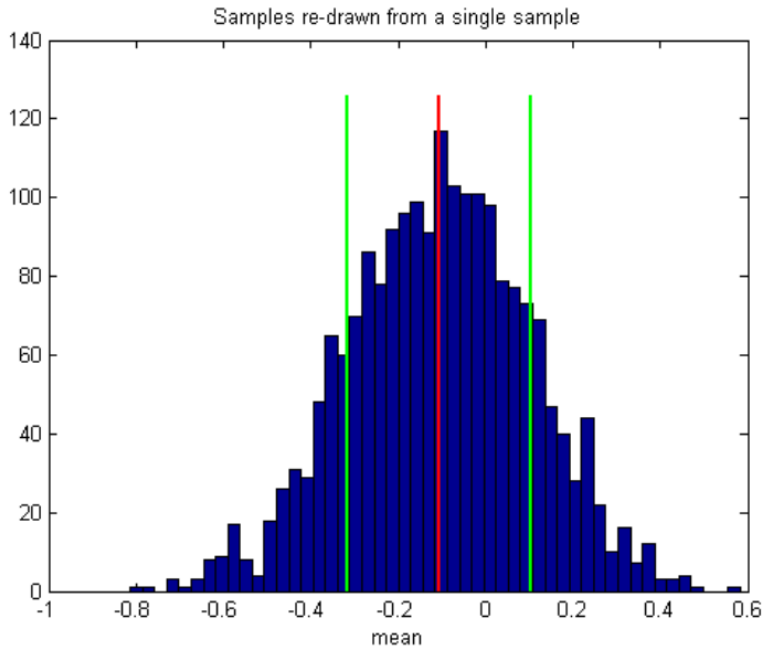
with time and computational power, empirical distributions can be built by resampling data, and used for

- confidence intervals
- hypothesis testing
- ...

## bootstrap

- we want to estimate the mean of a random variable with unknown distribution
- a $n$-dimensional sample $S := \{x_1, x_2, \ldots, x_n\}$ is observed
- many $m$-dimensional ($m \leq n$) sub-samples with repetition are extracted from $S$ and the mean is observed on each subsample to create an empirical distribution for the mean

$$\{x_2, x_5, x_{36}, x_1\} \rightarrow \mu_1$$
$$\{x_{21}, x_3, x_{12}, x_3\} \rightarrow \mu_2$$
$$\vdots$$
$$\{x_6, x_8, x_{15}, x_6\} \rightarrow \mu_k$$




- empirical distribution used to estimate confidence intervals or testing hypotheses

exercise 1.6

## reshuffling

- investigating the correlation of two random variables $X$ and $Y$
- Pearson's correlation coefficient $\rho_{XY} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_X \sigma_Y}$
- breaking the association structure between $X$ and $Y$ by randomly reshuffling many times one of the two samples and obtaining empirical distribution of the correlation to see if it is reliable

| | height | weight |
|---|---|---|
| person 1 | 1.875714232 | 109.8196777 |
| person 2 | 1.747060363 | 73.68895452 |
| person 3 | 1.882396677 | 96.58434842 |
| person 4 | 1.821966851 | 99.89928152 |
| person 5 | 1.774997615 | 93.68280948 |
| person 6 | 1.708226598 | 69.10431871 |
| person 7 | 1.747141064 | 83.50326143 |
| person 8 | 1.736052294 | 76.25888416 |
| person 9 | 1.702281321 | 79.87196594 |
| person 10 | 1.611794947 | 71.00545308 |
| person 11 | 1.80836271 | 84.7186362 |
| person 12 | 1.81967645 | 97.03849095 |
| person 13 | 1.64506476 | 75.87586733 |
| person 14 | 1.75978998 | 86.00856635 |
| person 15 | 1.758790799 | 84.64111228 |
| person 16 | 1.71819874 | 78.17286625 |
| person 17 | 1.839425242 | 88.99694187 |
| person 18 | 1.624947873 | 78.48909547 |
| person 19 | 1.768857521 | 84.43671674 |
| person 20 | 1.725574523 | 82.8216982 |
| | | |
| | corr | 0.818411356 |

| height | weight | height | weight | height | weight |
|---|---|---|---|---|---|
| 1.875714232 | 75.88 | 1.875714232 | 79.87 | 1.875714232 | 109.82 |
| 1.747060363 | 96.58 | 1.747060363 | 83.5 | 1.747060363 | 86.01 |
| 1.882396677 | 84.64 | 1.882396677 | 96.58 | 1.882396677 | 89 |
| 1.821966851 | 79.87 | 1.821966851 | 71.01 | 1.821966851 | 99.9 |
| 1.774997615 | 97.04 | 1.774997615 | 78.17 | 1.774997615 | 76.26 |
| 1.708226598 | 76.26 | 1.708226598 | 73.69 | 1.708226598 | 96.58 |
| 1.747141064 | 69.1 | 1.747141064 | 82.82 | 1.747141064 | 97.04 |
| 1.736052294 | 93.68 | 1.736052294 | 93.68 | 1.736052294 | 84.44 |
| 1.702281321 | 73.69 | 1.702281321 | 75.88 | 1.702281321 | 78.49 |
| 1.611794947 | 84.72 | 1.611794947 | 84.44 | 1.611794947 | 84.64 |
| 1.80836271 | 82.82 | 1.80836271 | 84.72 | 1.80836271 | 71.01 |
| 1.81967645 | 86.01 | 1.81967645 | 78.49 | 1.81967645 | 75.88 |
| 1.64506476 | 89 | 1.64506476 | 86.01 | 1.64506476 | 78.17 |
| 1.75978998 | 83.5 | 1.75978998 | 76.26 | 1.75978998 | 73.69 |
| 1.758790799 | 99.9 | 1.758790799 | 69.1 | 1.758790799 | 93.68 |
| 1.71819874 | 78.17 | 1.71819874 | 109.82 | 1.71819874 | 82.82 |
| 1.839425242 | 78.49 | 1.839425242 | 89 | 1.839425242 | 84.72 |
| 1.624947873 | 84.44 | 1.624947873 | 84.64 | 1.624947873 | 69.1 |
| 1.768857521 | 71.01 | 1.768857521 | 99.9 | 1.768857521 | 83.5 |
| 1.725574523 | 109.82 | 1.725574523 | 97.04 | 1.725574523 | 79.87 |
| | | | | | |
| | -0.129865955 | | -0.039975309 | | 0.384984573 |

## exercise 1.7

# Overview

# Multivariate statistics

some important questions...

- what data is?
- how we look at data?
- how we manipulate data?
- how we find things in data?

different combination of alternatives are possible!

data is composed by

- individuals/observations/units/samples
- features/characteristics/variables/attributes

minimal data requirements

- distinguishable individuals
- measurable features

many individuals, many features $\Rightarrow$ multivariate statistics

some assumptions...

- features can be measured on a continuous scale
- data is represented in a grid (matrix):
  rows = individuals
  columns = features
- matrix element $(i, j)$ is the measure of feature $j$ on individual $i$
- $m$ is the number of rows (individuals)
- $n$ is the number of columns (features)

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} \qquad x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{mi} \end{pmatrix} \qquad X = \left(x_1, x_2, \dots, x_n\right)$$

...data has been put into a "space"...

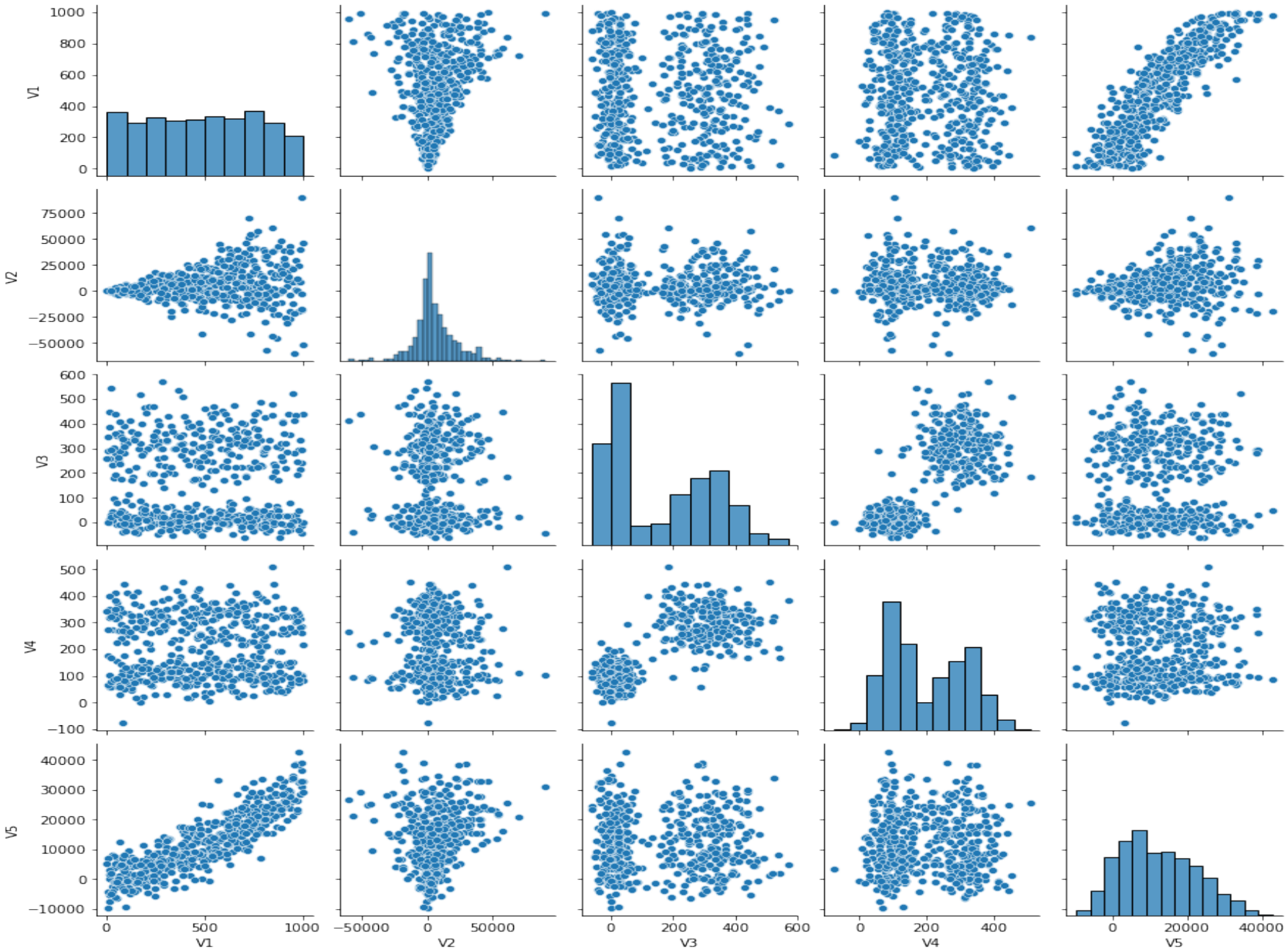
what do you see?

data represented in a space $\Rightarrow$ distance (metric) $d$

$d : X \times X \to [0, \infty)$ properties

1. $d(x, y) \geq 0$ (non-negativity)
2. $d(x, y) = 0 \iff x = y$ (identity of indiscernable)
3. $d(x, y) = d(y, x)$ (simmetry)
4. $d(x, y) + d(y, z) \geq d(x, z)$ (subadditivity)

exercise 2.1


V1
V2
V3
V4
V5
0
500
1000
−50000
0
50000
0
200
400
600
0
200
400
0
20000
40000

recall on correlation...

a statistical dependence (linear) between two random variables

- Pearson's correlation coefficient:

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X-\mu_x)(Y-\mu_y)\right]}{\sigma_X \sigma_Y}$$

- sample version:

$$r_{XY} = \frac{\frac{1}{n}\sum_{i=1,n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1,n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1,n}(y_i-\bar{y})^2}}$$

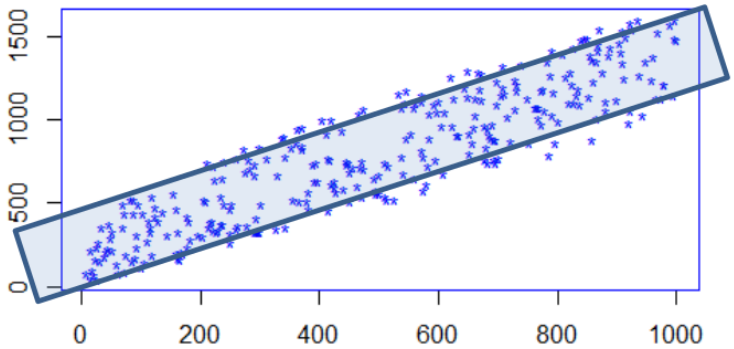
- $\rho_{XY}$ or $r_{XY} = \begin{cases} 1 & \text{positive correlation} \\ 0 & \text{linear independence} \\ -1 & \text{negative correlation} \end{cases}$

exercise 2.2

beyond correlation, much information in data is about distance between points

(the spread!)

- to extract information from data we have to focus on some aspects and discard the others
- data has many variables: is it possible to reduce the dimension of the data space, transform it, and preserve most of the distance between points? (e.g. humans like to see things in 2 dimensions)
- is it possible to change the system of coordinates to better capture the data variance?




$\Rightarrow$ orthogonality is a key aspect

- 2 vectors are linearly dependent if one can be obtained as a linear combination of the other

$$a = \begin{pmatrix} 2 \\ 1 \\ 4 \\ 3 \end{pmatrix}, b = \begin{pmatrix} 4 \\ 2 \\ 8 \\ 6 \end{pmatrix} \quad \Rightarrow \quad a = 0.5b$$

- $n$ vectors are linearly dependent if any of them can be obtained as a linear combination of the others

$$a = \begin{pmatrix} 2 \\ 1 \\ 4 \\ 3 \end{pmatrix}, b = \begin{pmatrix} 4 \\ 2 \\ 8 \\ 0 \end{pmatrix}, c = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 12 \end{pmatrix} \quad \Rightarrow \quad \begin{cases} a = 0.5b + 0.25c \\ c = 4a - 2b \end{cases}$$

- otherwise they are linearly independent

recall on linear algebra

- the inner product of 2 vectors $a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$, $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$ can be computed as

$$a^T b = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \|a\|\|b\| cos\alpha \quad (\alpha \text{ angle between } a \text{ and } b)$$

$$a = \begin{pmatrix} 2 \\ 1 \\ 4 \\ 3 \end{pmatrix}, b = \begin{pmatrix} 3 \\ 0 \\ 1 \\ 2 \end{pmatrix}, \; a^T b = \begin{pmatrix} 2\,1\,4\,3 \end{pmatrix} \times \begin{pmatrix} 3 \\ 0 \\ 1 \\ 2 \end{pmatrix} = 2\times3+1\times0+4\times1+3\times2 = 16$$

- 2 vectors are orthogonal if their inner product is zero (they are perpendicular)

$$a = \begin{pmatrix} 2 \\ 0 \\ -1 \\ 3 \end{pmatrix}, b = \begin{pmatrix} 0 \\ 6 \\ 3 \\ 1 \end{pmatrix}, \; a^T b = \begin{pmatrix} 2\ 0\ -1\ 3 \end{pmatrix} \times \begin{pmatrix} 0 \\ 6 \\ 3 \\ 1 \end{pmatrix} =$$

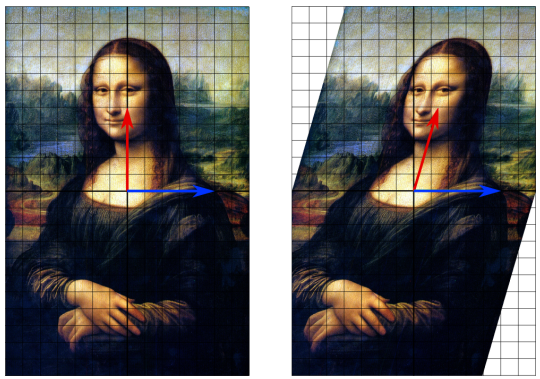
$$2 \times 0 + 0 \times 6 - 1 \times 3 + 3 \times 1 = 0$$

recall on linear algebra

- given a matrix $A^{m \times n}$, its rank $rk(A)$ is the maximum number of columns or rows linearly independent ($rk(A) \leq \min\{m, n\}$)
- given a square matrix $A^{n \times n}$, a scalar $\lambda$ and a vector $v \in \Re^n$ are an eigenvalue and an eigenvector if they satisfy

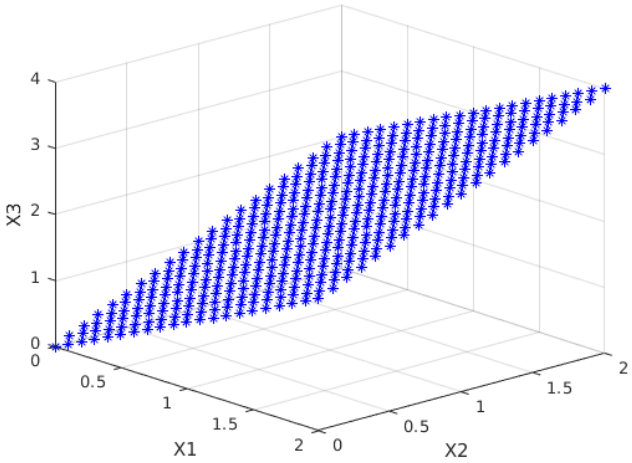
$$Av = \lambda v$$

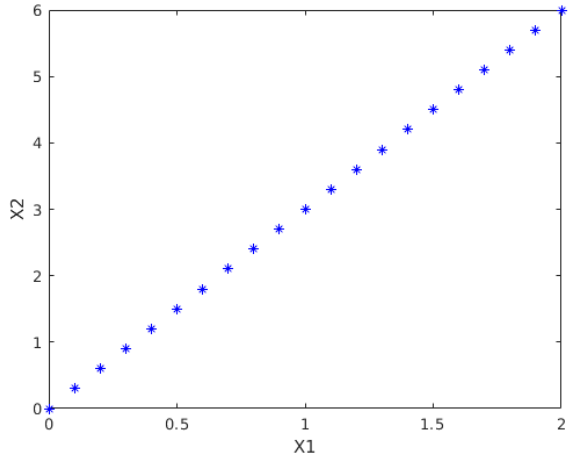
  - eigenvectors of $A$ are linearly independent
  - if $A$ is symmetric there are $n$ orthogonal eigenvectors (associated to $n$ eigenvalues)
  - provide an alternative coordinate system for the columns (rows) of $A$



why linear independence, orthogonality, matrix rank, eigenvalues/eigenvectors? vectors are measures of variables $\Rightarrow$ investigating vectors relations provide insights on data (dependence, redundancy,...)

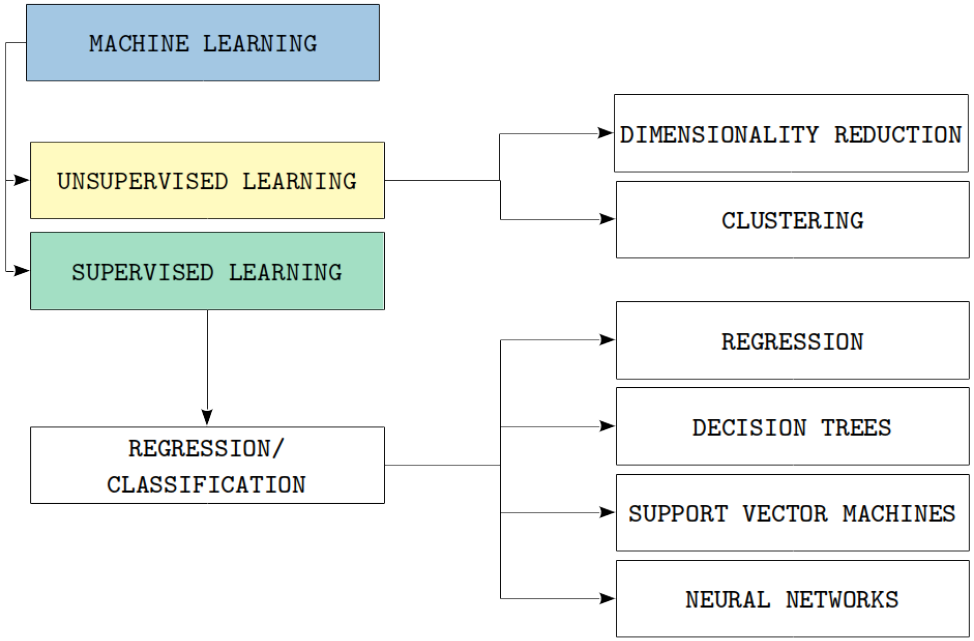
- a set of points with $n$ coordinates may lie in a space of dimension $\leq n$ (points in a plane can lie in a line, points in a 3d-space can lie in a plane...)




- given a set of $n$ dimensional points one may wonder what is the space of minimum dimension where they lie
- in such a space their mutual distance would be preserved
- maybe we are happy with something "less precise", ignoring the less relevant dimensions!

# Principal Component Analysis (PCA)




PCA is a dimensionality reduction technique used for

- exploratory data analysis (unsupervised learning)
- feature extraction (support to supervised learning)

context

- our data are represented with $X^{m \times n}$

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} \qquad x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{mi} \end{pmatrix} \qquad X = (x_1, x_2, \dots, x_n)$$

- data are in a $n$-dimensional space with coordinates $x_1, x_2, \dots, x_n$ (features)
- we seek a new system of orthogonal coordinates $z_1, z_2, \dots, z_p$ (principal components) with $p \leq n$ to represent orginal data such that
  1. it better captures the variance of data $\Rightarrow$ if $p < n$ the information loss is "minimal"
  2. each new coordinate $z_i$ is a linear combination of all the original coordinates

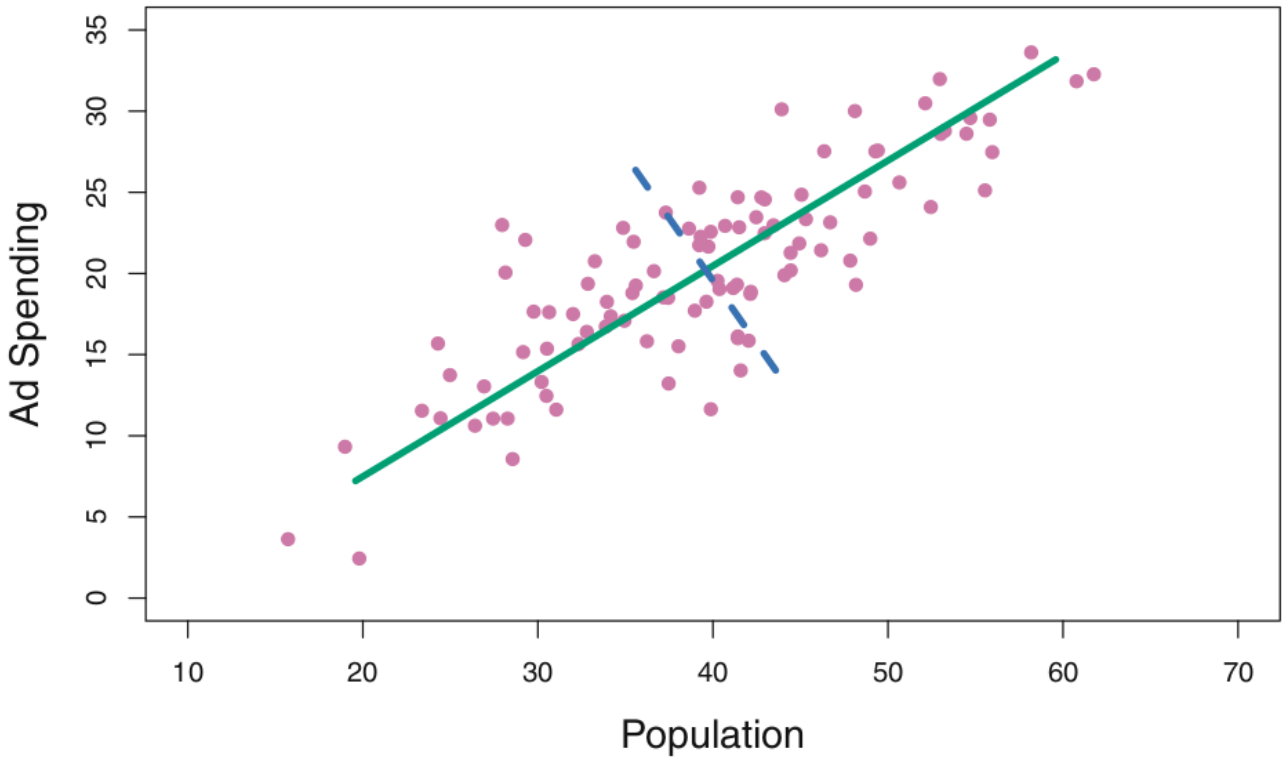
$$\begin{aligned} z_1 &= \phi_{11}x_1 + \phi_{21}x_2 + \dots + \phi_{n1}x_n \\ z_2 &= \phi_{12}x_1 + \phi_{22}x_2 + \dots + \phi_{n2}x_n \\ &\vdots \\ z_p &= \phi_{1p}x_1 + \phi_{2p}x_2 + \dots + \phi_{np}x_n \end{aligned} \qquad (x_1, x_2, \dots, x_n) \Rightarrow (z_1, z_2, \dots, z_p)$$

  3. thay can be ranked in order of "importance"

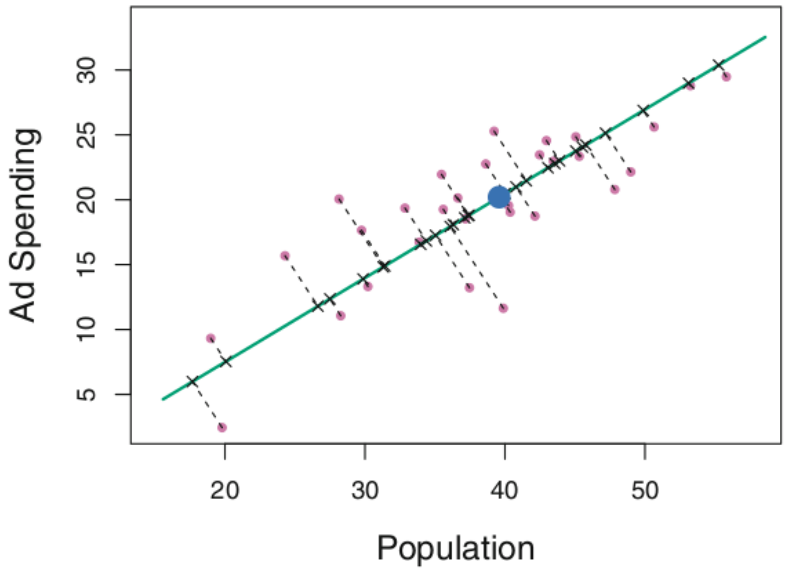
geometric interpretation

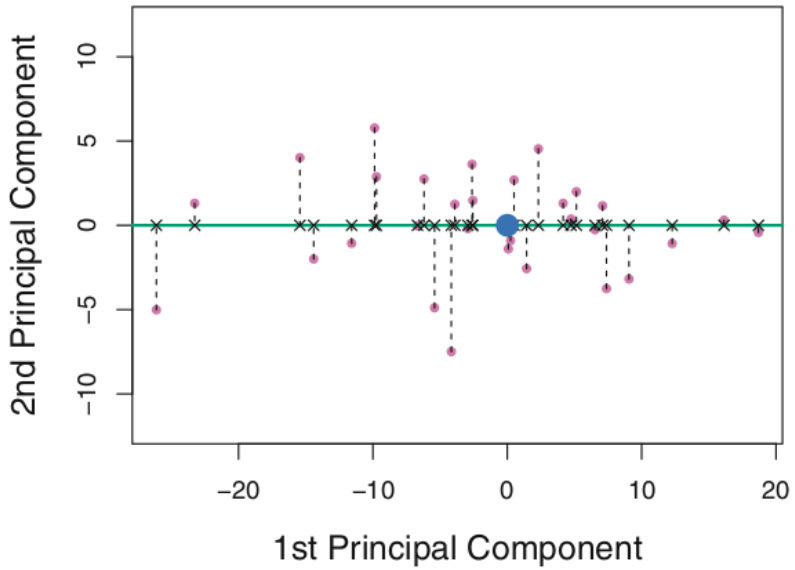
"Principal components are a sequence of projections of the data, mutually uncorrelated and ordered in variance."
('The Elements of Statistical Learning' page 534)

example: population size (feature 1 or $x_1$) and ad spending (feature 2 or $x_2$) over 100 cities (individuals)



geometric interpretation




$$\begin{aligned} z_1 &= \phi_{11}x_1 + \phi_{21}x_2 \\ z_2 &= \phi_{12}x_1 + \phi_{22}x_2 \end{aligned} \quad \Rightarrow \quad \begin{aligned} z_1 &= \phi_{11}\text{pop} + \phi_{21}\text{ad} \\ z_2 &= \phi_{12}\text{pop} + \phi_{22}\text{ad} \end{aligned}$$

coordinates $(z_1, z_2)$ (Principal Component scores) determine the position of individuals in the new system!

$\begin{pmatrix} \phi_{11} \\ \phi_{21} \end{pmatrix}, \begin{pmatrix} \phi_{12} \\ \phi_{22} \end{pmatrix}$: loadings vectors associated to 1st and 2nd Principal Components

$$z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}})$$

mathematical interpretation

we sequentially determine new coordinates (directions) maximizing the overall variance of data

- center data by columns as $x_i = x_i - \bar{x}_i \Rightarrow \bar{x}_i = 0$
- determine the first Principal Component by solving

$$\max_{\phi_{11},\phi_{21},\ldots,\phi_{n1}} \quad \frac{1}{m} \sum_{i=1,m} \sum_{j=1,n} (\phi_{j1} x_{ij} - \bar{x}_j)^2 = \frac{1}{m} \sum_{i=1,m} \sum_{j=1,n} (\phi_{j1} x_{ij})^2$$

$$\text{subject to} \qquad \sum_{j=1,n} \phi_{j1}^2 = 1$$

- determine the second Principal Component by solving

$$\max_{\phi_{12},\phi_{22},\ldots,\phi_{n2}} \quad \frac{1}{m} \sum_{i=1,m} \sum_{j=1,n} (\phi_{j2} x_{ij})^2$$

$$\text{subject to} \qquad \sum_{j=1,n} \phi_{j2}^2 = 1$$
$$\sum_{j=1,n} \phi_{j1}\phi_{j2} = 0$$

- ...

## covariance interpretation

- center data by columns as $x_i = x_i - \bar{x}_i \Rightarrow \bar{x}_i = 0$
- construct the covariance matrix as

$$C = \begin{pmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \dots & cov(x_1, x_n) \\ cov(x_2, x_1) & cov(x_2, x_2) & \dots & cov(x_2, x_n) \\ \vdots & \vdots & \dots & \vdots \\ cov(x_n, x_1) & cov(x_n, x_2) & \dots & cov(x_n, x_n) \end{pmatrix}$$

$C$ symmetric, positive semidefinite

$$\Rightarrow \begin{cases} n \text{ orthogonal eigenvectors } v_1, v_2, \dots, v_n \\ n \text{ nonnegative eigenvalues } \lambda_1 > \lambda_2 > \dots > \lambda_n \geq 0 \end{cases}$$

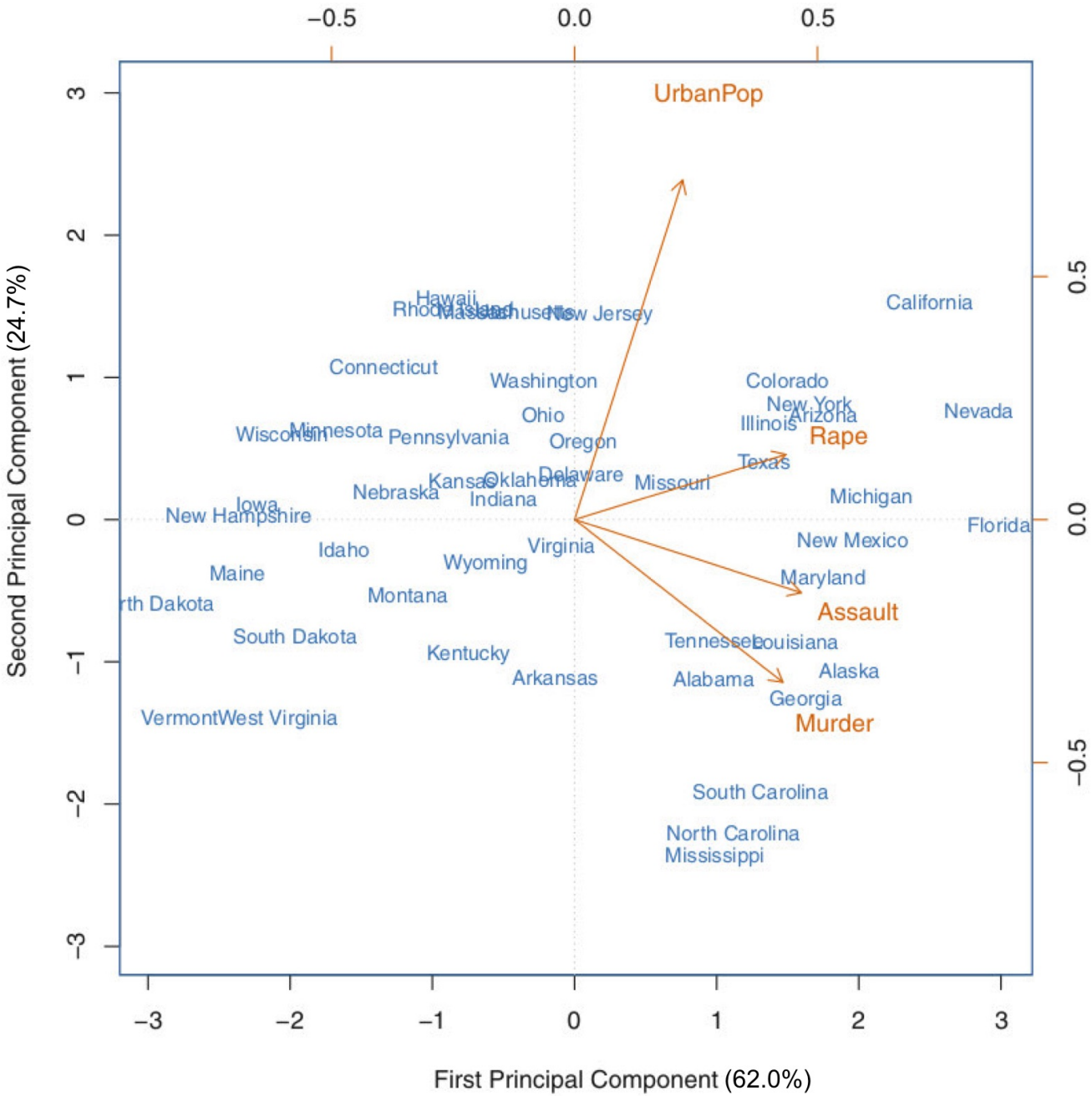
- eigenvectors represent the Principal Components

$$v_1 = \begin{pmatrix} \phi_{11} \\ \phi_{21} \\ \vdots \\ \phi_{n1} \end{pmatrix}, v_2 = \begin{pmatrix} \phi_{12} \\ \phi_{22} \\ \vdots \\ \phi_{n2} \end{pmatrix}, \dots, v_n = \begin{pmatrix} \phi_{1n} \\ \phi_{2n} \\ \vdots \\ \phi_{nn} \end{pmatrix}$$

proportion of variance explained (PVE) by Principal Component $i$

$$\frac{\lambda_i}{\sum_{j=1,n} \lambda_j}$$

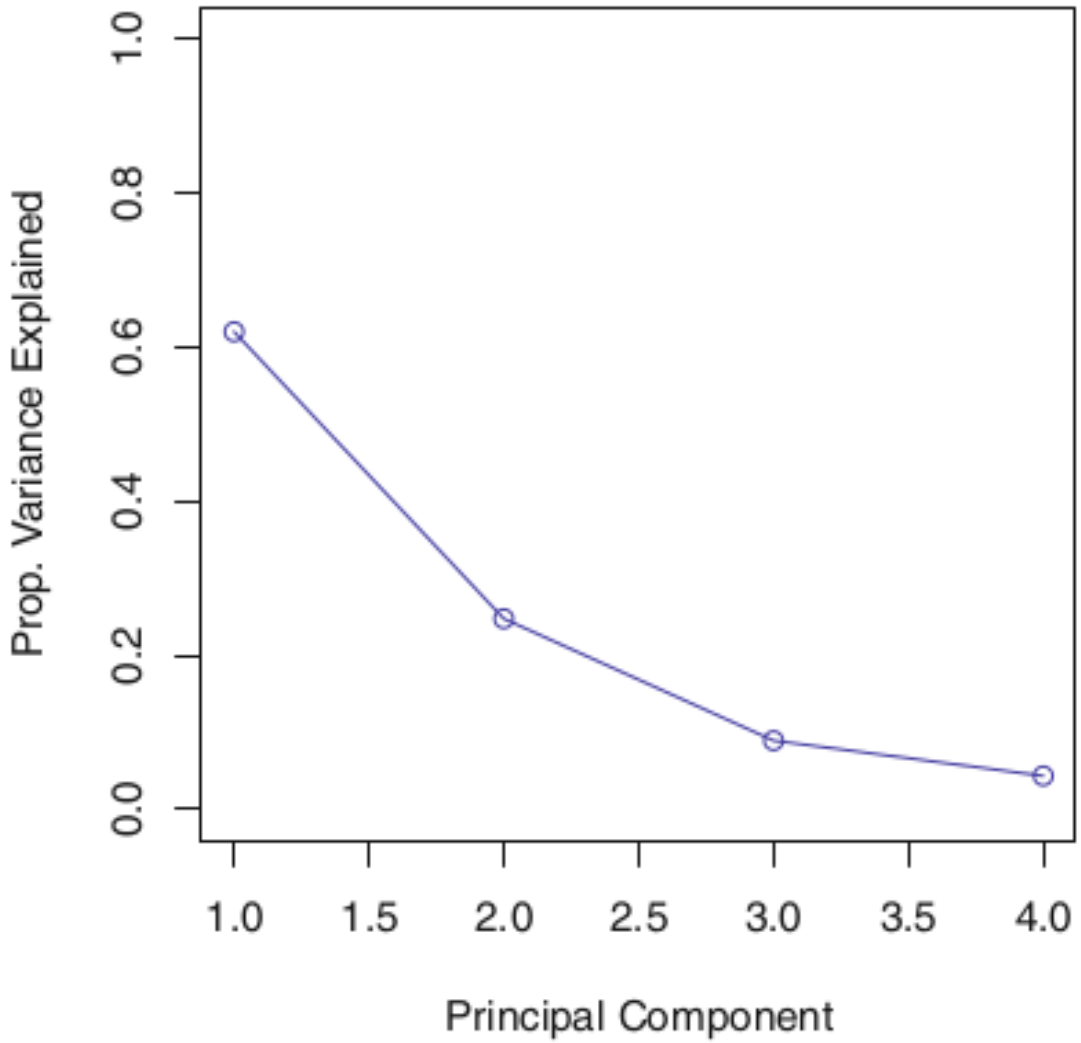
# biplot




example:
Assault, Murder, Rape and Urban Population measured on 50 US states

| | PC1 | PC2 |
|---|---|---|
| Murder | 0.5358995 | −0.4181809 |
| Assault | 0.5831836 | −0.1879856 |
| UrbanPop | 0.2781909 | 0.8728062 |
| Rape | 0.5434321 | 0.1673186 |

scree plot



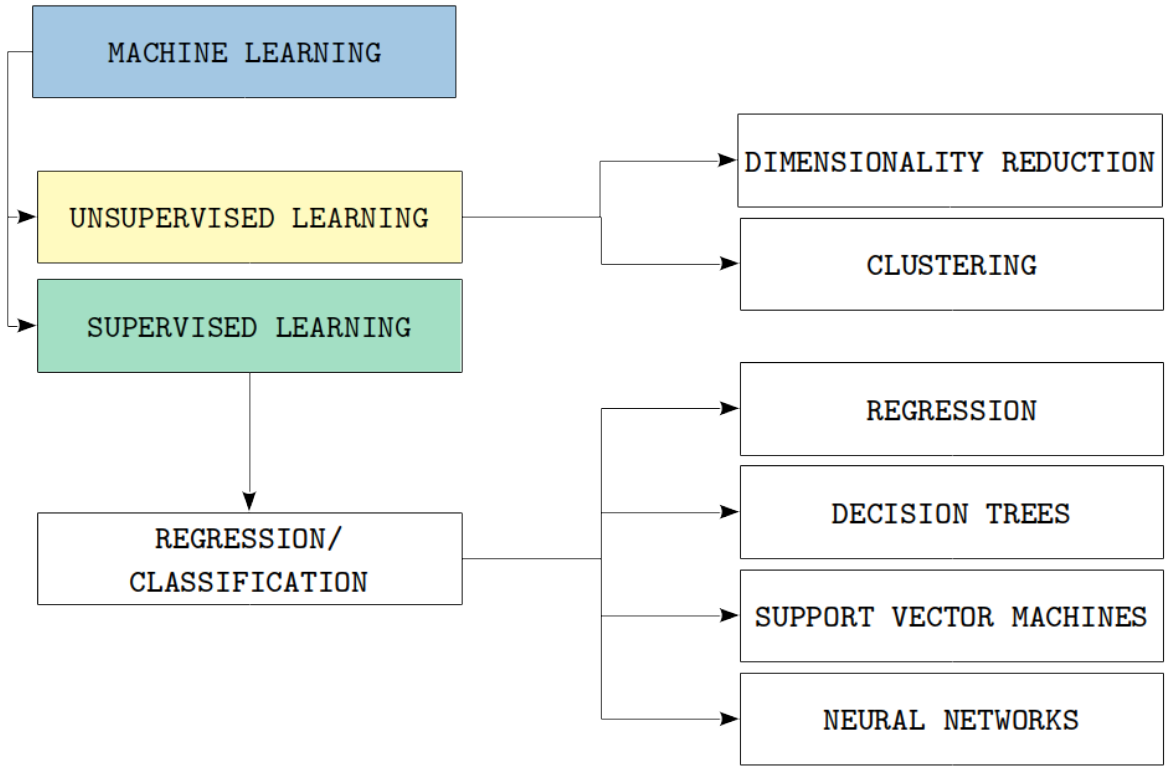

exercise 2.3

# references

- 'An Introduction to Statistical Learning', Section 6.3 (pages 228-233), 10.2
- 'The Elements of Statistical Learning', Section 14.5.1

# Overview




MACHINE LEARNING
UNSUPERVISED LEARNING
SUPERVISED LEARNING
REGRESSION/
CLASSIFICATION
DIMENSIONALITY REDUCTION
CLUSTERING
REGRESSION
DECISION TREES
SUPPORT VECTOR MACHINES
NEURAL NETWORKS

# unsupervised vs supervised learning

- unsupervised learning

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix}$$

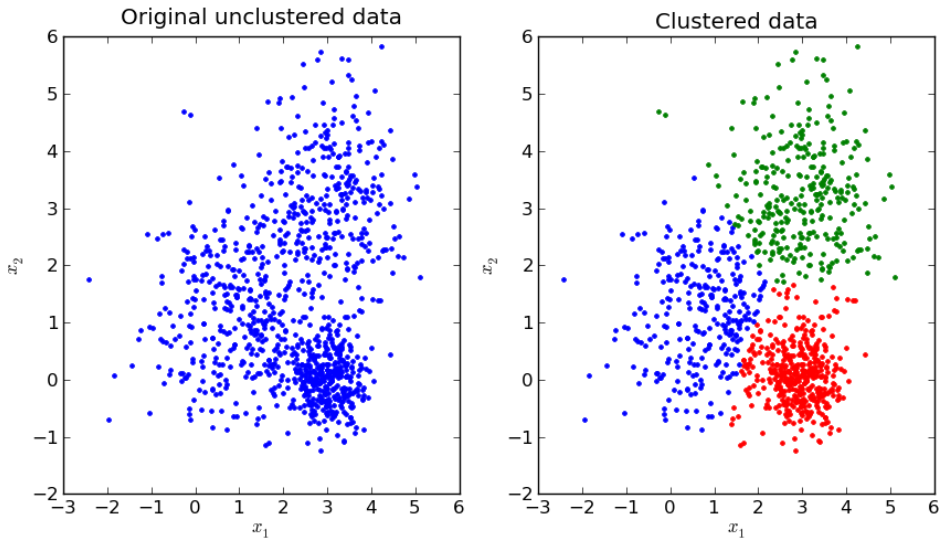
task: to discover interesting things about data without knowing the true answer

- supervised learning

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \dots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

task: prediction of the targets

# clustering




- find clusters (subgroups) of observations having similar features
- clusters $C_i, i = 1, \ldots, K$ are a partition of the observations:

$$C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, m\}, \quad C_i \cap C_j = \emptyset, \forall i \neq j$$

- when observations are similar (dissimilar)? domain-specific
- PCA and clustering simplify data via smaller summeries
  - PCA via low-dimension features representations
  - clustering via homogenous subgroups among observations

clustering methods

- hierarchical clustering:

  the number of clusters is iteratively modified

- K-means clustering:

  the number of clusters is decided a priori

- fuzzy clustering:

  observations are assigned to all clusters with a certain probability (not to a single one)

# hierarchical agglomerative clustering

- idea: iteratively aggregate $m$ initial clusters (one per observations) to a final single cluster (including all observations)
- dendrogram: attractive tree-based representation of observations and clusters
- the number of clusters is decided ex post
- requires
  - a point to point distance (dissimilarity) $d(x_i, x_j)$ with $x_i, x_j \in \Re^n$
  - a cluster to cluster distance (linkage) $D(C_s, C_t)$ with $s, t \in \{1, \ldots, K\}$

# agglomerative clustering scheme and dendrogram

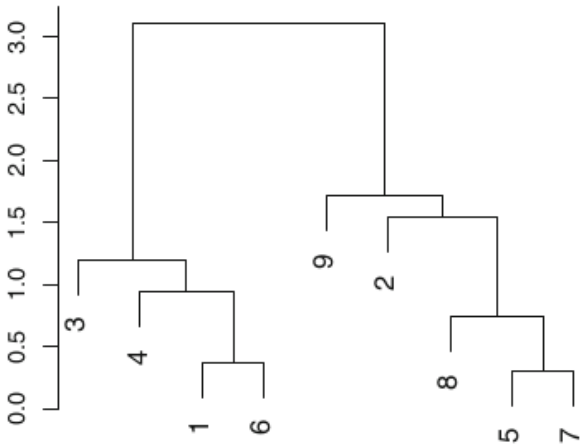
0 select dissimilarity $d$ and linkage $D$

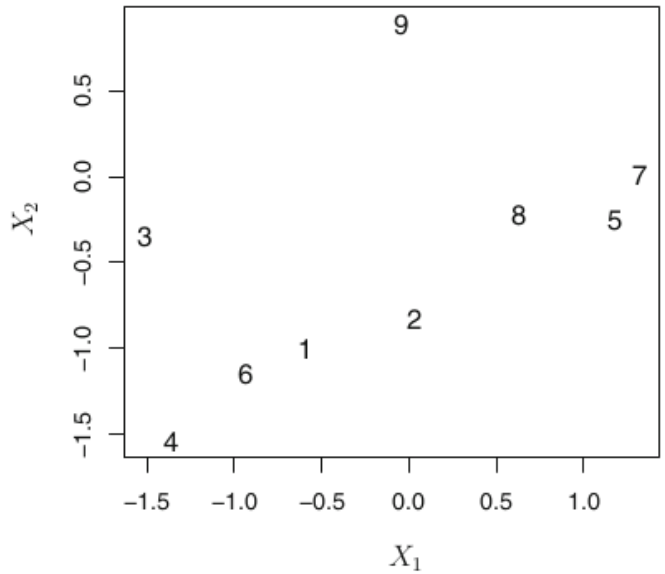
1 set the $m$ initial clusters as the $m$ observations and compute the $\binom{m}{2} = m(m-1)/2$ pairwise dissimilarities

2 for $i = m, m-1, \ldots, 2$

a identify the pair of closest clusters according to $D$ (or $d$ if single observations) and fuse them

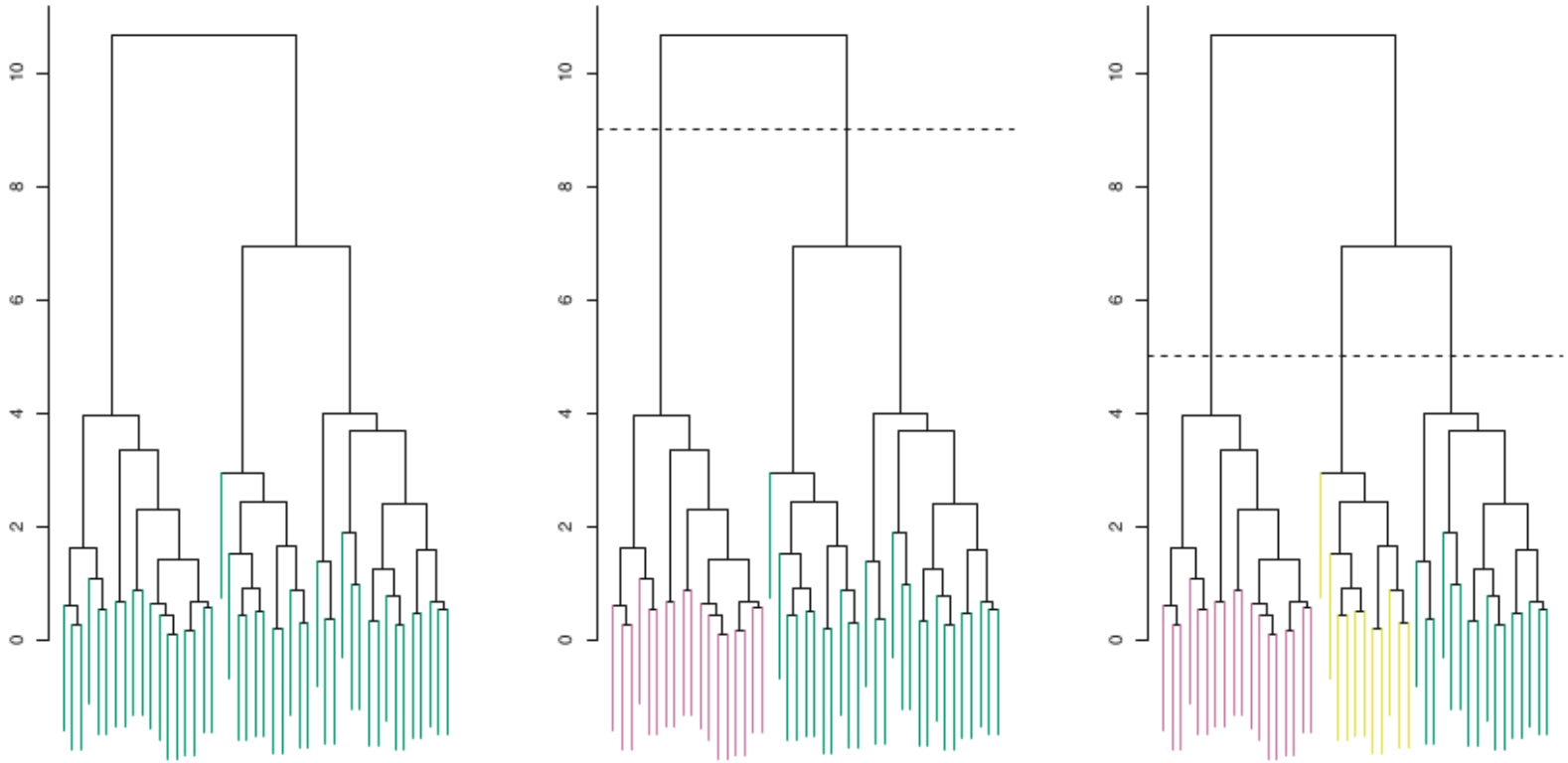
b compute the pairwise linkage $D$ for the remaining $i-1$ clusters



how to select the final clustering?

- make horizontal cut on the dendrogram
- subgroups below the cut are the clusters
- number of clusters = number of vertical lines crossed by the cut (height of the cut determine the number of clusters $K$)



## linkage choices

- SINGLE LINK: $D(C_s, C_t) = \min_{x_i \in C_s, x_j \in C_t} d(x_i, x_j)$
- COMPLETE LINK: $D(C_s, C_t) = \max_{x_i \in C_s, x_j \in C_t} d(x_i, x_j)$
- AVERAGE LINK: $D(C_s, C_t) = \frac{1}{|C_s||C_t|} \sum_{x_i \in C_s, x_j \in C_t} d(x_i, x_j)$
- CENTROID LINK: $D(C_s, C_t) = d(\bar{x}_s, \bar{x}_t), \quad \bar{x}_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$

example: apply the agglomerative clustering with single link to the 5 points characterized by this pairwise Euclidean distances matrix.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | | | | |
| 2 | 9 | 0 | | | |
| 3 | 3 | 7 | 0 | | |
| 4 | 6 | 5 | 9 | 0 | |
| 5 | 11 | 10 | 2 | 8 | 0 |

| | $\{1,3,5\}$ | 2 | 4 |
|---|---|---|---|
| $\{1,3,5\}$ | 0 | | |
| 2 | 7 | 0 | |
| 4 | 6 | 5 | 0 |

| | 1 | 2 | $\{3,5\}$ | 4 |
|---|---|---|---|---|
| 1 | 0 | | | |
| 2 | 9 | 0 | | |
| $\{3,5\}$ | 3 | 7 | 0 | |
| 4 | 6 | 5 | 8 | 0 |

| | $\{1,3,5\}$ | $\{2,4\}$ |
|---|---|---|
| $\{1,3,5\}$ | 0 | |
| $\{2,4\}$ | 6 | 0 |

some considerations on agglomerative clustering

good aspects:

- provides a nested sequence of clusterings (more informative)

bad aspects:

- sensitive to anomalous points/outliers
- irreversible: "bad" early fusions affect the whole structure of nested clusters
- arbitrary fusion choices of equally distant clusters affect the rest of the sequence
- time consuming for large datasets

## K-means clustering

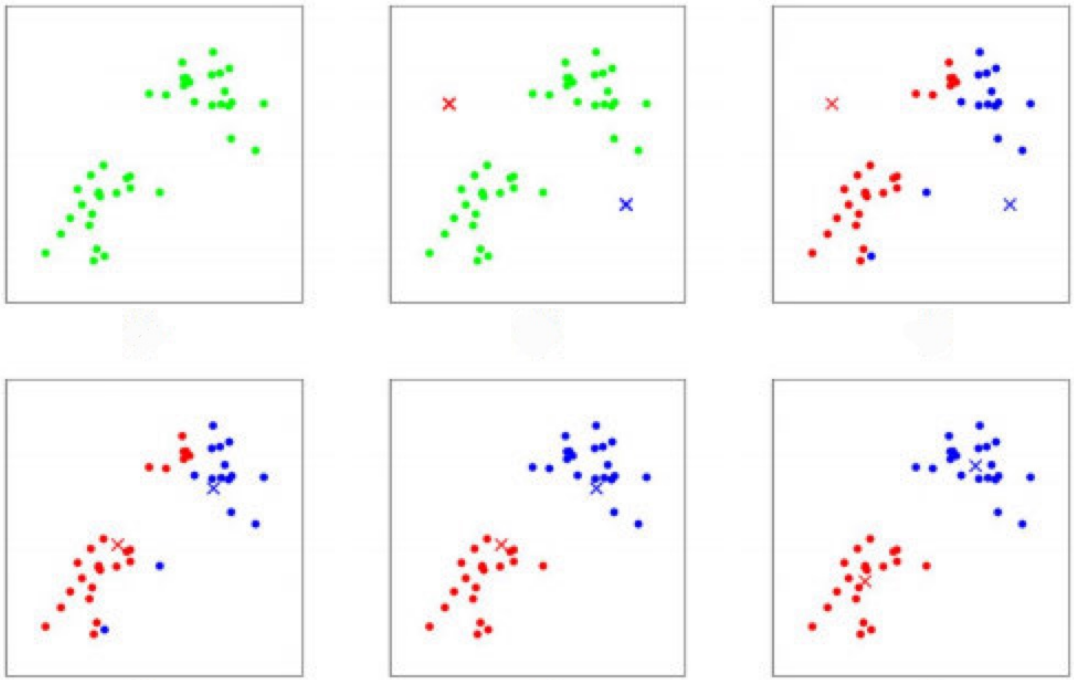
- the number of clusters $K$ is given
- idea: a "good" clustering is one for which the within-cluster variation is a small as possible
- each cluster $C_k$ is characterize by its centroid $\bar{x}_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$
- tends to minimize the overall distance between each point and its clusters' centroid:

$$\min_{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K} W(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K) = \sum_{k=1,K} \sum_{x_j \in C_k} d(x_j, \bar{x}_k)$$

- the algorithm is an alternation of two main steps
  a. assign the points to the cluster with the closest centroid
  b. determine the centroids of each cluster
- the algorithm stops when the two steps do not change the assignment (or do not improve $W(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K)$) anymore

## K-means clustering scheme

0. select a distance $d$ and a number $K$
1. (initialization) randomly determine $K$ centroids $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K \in \Re^n$
2. iterate until the cluster assignment stop changing
   a. (cluster assignments) assign each observation to the closest centroid
   b. (centroids update) update the centroids of each cluster as $\bar{x}_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$



## some considerations on K-means

- sensitive to anomalous points/outliers
- points can move from a cluster to another
- converges to a sub-optimal solution $\Rightarrow$ the final solution depends on the random intialization
- if points have not a strong "clustered" structure or $K$ is "wrong" the algorithm is more sensitive to initialization
- good practice: mulitple runs of K-means for different initialization and different $K$ and select the best solution (in terms of $W(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K)$)

how to verify the quality of a clustering

have we obtained relevant structures? is $K$ right?

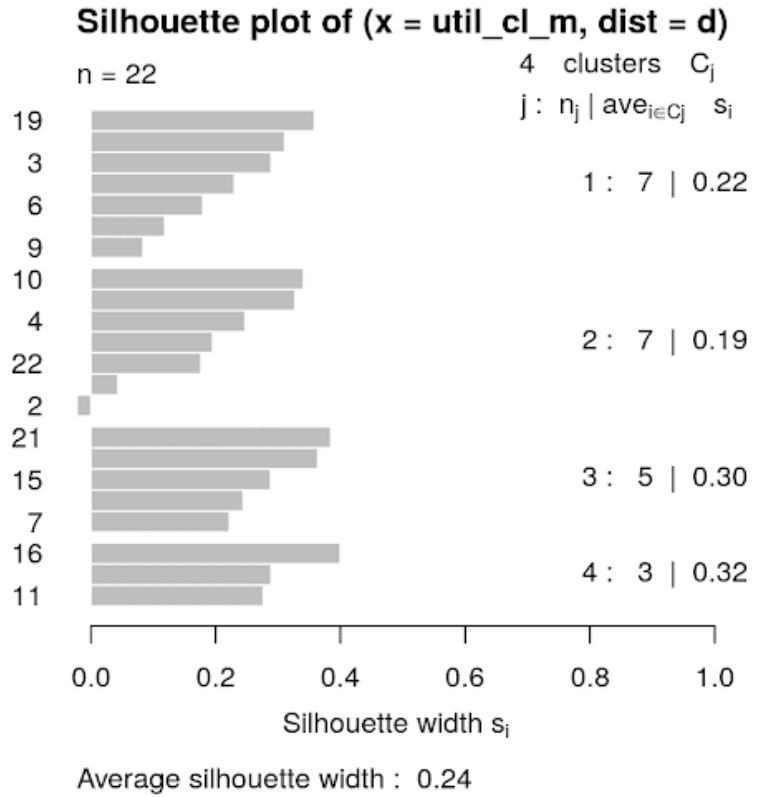
we can perform computational assessments

- by perturbating data and verifying if the clustering is stable
  - add noise to data
  - delete points (resample without replacement)
  - bootstrapping (resemple with replacement)
- use the silhouette measure and plot...

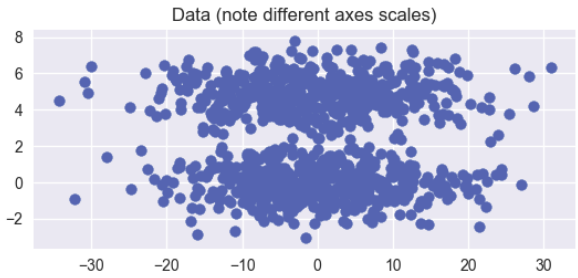
the silhouette of clusters

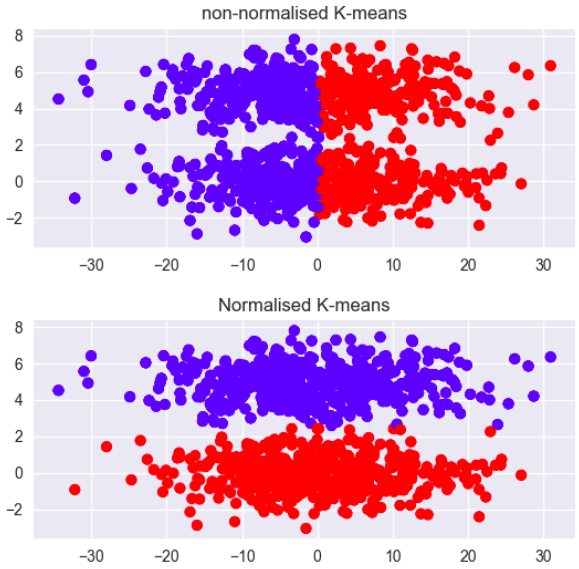
suppose point $i$ belongs to $C_i$

- mean distance from $x_i$ to points of cluster $C$
  $d_{iC} = \frac{1}{|C|} \sum_{x_j \in C} d(x_i, x_j)$
- mean distance from $x_i$ to points of its cluster
  $a_i = d_{iC_i}$
- minimum mean distance from $x_i$ to points of another cluster
  $b_i = \min_{C \neq C_i} d_{iC}$
- silhouette of point $i$
  $SIL_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$
- overall silhouette
  $SIL = \frac{1}{m} \sum_{i=1,m} SIL_i$



# the issue of scaling data




features with different locations or scales may influence the distance of two points (e.g. age and salary) and bias clustering results (the same holds for PCA!)

often it is better to center and/or scale features

- Z-score scaler: $\frac{x-\bar{x}}{\sigma} \Rightarrow$ mean=0, st.dev.=1
- unit scaler: $\frac{x-\min_x}{\max_x - \min_x} \Rightarrow$ values in the interval $[0, 1]$

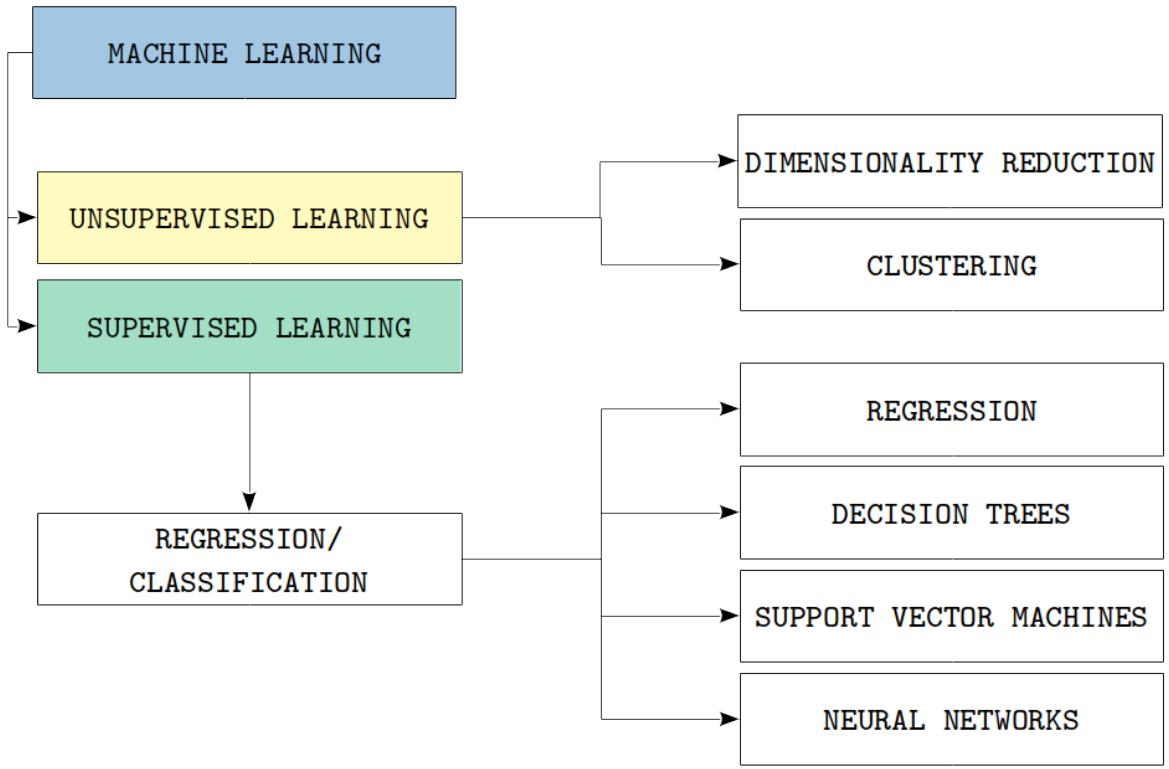
exercise 3.1

# references

- 'An Introduction to Statistical Learning', Section 10.3

# Overview






supervised learning $\equiv$ learning with a teacher

idea: observe many features-target pairs to reconstruct the function mapping features to target

# supervised learning

- consider a process producing target $y$ in correspondence to features $x \in \Re^n$ according to an unknown function

$$y = f(x)$$

- supervised learning determines a model $m(\cdot)$ approximating $f(\cdot)$ on the basis of a set of known features-target pairs denoted as training set ($TR$)

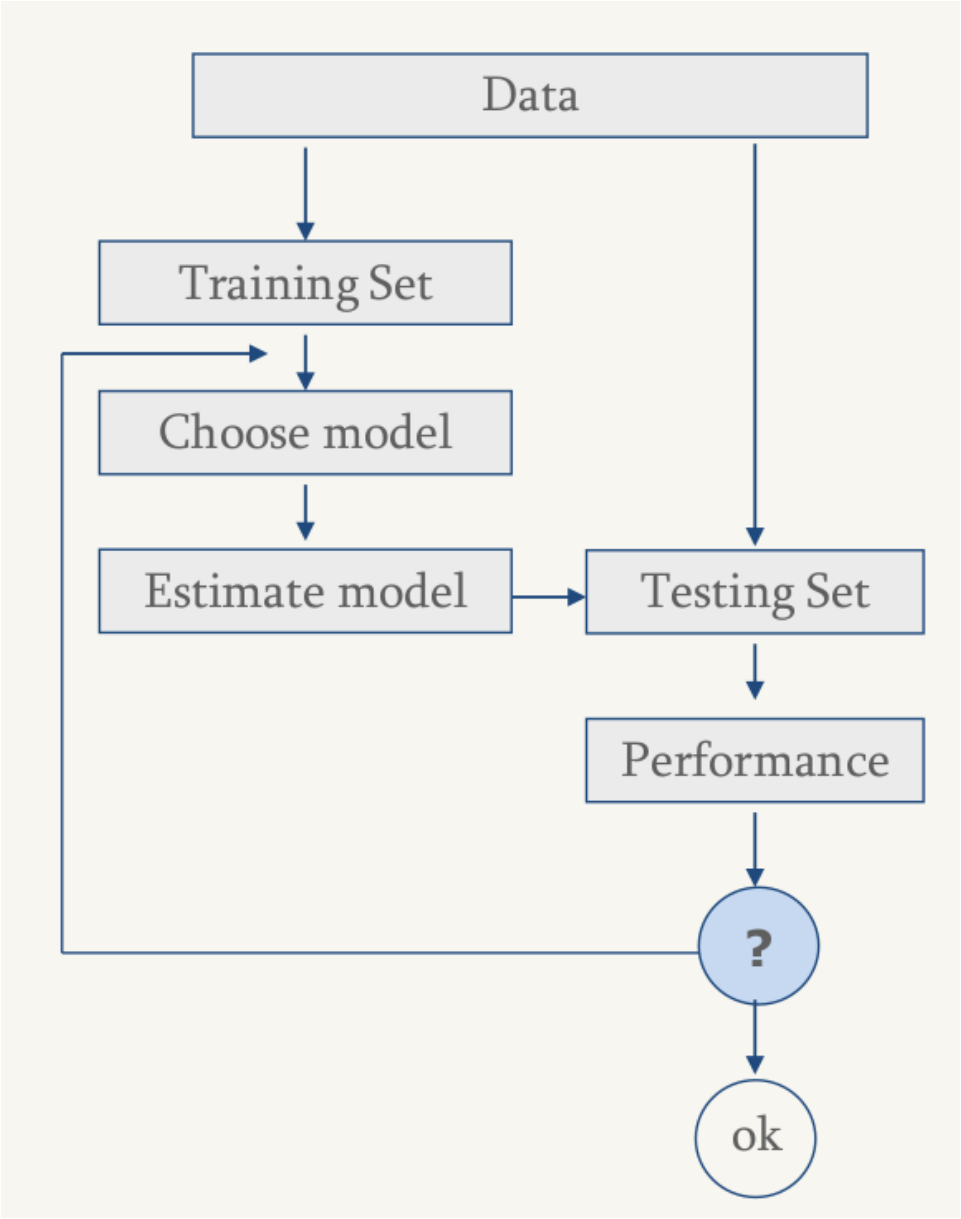
$$TR = \{(x^i, y^i) : x^i \in \Re^n, y^i \in ?, i = 1, \ldots, m\}$$

- in the training phase the parameters of $m(\cdot)$ are tuned so that the model fits "as much as possible" $TR$

- once trained the model is tested on a testing set ($TS$) of known features-target pairs not used in the training, and if "sufficiently good" it can be used to predict the target $\tilde{y}$ for unseen observations $\tilde{x}$ according to

$$\tilde{y} = m(\tilde{x})$$

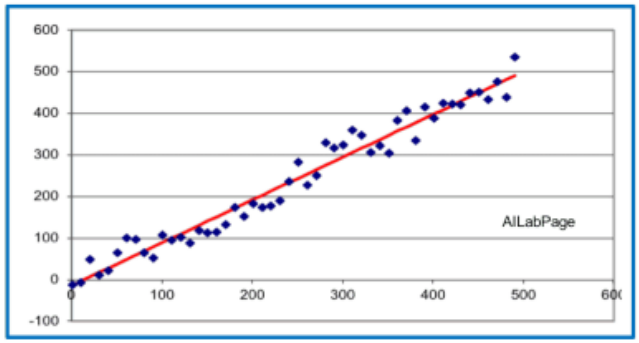
# supervised learning scheme



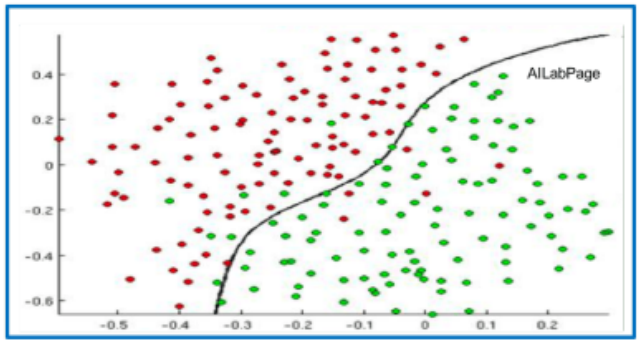

- generalization: capability of the trained model $m(\cdot)$ to correctly predict the target of unseen observations
- overfitting: when $m(\cdot)$ fits "too much" the training data and it is not able to make good predictions for unseen observations

- the form of $f(\cdot)$ may be of any type
  - mathematical function
  - logic formula
  - the result of an algorithm
  - a black-box system
  - ...
- the form of $m(\cdot)$ may be of different types
  - mathematical function
  - a structured object (e.g., a tree, a neural network,...)
  - a set of rules
  - ...
- learning tasks categorized according to the domain of target $y$
  - $y$ continuous $\Rightarrow$ regression
  - $y$ discrete $\Rightarrow$ classification (binary or multi-class)

regression




binary classification



many issues in supervised learning...

- how to select the family of functions of model $m(\cdot)$
- how to determine hyperparameters of $m(\cdot)$ (model parameters which cannot be tuned during the training phase and must be set a priori)
- which algorithm to adopt in the training phase
- how to split training and testing sets
- how to measure the performance of the trained model
- how to choose/control the model complexity (complex models fit training data well, simple models tend to better generalize)

... many alternative machine learning approaches

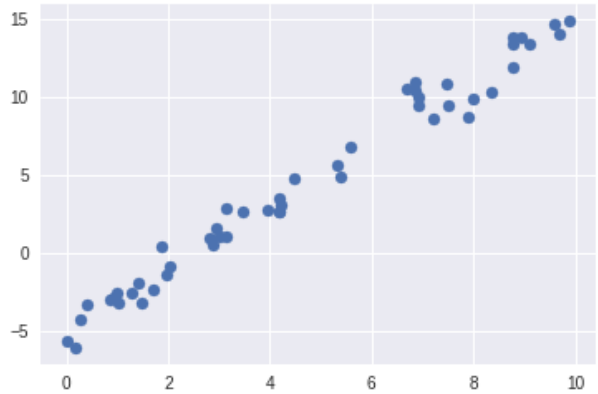
# linear regression

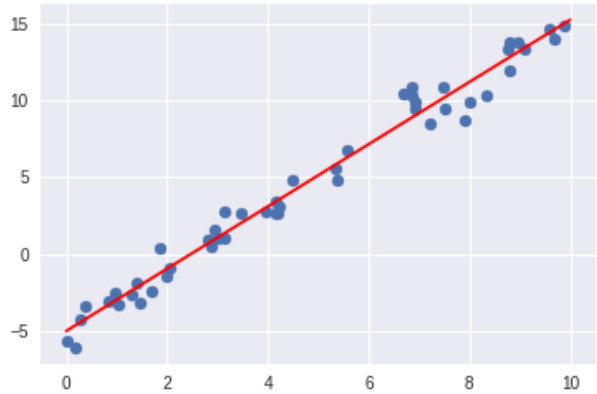
linear regression fast and interpretable

simple linear regression model (single feature)

$$y = \beta_0 + \beta_1 x \text{ (regression line)}$$

- $y \in \Re$ (regression task)
- $\beta_0$ intercept
- $\beta_1$ slope




training a simple linear regression consists in determinining $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the regression line fits better the points

training simple linear regression

- given a $TR$ of pairs $(x^i, y^i)$ with $i = 1, \ldots, m$, find estimates $\hat{\beta}_0, \hat{\beta}_1$ such that

$$\hat{\beta}_0 + \hat{\beta}_1 x^i \approx y^i \quad \text{with } i = 1, \ldots, m$$

- given $\hat{\beta}_0, \hat{\beta}_1$, the prediction $\hat{y}^i$ for point $x^i$ is

$$\hat{y}^i = \hat{\beta}_0 + \hat{\beta}_1 x^i$$

and the $i$-th error (residual) is computed as

$$e^i = y^i - \hat{y}^i$$

- $\hat{\beta}_0, \hat{\beta}_1$ are obtained by solving the following optimization problem where a loss function called residual sum of squares ($RSS$) is minimized

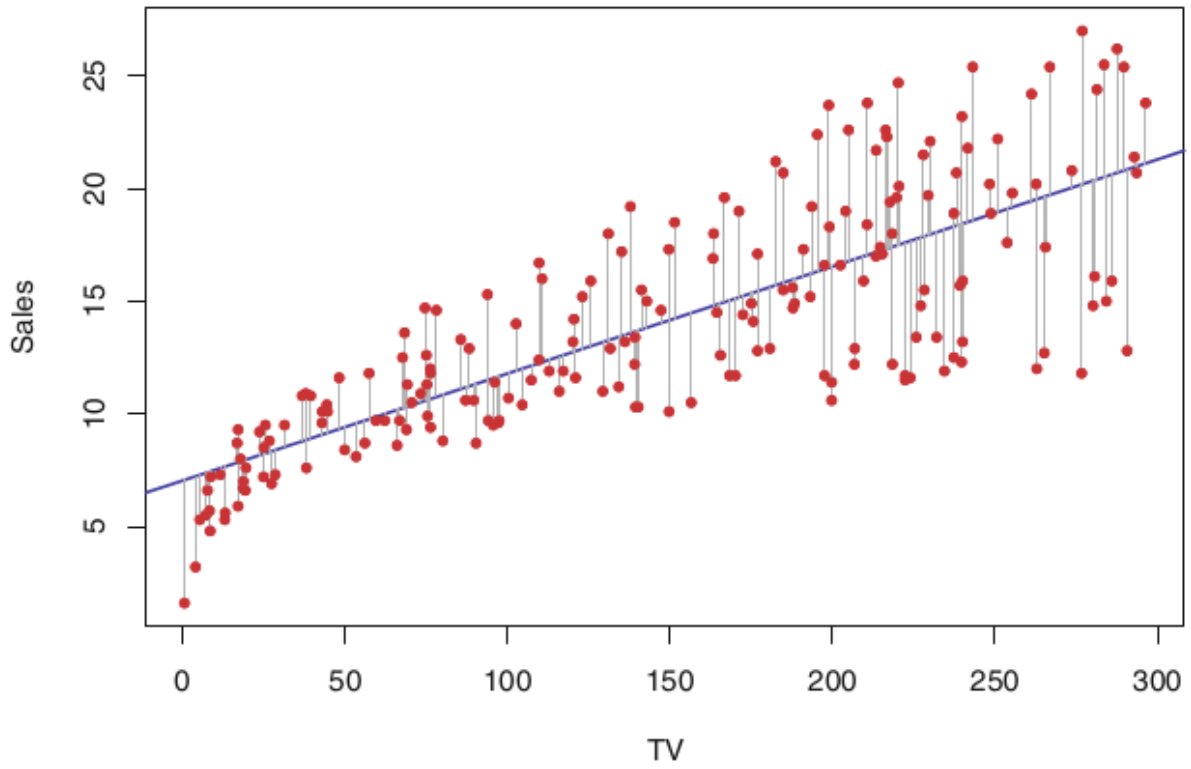
$$\min_{\beta_0, \beta_1 \in \Re} RSS = \sum_{i=1,m} (y^i - \beta_0 - \beta_1 x^i)^2 = \sum_{i=1,m} (y^i - \hat{y}^i)^2 = \sum_{i=1,m} (e^i)^2$$

- $RSS$ is convex $\Rightarrow$ the solution is unique:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1,m} (x^i - \bar{x})(y^i - \bar{y})}{\sum_{i=1,m} (x^i - \bar{x})} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

- prediction for an unseen $\tilde{x}$: $\tilde{y} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}$

example: `Sales` of a product as a function of the `TV` advertisings over 200 different markets




$$\hat{\beta}_0 = 7.03, \hat{\beta}_1 = 0.0475 \Rightarrow \texttt{Sales} = 7.03 + 0.0475 \ \ \texttt{TV}$$

by spending 1000 more euro in `TV` advertisings we sell 47.5 more units of product!

multivariate linear regression: predictors (features) are $n$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

- given $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n$ , the prediction $\hat{y}^i$ for point $x^i \in \Re^n$ is

$$\hat{y}^i = \hat{\beta}_0 + \hat{\beta}_1 x_1^i + \hat{\beta}_2 x_2^i + \cdots + \hat{\beta}_n x_n^i$$

- $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n$ are obtained by solving

$$\min_{\beta_0, \beta_1, \beta_2, \ldots, \beta_n} RSS = \sum_{i=1,m} (y^i - \beta_0 - \beta_1 x_1^i - \beta_2 x_2^i - \cdots - \beta_n x_n^i)^2 =$$

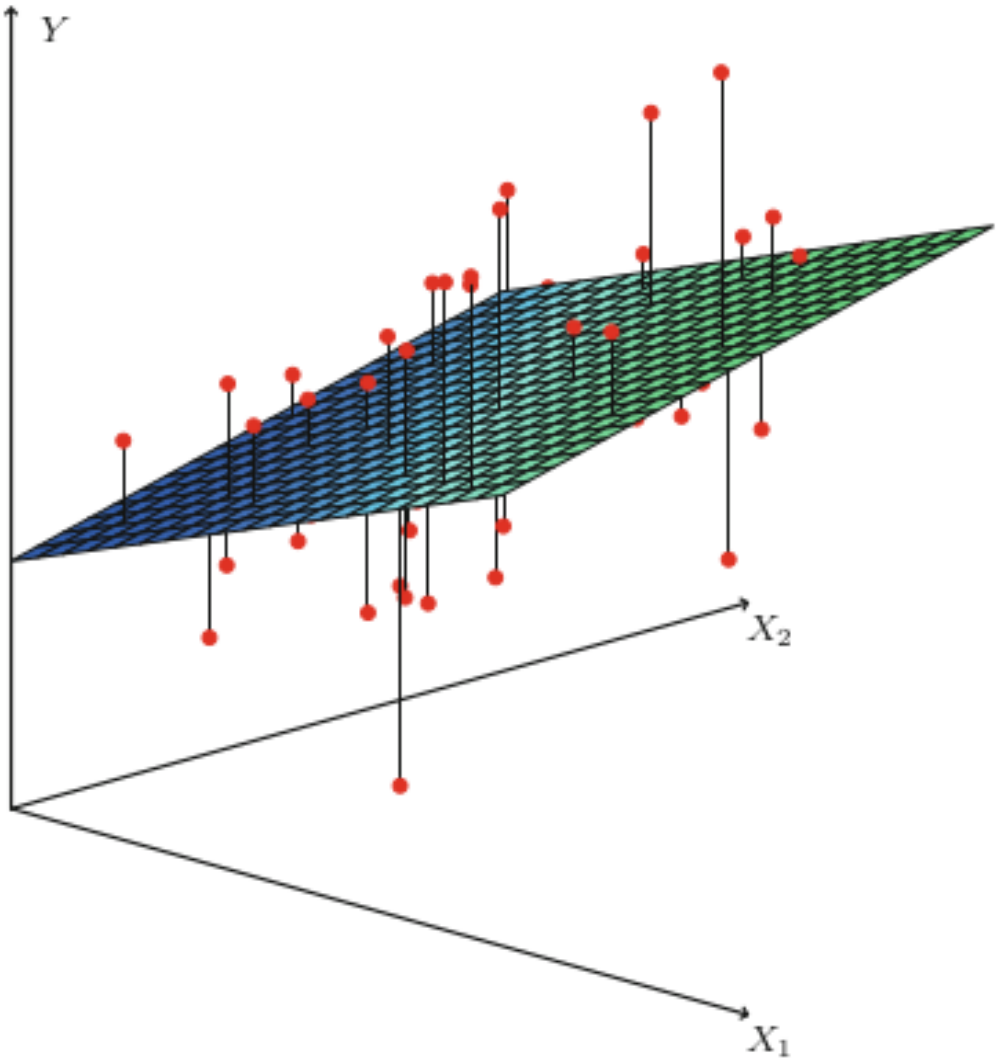
$$= \sum_{i=1,m} (y^i - \hat{y}^i)^2 = \sum_{i=1,m} (e^i)^2$$

- the solution $\hat{\beta} \in \Re^n$ can be computes as

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- prediction for an unseen $\tilde{x}$: $\tilde{y} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}_1 + \hat{\beta}_2 \tilde{x}_2 + \cdots + \hat{\beta}_n \tilde{x}_n$

# a 2-dimensional linear regression



how to measure the regression accuracy (on the training or testing set)?

- residual sum of squares ($RSS$)

$$RSS = \sum_{i=1,m} (y^i - \hat{y}^i)^2$$

- mean square error ($MSE$)

$$MSE = \frac{1}{m} \sum_{i=1,m} (y^i - \hat{y}^i)^2$$

(root mean square error ($RMSE$) $= \sqrt{MSE}$)

- coefficient of determination ($R^2$)

$$\text{total sum of squares } TSS = \sum_{i=1,m} (y^i - \bar{y})^2$$

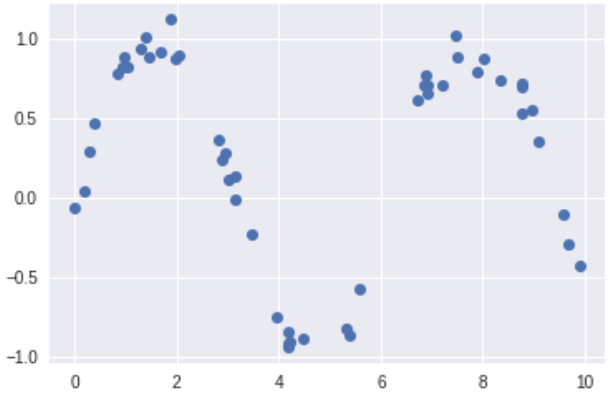
$$\text{explained sum of squares } ESS = \sum_{i=1,m} (\hat{y}^i - \bar{y})^2$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

# polynomial regression

sometimes linear regression is not suited...

...but linearity is desirable (simple training problem)

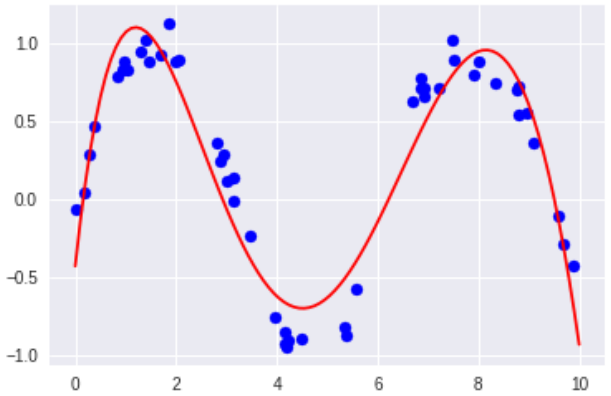



- idea: mantain the linearity of the model with respect to the training variables $\beta$ and polynomially map the features in a larger space (of dimension $q$)
- example of $q = 3$ polynomial transformation $x \Rightarrow \begin{pmatrix} x \\ x^2 \\ x^3 \end{pmatrix}$

$y = \beta_0 + \beta_1 x \Rightarrow y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

features are nonlinearly transformed (in a 3D space) but the model remains linear in parameters $\beta$!

exercise 4.1



typical supervised learning scheme

1. selection of the class of model $m(\cdot)$ (e.g. polynomial regression)
2. hyperparameters selection (e.g. polynomial degree)
3. training the model on $TR$
4. use of the trained model to make prediction on unseen new data

steps (1) and (2) fundamentals for the prediction performance of the model

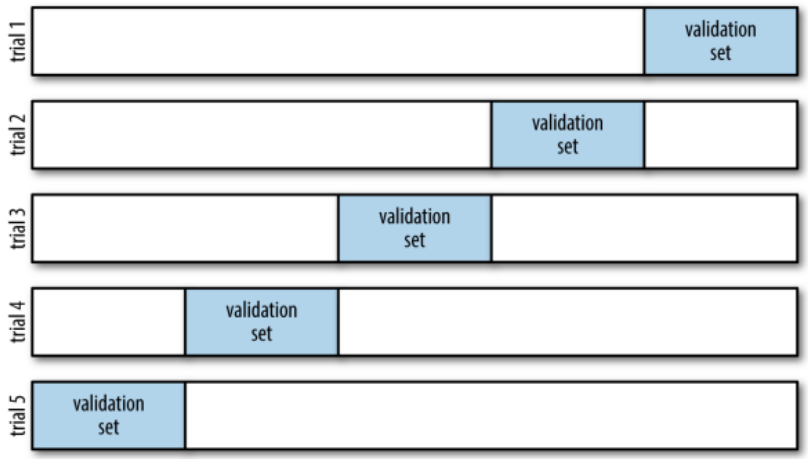
- how to select the model?
- how to select hyperparameters?

a model validation is needed!

model validation

- holdout set: remove a subset of available data to create a testing set(not used in the training!)
  if testing set small scarce significativity, if big too much data substracted to the training

- k-fold cross-validation: train the model $k$ times (trials or folds), at each trial the dataset is divided in a training and a validation set (used to measure the accuracy), the validation set is a $\frac{1}{k}$ fraction of the number of available data, the final performance is the average over all $k$ folds




  - all data considered one time in the validation
  - time consuming

which model/hyperparameters to select?

optimization of the model/hyperparameters is too expensive!

model/hyperparameters selection done heuristically via grid-search

1. for every hyperparameter (including model structure) choose a set of reasonable values (all possibly)
2. perform a k-fold cross-validation for every combination of hyperparameters values
3. select the hyperparameters values in which the best model accuracy is achieved

example: polynomial regression $\Rightarrow$ $\begin{cases} q: & \{1,2,3\} \\ \text{fit\_intercept}: & \{True, False\} \end{cases}$

| | | |
|---|---|---|
| (1,True) | (2,True) | (3,True) |
| (1,False) | (2,False) | (3,False) |

very time consuming!

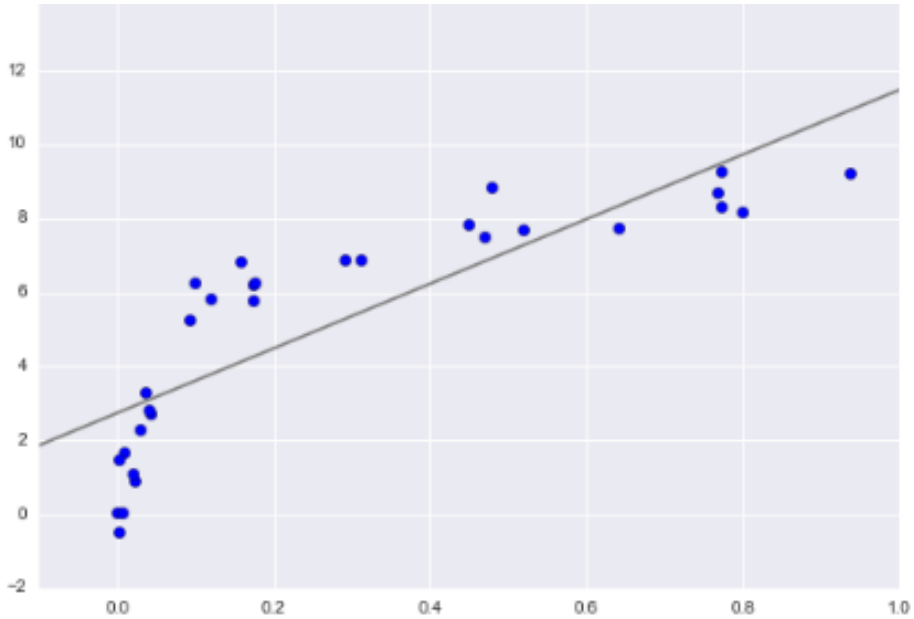
`sklearn.model_selection.GridSearchCV()`

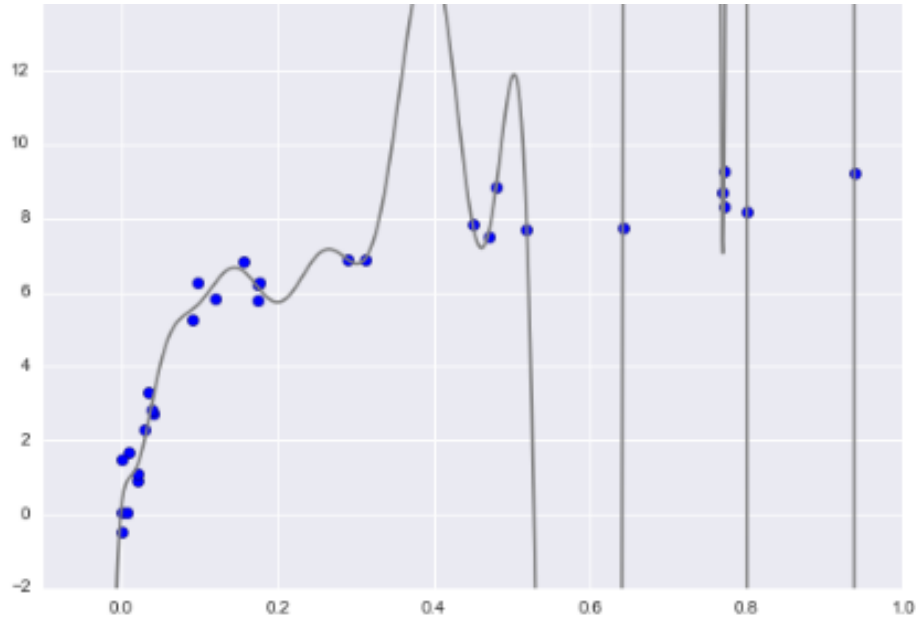
what if the trained model does not work well?

- adopt a more complex/expressive model?
- adopt a less complex/expressive model?
- collect more training data?
- increase the number of features?
- decrase the number of features?

often in machine learning the answer is counterintuitive!

# let's focus on model complexity

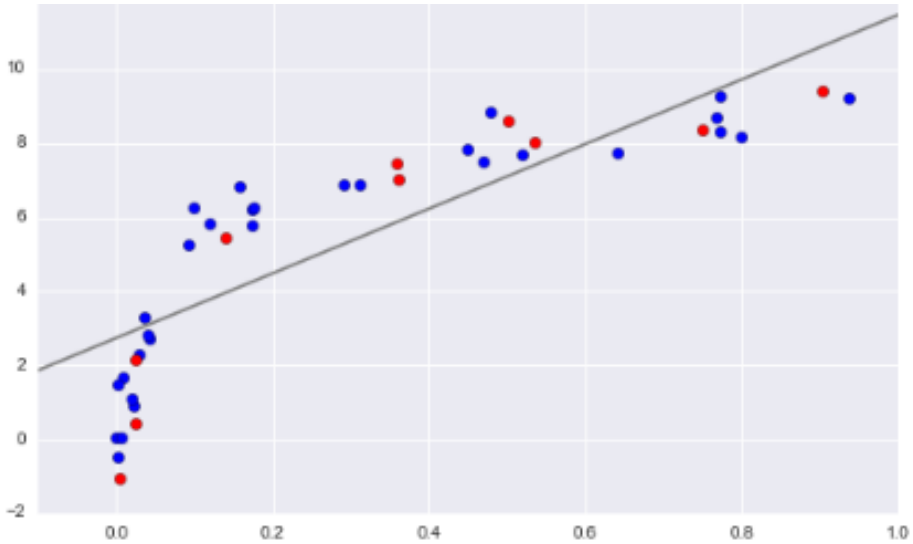


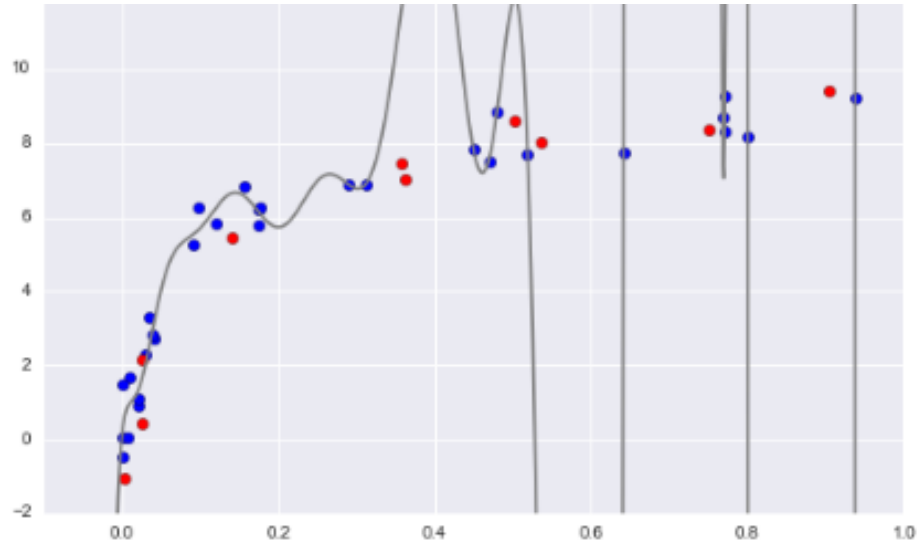




linear model does not fit data as data complexity is higher than the model complexity, underfitting occurs $\Rightarrow$ bias model

highly nonlinear model has enough complexity to fit data but it does not describe well the process, overfitting occurs $\Rightarrow$ variance model

let's focus on model complexity

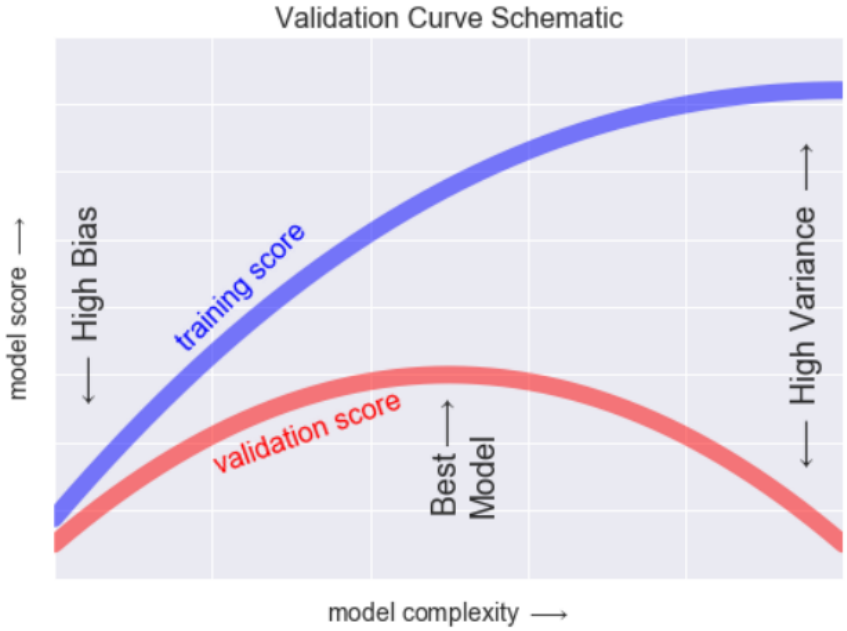


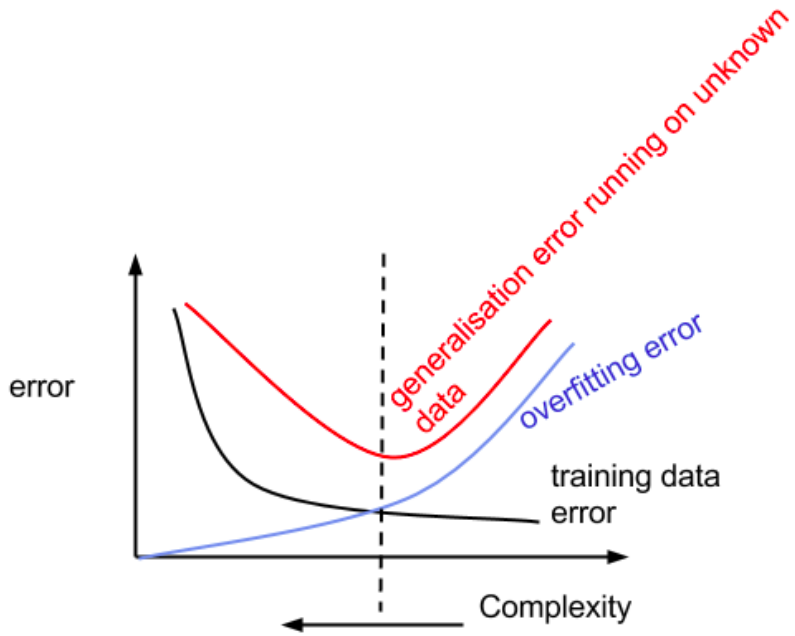




neighter of the two models is able to predict the testing points!

- for highly bias models low accuracy on both training and testing sets
- for highly variance models high accuracy on training set, low on testing set

a trade-off between bias and variance is required!

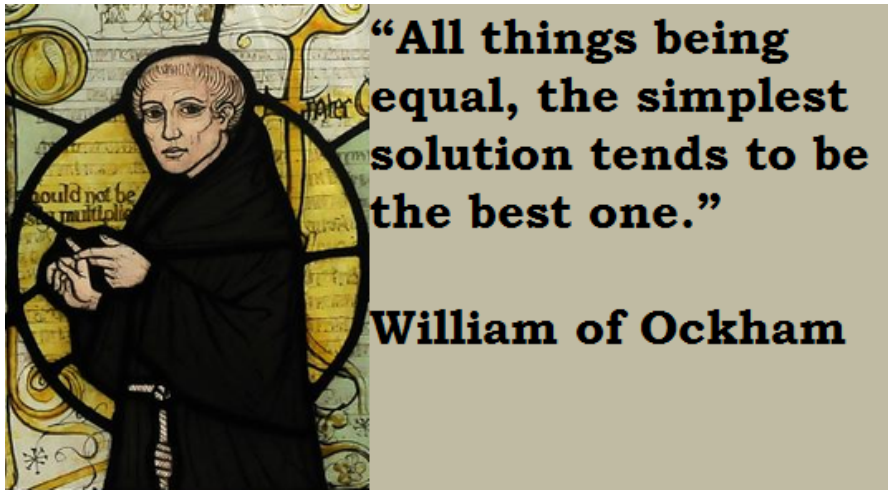


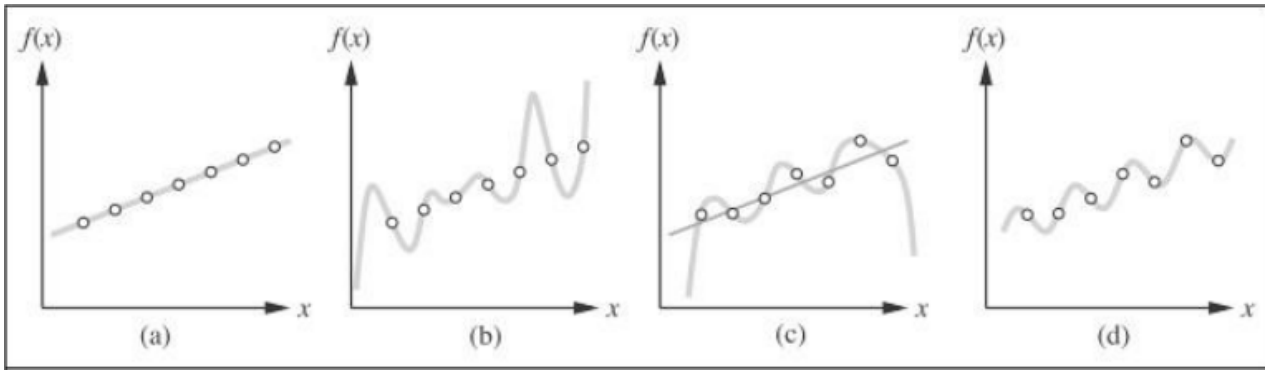

the goal is to obtain a model near the maximum of the validation score curve!

grid-search+cross-validation may help...

follow the Occam's razor principle...




prefer the simplest hypothesis consistent with data!



data are often in a format not suited for the learning...
data preprocessing may help in improving the prediction model

- image transformation (e.g. pixel matrix for handwritten digit)
- categorical data transformations
  - bad method: transform categories into integer numbers
    L'Aquila $\rightarrow$ 1, Rome $\rightarrow$ 2, Milan $\rightarrow$ 3
    inherits a meaningless order (L'Aquila $<$ Rome $<$ Milan)
  - good method: one-hot encoding transforms categorical data in array with a number of binary elements equal to the number of categories (-1)
    L'Aquila$\rightarrow$ [1 0], Rome$\rightarrow$ [0 1], Milan$\rightarrow$ [0 0]
    features grow too much with many categories
- text data transformations
  - words count: count the number of occurrences of each word in a text
    sample = ['problem of evil','evil queen','horizon problem']

| evil | horizon | of | problem | queen |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 |

- derivations (e.g. polynomial transformation)
- NA values imputations
- standardization/normalization (target normalization very important in regression!)

`sklearn.preprocessing`, `sklearn.feature_extraction`

feature selection may be useful to limit the complexity by reducing the actually used features

all possible subsets of features are $2^n$, too expensive training $2^n$ models...

...heuristics!

- forward selection
  1. start with the null-model, train all $n$ single feature models and add to the null-model the feature corresponding to the best performing single feature model
  2. train all the $n-1$ two features models and add to the previous single features model the feature corresponding to the best performing two features model
  3. iterate until some stopping condition is reached
- backward selection
  1. start with whole $n$ feature model and train it
  2. iteratively remove the feature which is "statistically less significant" until some stopping condition is reached
- mixed selection: combine forward and backward selection

`sklearn.feature_selection`

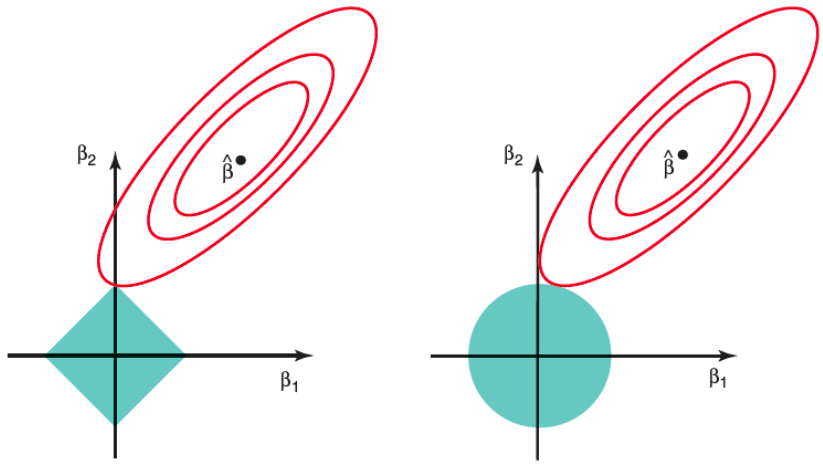
complexity can be also controlled by adding a regularization term to the loss function

example: multivariate regression

$$\min \quad loss + \lambda \cdot regularization = \sum_{i=1,m} (y^i - \hat{y}^i)^2 + \lambda \|\beta\|_p^p$$

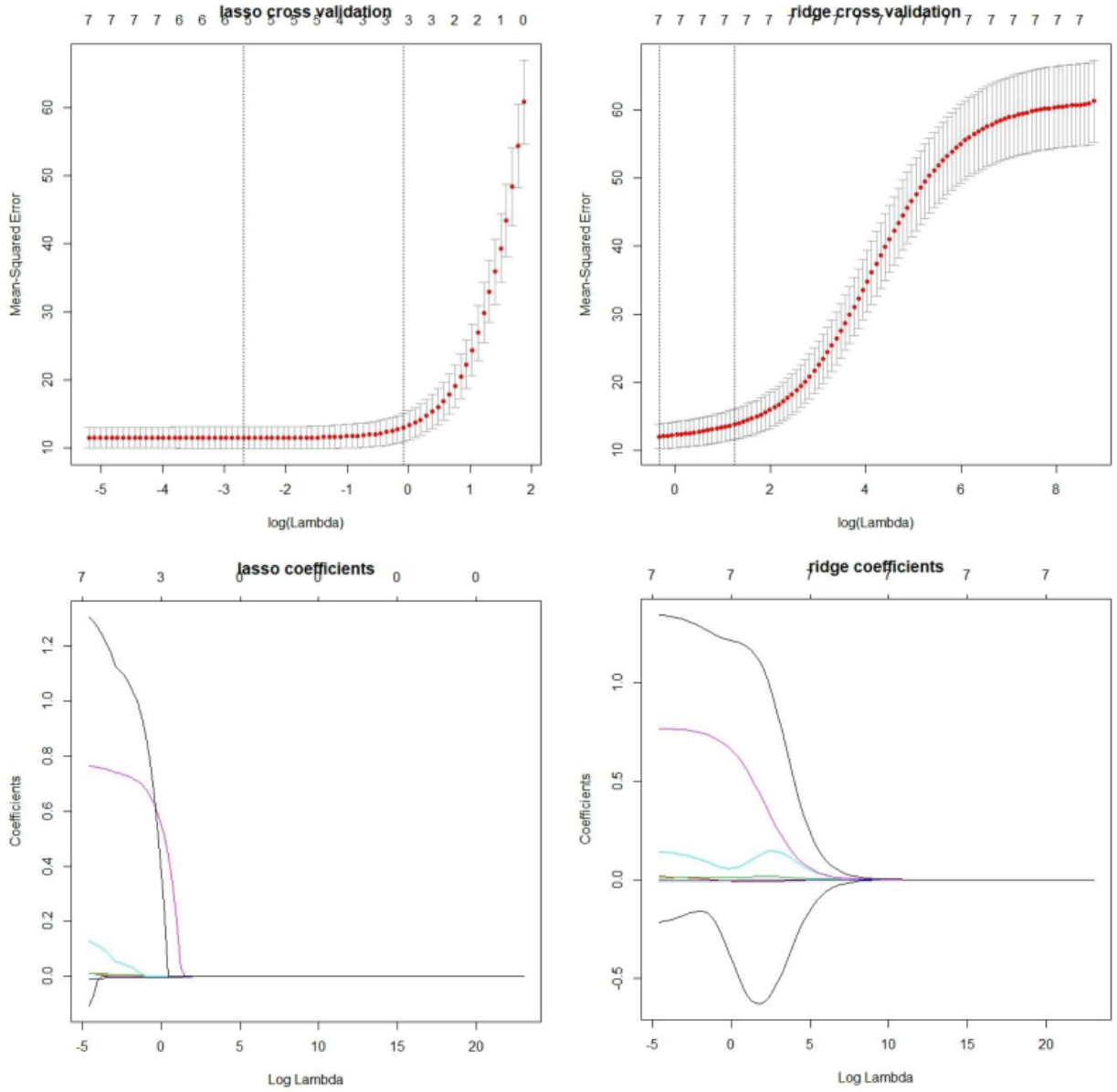
typical choices

- $p = 1$: lasso regression $\rightarrow$ min $\sum_{i=1,m} (y^i - \hat{y}^i)^2 + \lambda \sum_{j=0,n} |\beta_j|$
- $p = 2$: ridge regression $\rightarrow$ min $\sum_{i=1,m} (y^i - \hat{y}^i)^2 + \lambda \sum_{j=0,n} |\beta_j|^2$




lasso⇒ more sparse models ridge⇒ more smooth models

example: grid-search cross-validation results for lasso regression and ridge regression

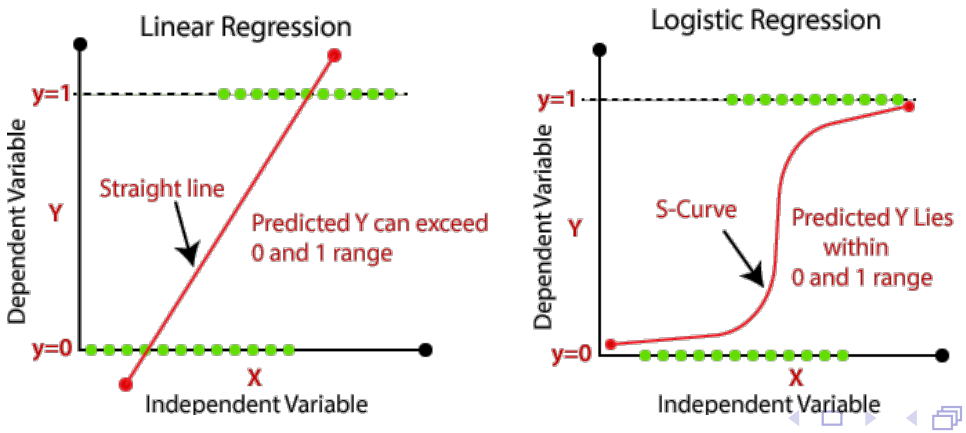


# logistic regression

intermediate bewtween regression and binary classification tasks

- consider a binary classification dataset ($y^i \in \{0, 1\}$) with single feature, assume we want to estimate for each feature $x^i$ the probability $p^i = P(y^i = 1)$
- in principle one could fit a line to data (linear regression) and set $p(x^i) = \hat{\beta}_0 + \hat{\beta}_1 x^i$ but $p(x^i)$ is not in the interval $[0, 1]$...
- better to fit a $S$-shaped nonlinear function ranging in the interval $[0, 1]$ like

$$p(x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x}}$$



- logistic regression training consists in determining the parameters estimates $\hat{\beta}_0, \hat{\beta}_1$ by maximizing the likelihood function over $TR$

$$\ell(\beta_0, \beta_1) = \prod_{i:y^i=1} p(x^i) \prod_{i:y^i=0} (1 - p(x^i))$$

- once trained the model one can use a threshold (typically 0.5) to classify data according to

$$\tilde{y} = \begin{cases} 1 \text{ if } p(\tilde{x}) \geq 0.5 \\ 0 \text{ if } p(\tilde{x}) < 0.5 \end{cases}$$

- logistic regression easily extends to the multivariate case

$$p(x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_n x_n}}$$

how to measure binary classification accuracy?

assume targets are {positive,negative} and the numer of samples of the two classes are $P$ and $N$

confusion matrix

| | predicted positive | predicted negative |
|---|---|---|
| actual positive | true positive (TP) | false negative (FN) |
| actual negative | false positive (FP) | true negative (TN) |

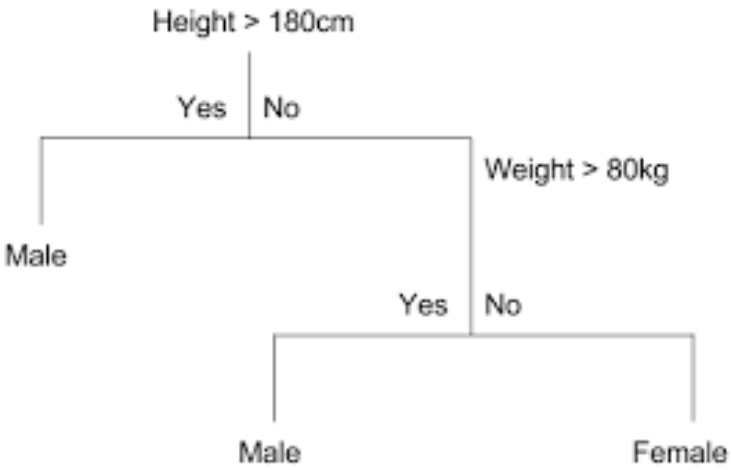
- $ACC = \frac{TP+TN}{P+N}$ (accuracy)
- $TPR = \frac{TP}{P}$ (true positive rate, sensitivity, recall...)
- $TNR = \frac{TN}{N}$ (true negative rate, specificity)
- $PPV = \frac{TP}{TP+FP}$ (positive predicted value, precision)
- $BA = \frac{TPR+TNR}{2}$ (balanced accuracy)

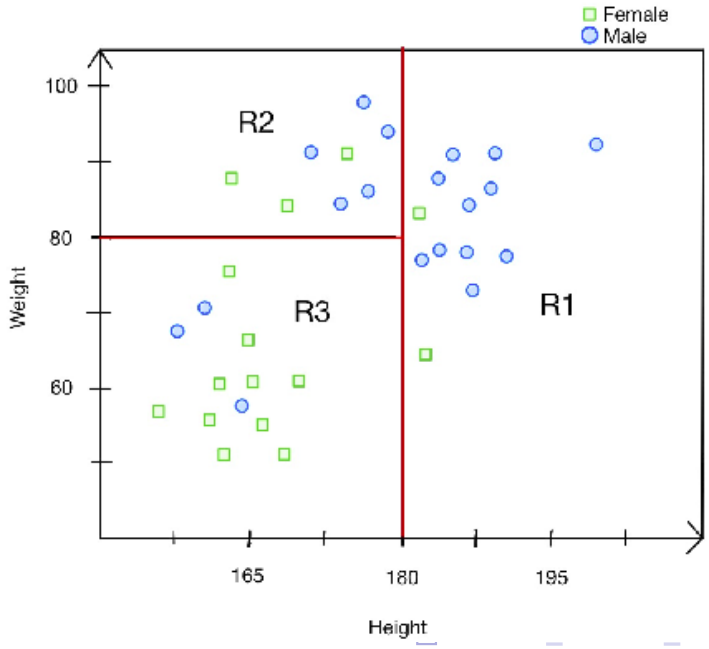
exercise 4.2

# decision trees

machine learning model for classification (binary and multiclass) and regression

- easily interpretable models and representable with a tree structure (widely used in medical applications)
- idea: the tree represents a sequence of splitting rules operated on the features' domain
- branch nodes (splitting decisions) and leaf nodes (final target assignment)
- the class of leaf nodes is the one of the largest number of training observations falling into the node



- multiclass classification context

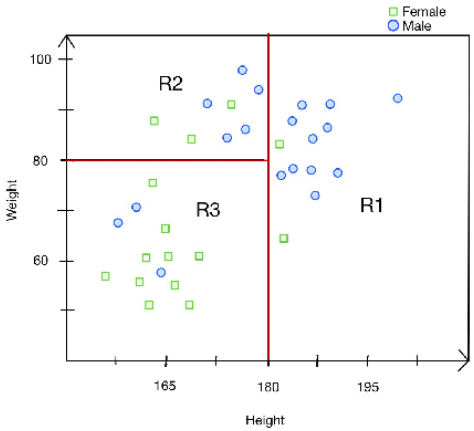
$$TR = \{(x^i, y^i) : x^i \in \Re^n, y^i \in \{1, 2, \ldots, K\}, i = 1, \ldots, m\}$$

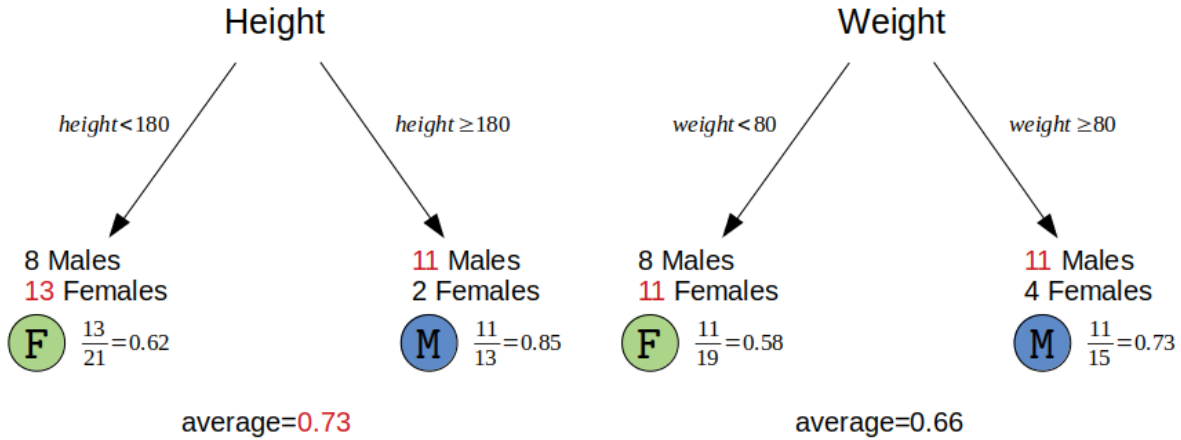
- given feature $x_j$ and threshold $s$, the splitting regions $R_1(j, s), R_2(j, s)$ are

$$R_1(j, s) = \{x : x_j < s\} \text{ and } R_2(j, s) = \{x : x_j \geq s\}$$

no need of threshold for binary features!

- the class of a region is its majority class
- the "quality" of a splitting is the ability to separate observations of different classes according the purity ratio $\frac{\#\text{ majority class}}{\#\text{ observations in the region}}$



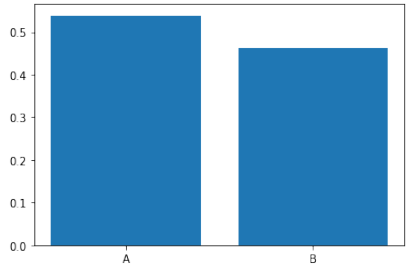



often entropy is used as impurity measure (instead of the purity ratio)

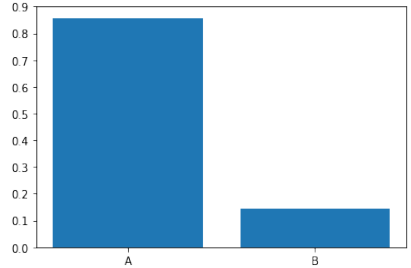
$$H = -\sum_{k=1,K} f_k \log(f_k)$$

where $f_k$ is the frequence of occurrences of class $k$ in the region

bad splitting




good splitting




how to train a classification tree?

in principle determine regions and thresholds $R_\ell(j, s)$ with $j = 1, \ldots, n, \ell = 1, \ldots, M$ minimizing the overall entropy of all splittings

too expensive to consider every partition of the feature space in $M$ regions!

$\Rightarrow$ a top-down greedy approach: recursive binary splitting

1. at each step determine feature $x_j$ and threshold $s$ such that the splitting

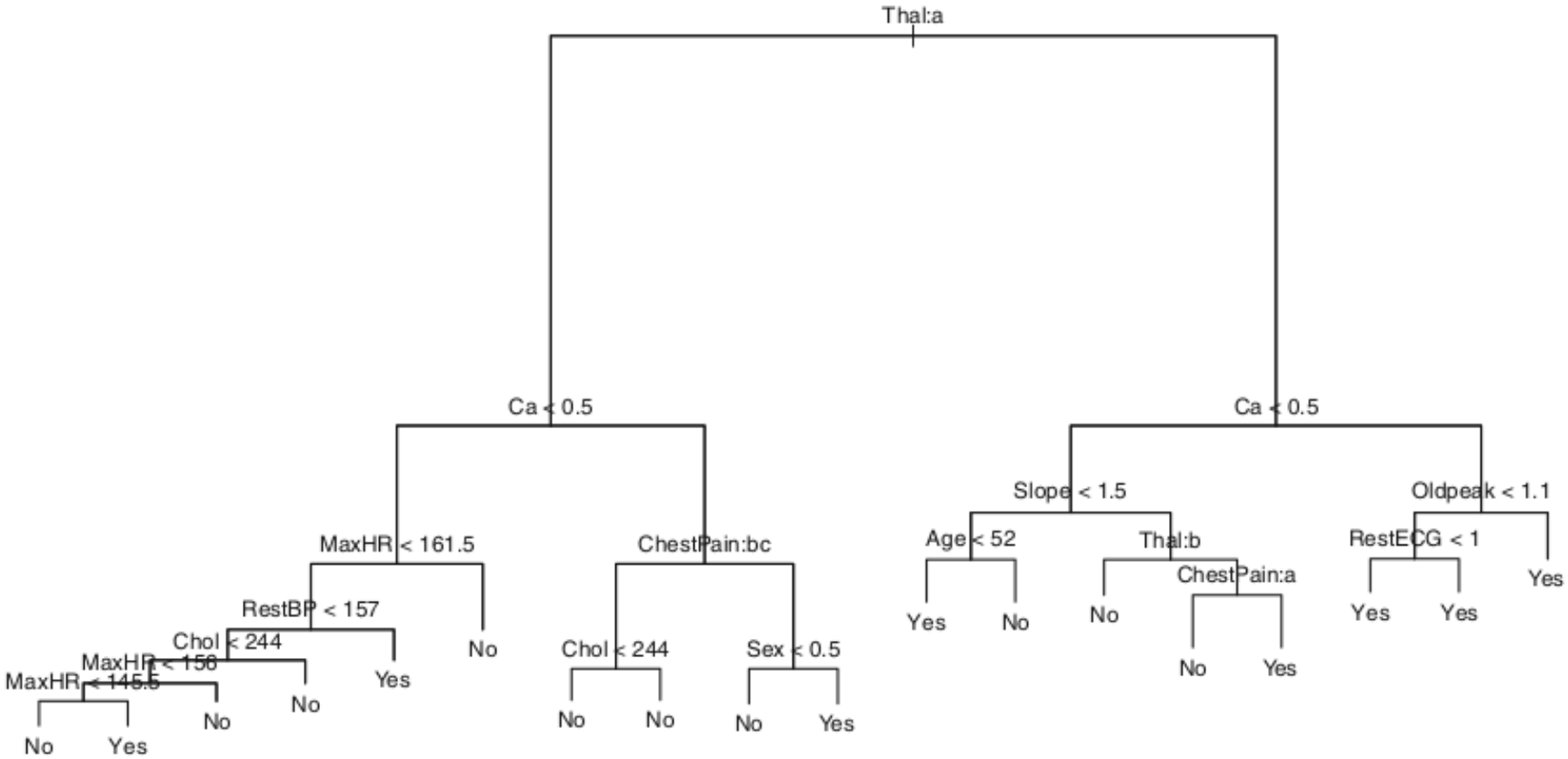
$$R_\ell(j, s) = \{x : x_j < s\} \text{ and } R_{\ell+1}(j, s) = \{x : x_j \geq s\}$$

   minimize the average entropy of the two formed regions among all possible features and thresholds
2. stop if some condition is reached (e.g. number of points in a region, number of nodes, entropy of the splittings)

recursive binary splitting

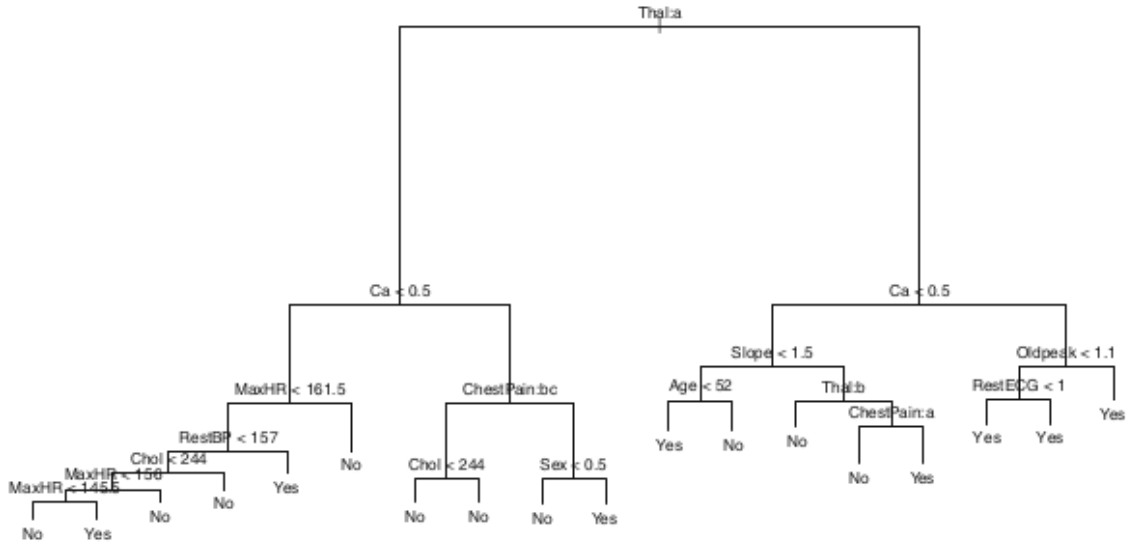
- greedy $\Rightarrow$ at each step the best current split is added not considering the split leading to the best final tree
- computing the best current split ($x_j$ and $s$) can be done efficiently
- can be used for mixed features (continuous, categorical,...)

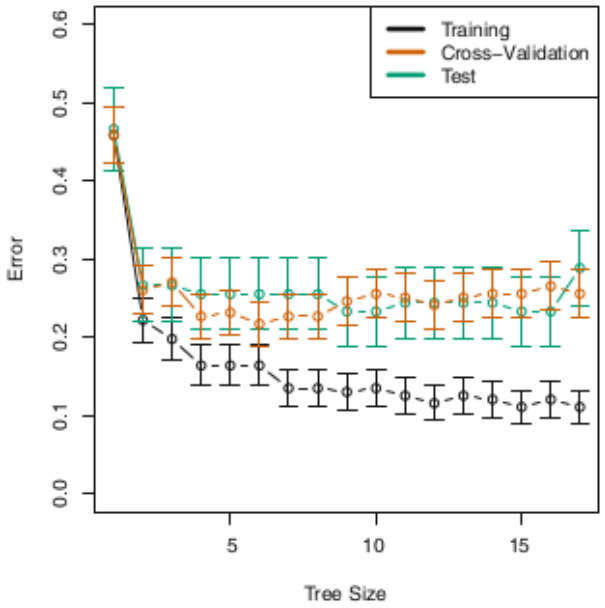
example: dataset of 303 patients with binay target indicating the presence of heart disease (the specified value in the branch node corresponds to left branch)

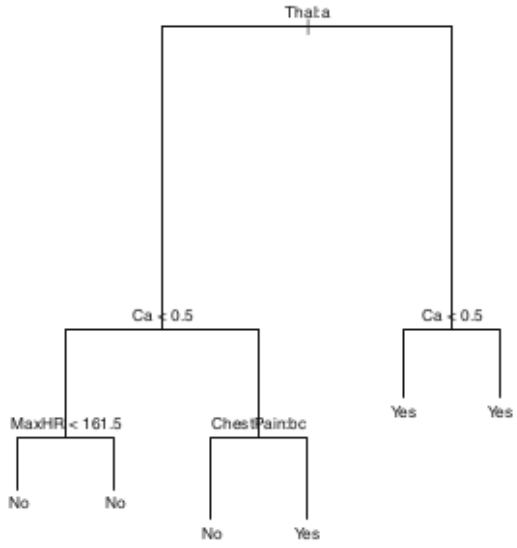


- recursive binary splitting may overfit data with too complex trees
- smaller trees may have lower variance with slightly larger bias
- bad strategy $\Rightarrow$ stop splits early when the entropy reduction is below an high threshold ('better' splits can occur later in the tree)
- good strategy $\Rightarrow$ pruning:
  1. grow a large initial tree $T_0$ where $|T_0|$ is the number of leaf nodes
  2. for different values of coefficient $\alpha$ determine a sequence of sub-trees $T \subseteq T_0$ minimizing
$$\sum_{p=1}^{|T|} \Big( - \sum_{k=1,K} f_k \log(f_k) \Big) + \alpha |T|$$
  3. operate a cross-validation to determine the best $\alpha$ value and so the best sub-tree

# a pruned classification tree



## some decision trees considerations

- Gini coefficient can be used as a impurity measure
- in the regression version the predicted target of leaf nodes is the mean value of training observations falling in that node
- decision trees are close to human decision-making mechanisms
- are easy to depict and interpret
- can deal with mixed features
- no high level of accuracy compared to other models
- not robust to data

$\Rightarrow$ to overcome the previous drawbacks multiple trees are combined together (random forests)

exercise 4.3

# Support Vector Machines

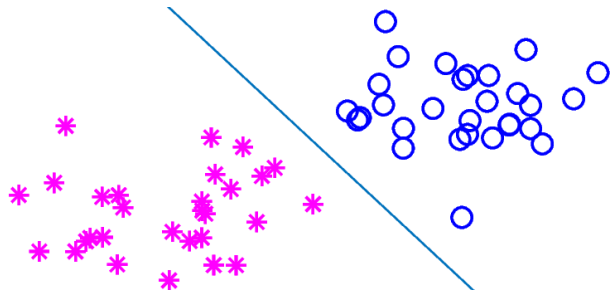
SVMs are machine learning models for binary classification (adaptable to multiclass and regression)

- binary classification problem

$$TR = \{(x^i, y^i) : x^i \in \Re^n, y^i \in \{-1, +1\}, i = 1, \ldots, m\}$$

- assume $TR$ is linearly separable, i.e. the points of the two classes can be separated by a $(n-1)$-dimensional separating hyperplane

$$w^T x + b = 0, \quad w \in \Re^n, b \in \Re$$



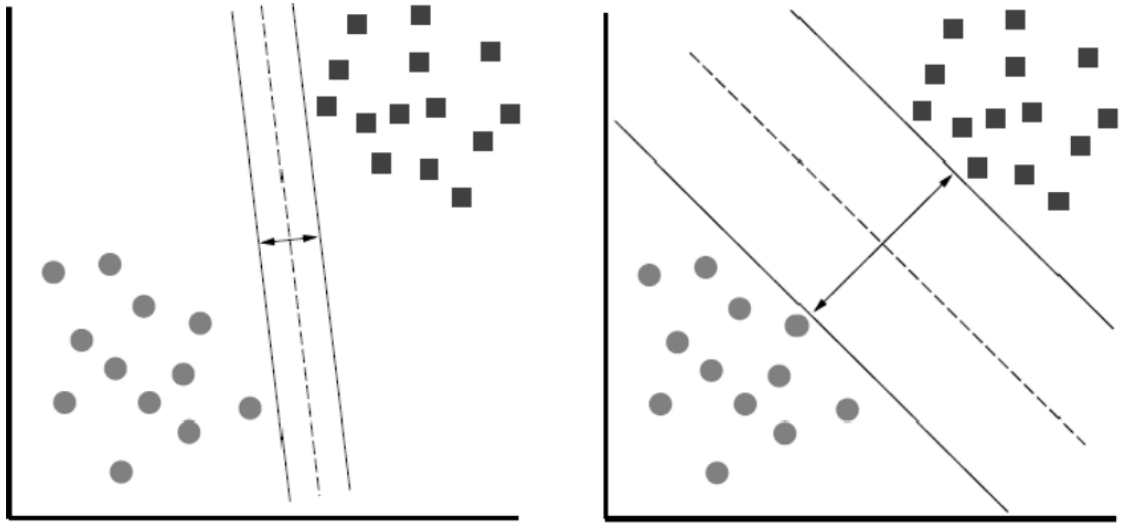
- training a SVM $\Rightarrow$ determine optimal parameters $(\hat{w}, \hat{b})$ of the hyperplane separating samples in $TR$ according to their class
- once trained the model, the decision function $h(\cdot)$ to classify unseen observations is

$$h(\tilde{x}) = sign(\hat{w}^T \tilde{x} + \hat{b})$$

- if $TR$ is linearly separable, a separating hyperplane $(w, b)$ divides data with a tolerance gap $\rho$ according to the separating conditions

$$\begin{cases} w^T x^i + b \geq \rho & \text{for } y^i = +1 \\ w^T x^j + b \leq -\rho & \text{for } y^j = -1 \end{cases} \Rightarrow \begin{cases} w^T x^i + b \geq 1 & \text{for } y^i = +1 \\ w^T x^j + b \leq -1 & \text{for } y^j = -1 \end{cases}$$

- the separating hyperplane are infinitely many! which one is better?



- the maximum margin hyperplane (the one with maximum distance from the closest points) is better!
  - hyperplanes close to training samples are more sensitive to noise
  - intuitively/theoretically a big margin hyperplane is more likely to better classifiy unseen data

a maximum margin training problem...

- the distance of a point $x^i$ from hyperplane $(w, b)$ is

$$d(x^i; w, b) = \frac{|w^T x^i + b|}{\|w\|}$$

- the margin of a hyperplane $(w, b)$ for $TR$ is

$$\rho(w, b) = \min_{i=1,\ldots,m} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\}$$

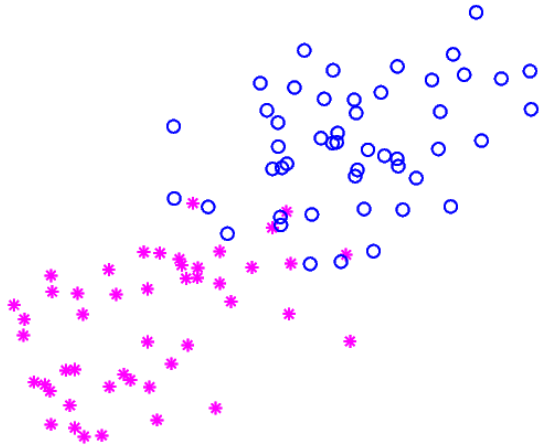
- it can be shown that: maximizing $\min_{i=1,\ldots,m} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\} \equiv$ minimizing $\|w\|$
- SVM training optimization problem

$$\min_{w \in \Re^n, b \in \Re} \quad \frac{\|w\|^2}{2}$$

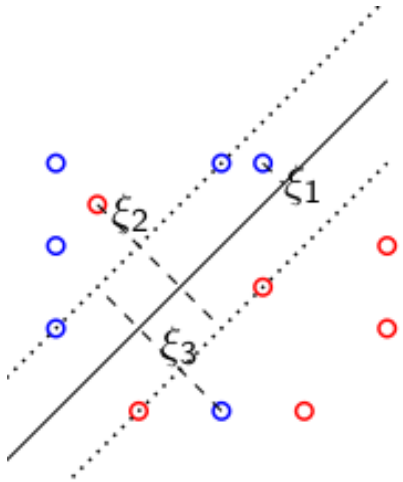
$$\text{subject to} \quad y^i(w^T x^i + b) \geq 1 \quad i = 1, \ldots, m$$

- it is a convex problem!

what if $TR$ is not linearly separable?



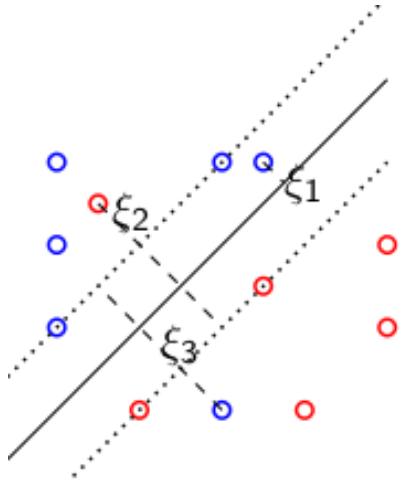
- in such a case there are no separating hyperplanes but we may wish to mantain linearity of the spearating surface
- idea: for each separating constraint $y^i(w^T x^i + b) \geq 1$ add a slack variable $\xi_i$ with $\xi_i \geq 0$ allowing a potential violation of it



# the SVM soft-margin formulation

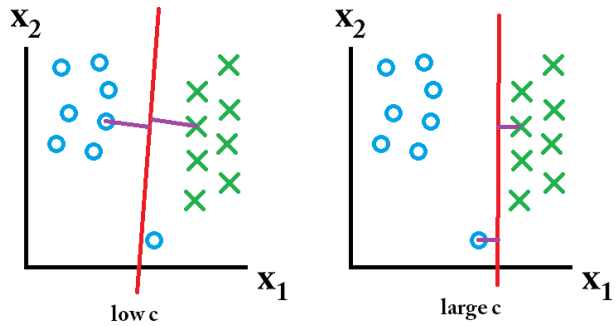
$$\min_{w\in\Re^n, b\in\Re, \xi\in\Re^m} \frac{\|w\|^2}{2} + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \quad y^i(w^T x^i + b) \geq 1-\xi_i \quad i = 1,\ldots,m$$
$$\xi_i \geq 0 \quad i = 1,\ldots,m.$$

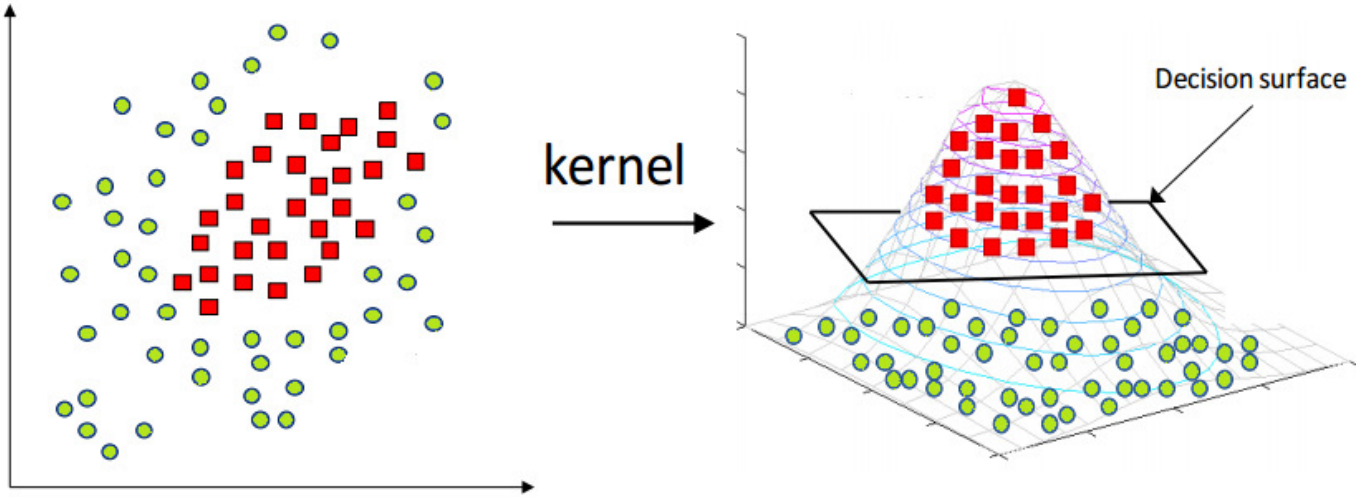



- $\xi_i$ measures how much constraint $i$ is violated
- $0 \leq \xi_i < 1$ the point is correctly classified although violating the constraint
- $\xi_i \geq 1$ the point is misclassified
- $C$ is the weight of the term penalizing the misclassifications in the objective
- it is still a convex problem!

$C$ values comparison

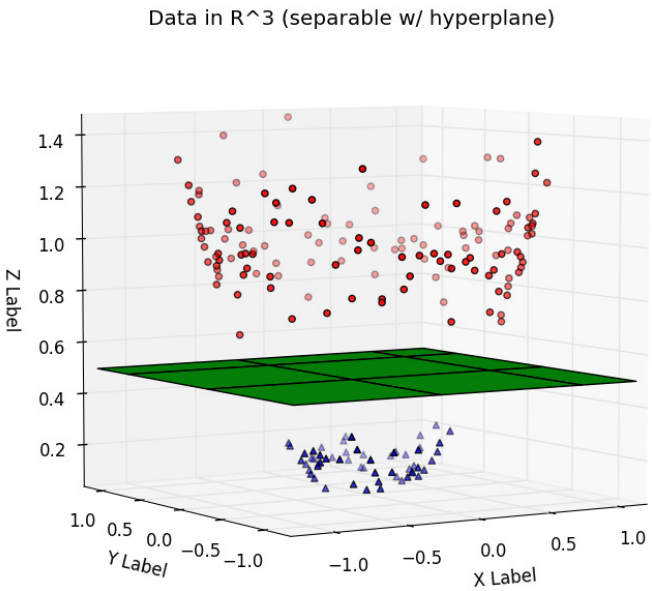


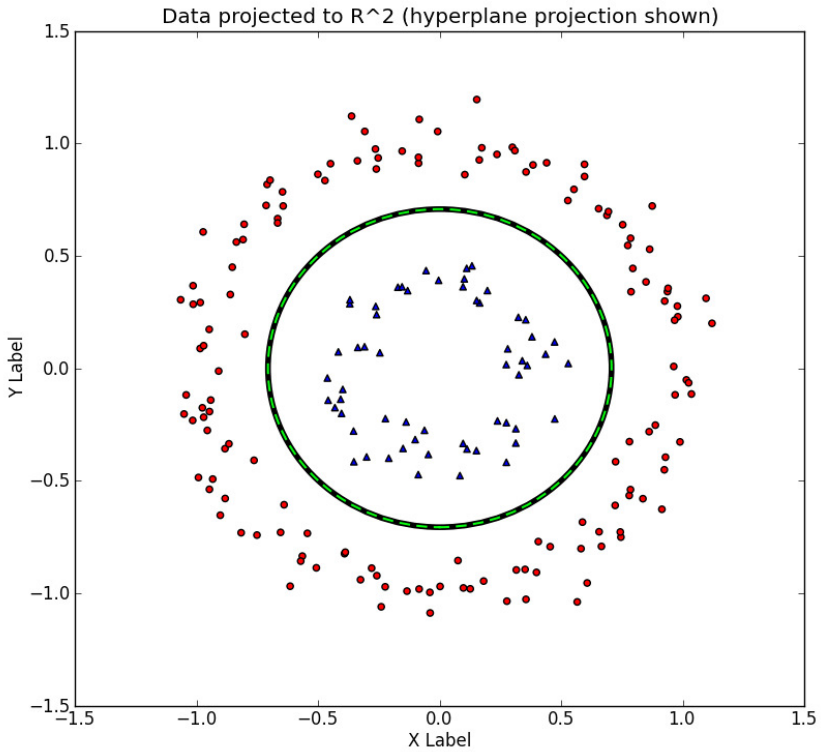
what if $TR$ is not suited for a linear decision function?




- idea: mapping input data $x$ from their original input space to a higher dimensional space (feature space) where data are more likely to be linearly separable
- the nonlinear mapping $\phi : \Re^n \rightarrow \mathcal{H}$ has 2 main peculiarities
  - $dim(\mathcal{H}) >> n$
  - it is known an analytical formula to efficiently compute $\phi(x^i)^T \phi(x^j)$ (kernel)

a linear separating surface in the feature space is nonlinear in the original input space




example:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \stackrel{\phi}{\rightarrow} \quad \begin{pmatrix} -x_1 \\ x_2 \\ -x_1^2 \end{pmatrix} \equiv \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

$$-w_1 x_1 + w_2 x_2 - w_3 x_1^2 + b = 0 \quad \leftarrow \quad w_1 z_1 + w_2 z_2 + w_3 z_3 + b = 0$$

the nonlinear SVM formulation

$$\min_{w \in \Re^{dim(\mathcal{H})}, b \in \Re, \xi \in \Re^m} \frac{\|w\|^2}{2} + C \sum_{i=1}^{m} \xi_i$$

$$\text{subject to } y^i(w^T \phi(x^i) + b) \geq 1 - \xi_i \quad i = 1, \ldots, m$$

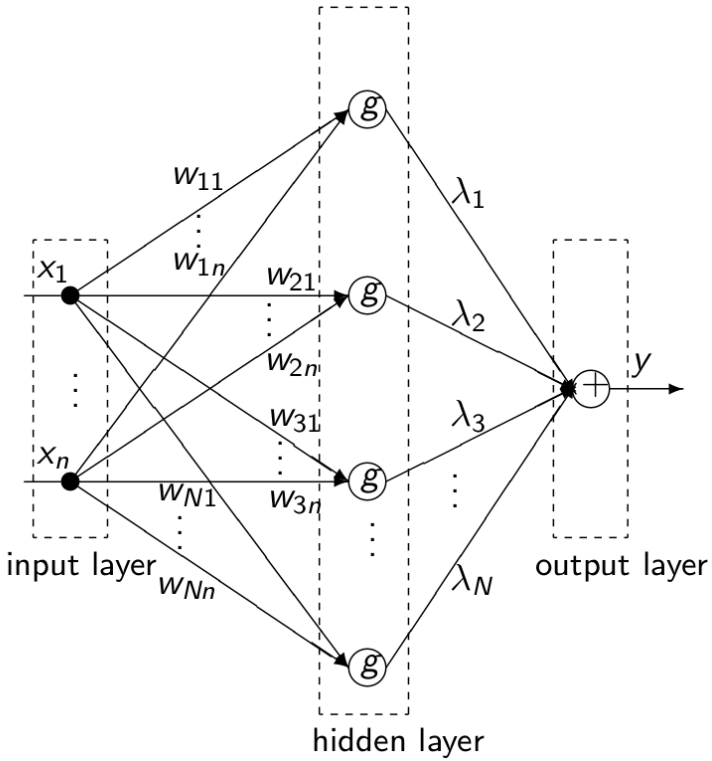
$$\xi_i \geq 0 \quad i = 1, \ldots, m$$

- it is still a convex problem!
- the explicit mapping $\phi$ may not be known or computable (the feature space may be of infinite dimension) $\Rightarrow$ in the dual reformulation of the problem mapping $\phi$ appears only as kernel terms $\phi(x^i)^T \phi(x^j)$ which can be efficiently computed (kernel trick)
- typical kernel functions:
  - linear: ${x^i}^T x^j$ (corresponding to a linear SVM)
  - gaussian (rbf): $e^{-\gamma \|x^i - x^j\|^2}$ for $\gamma > 0$
  - polynomial: $({x^i}^T x^j + 1)^p$ for $p \geq 1$ integer


exercise 4.4

# neural networks

neural networks are regression/classification machine learning models inspired by the functioning of the biological neuron

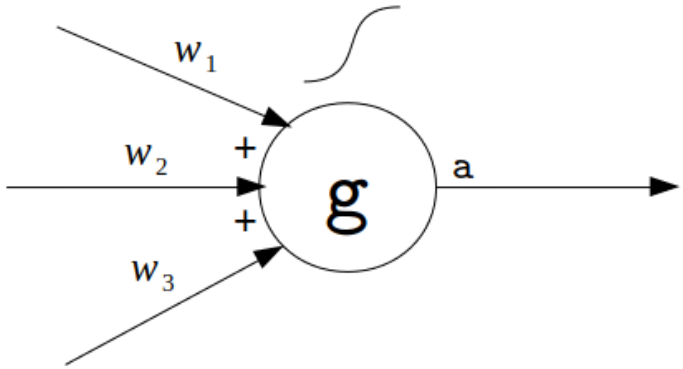



neural networks are input-output systems with a layered structure

- input layer propagating the input signals (features)
- hidden layers made up of the computational units neurons elaborating the signals from the previous layers
- output layer providing the final output signal

weighted connections link the layers

each neuron is in turn an input-output system...




- taking as input the weighted incoming signals
- elaborating their sum by means of a nonlinear activation function $g$ (typicall sigmoidal) to produce an output signal $a$

training a neural network consists in tuning the weights of the connections so as to minimize a loss function representing the overall discrepancy between the outputs $\hat{y}^i$ produced by the network for points $x^i$ in $TR$ and the actual outputs $y^i$ like $RSS = \sum_{i=1,m} (y^i - \hat{y}^i)^2$

some neural networks considerations...

- difficult to train (training optimization problem is not convex!)
- hyperparameters are difficult to set (e.g. number of hidden layers, number of neurons on the layers, type of activation function,...)
- once trained they act as a black-box (not interpretable)
- big representative power
- can approximate any contnuous function with arbitrary precision
- are the basis of deep learning models for e.g. image and speech recognition
- in `scikit-learn`
  - `sklearn.neural_network.MLPClassifier()` (classification model)
  - `sklearn.neural_network.MLPRegressor()` (regression model)

# references

- linear/polynomial regression
  - 'An Introduction to Statistical Learning', Sections 3.1-3.3, 7.1
- logistic regression
  - 'An Introduction to Statistical Learning', Section 4.3
- decision trees
  - 'An Introduction to Statistical Learning', Section 8.1
- support vector machines
  - 'An Introduction to Statistical Learning', Sections 9.1-9.3
- neural networks
  - 'The Elements of Statistical Learning', Section 11