Probability and Statistics for Machine Learning

Mark Asch - IMU/VLP/CSU

2023

ML = STATISTICAL learning

- all Machine Learning methods are statistical in nature, since we learn general relationships from a sample/data/observations/measurements

- in order to not just learn "cooking recipes", we will use a minimal mathematical formalism that gives us a uniform and coherent representation of statistical learning

- we have:

  $\Rightarrow$ independent variables x (inputs, features, attributes, explanatory variables)
  $\Rightarrow$ dependent variables y (outputs, responses, explained variables)
  $\Rightarrow$ an unknown relationship, f, that links inputs to outputs, and that we want to learn from the available data
    $\rightarrow$ for predictions
    $\rightarrow$ for inference

Populations and Samples

Definition 1. A population is the set of all objects (observations) being studied. Their number is denoted by N.

Definition 2. A sample is a subset, of size n, $n \leq N$, drawn from the population. We examine these observations to draw conclusions and to make inferences about the population.

For Big Data:

✘ $N =$ALL ??? No.

✘ Correlation $\Longrightarrow$ Causation ??? No.

Pre-requisite: the mathematical framework

- Suppose we have :

  $\Rightarrow$ a response variable (to explain), Y,
  $\Rightarrow$ p explanatory, variables, $X = (X_1, X_2, \ldots, X_p)$,
  $\Rightarrow$ n samples of data, giving an $(n \times p)$ matrix,

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & & \\ \vdots & & \ddots & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

  $\Rightarrow$ a relationship between Y and X of the form

$$Y = f(X) + \epsilon$$

  where
  $\rightarrow$ f is an unknown function of $X_1, X_2, \ldots, X_p$

$\rightarrow$ ϵ is a random error term, independent of X, and with zero mean

- ML is then an ensemble of approaches for estimating f with the objectsivesof

  $\Rightarrow$ Prediction : $\hat{Y} = \hat{f}(X)$ where $\hat{f}$ is an estimation for f and $\hat{Y}$ is the resulting prediction

  $\Rightarrow$ Inference: to understand how Y varies as a function of X (correlations, importances, linearity, etc.)

Step 1: Exploratory Data Analysis (EDA)

✔ An initial, critical step of the «data science» process

✔ There are neither hypotheses, nor models - we explore and we try to understand the problem!

✔ The tools of EDA are :

- summary statistics
- basic plots
- graphics

✔ The methodology :

- systematic passage over all the data
- plot all distributions of all the variables («box plots»)
- plot all the time series
- try changes of variables (usually logs or powers)
- look at all the relations two-by-two («scatterplots»)

- calculate all the summary statistics: mean, minumum, maximum, quartiles, outliers

SUMMARY Statistics

- measures of
  - $\Rightarrow$ central tendancy
  - $\Rightarrow$ dispersion around the centre

Measures of Central Tendancy

- mean:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{m} x_{ij}, \quad j = 1, \ldots, p$$

```
> Xj    = c(1,2,3,4,5)
> Xbarj = mean(Xj)
```

- median : value for which at most the half of the population is less than, and at least half is greater than,

$$\text{median}(x) = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ odd} \\ \dfrac{x_{(n/2)} + x_{(n/2)+1}}{2}, & \text{if } n \text{ even} \end{cases}$$

```
> Xmedj = median(Xj)
```

- mode : the most frequent value (for which the frequency/probability is maximal)

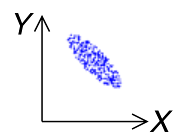
Measures of Dispersion

- variance and standard deviation:

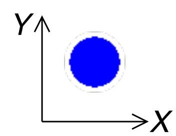
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

```
> Xvar = var(Xj)
> Xstd = sd(Xj)
```

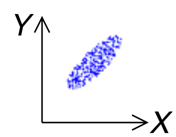
- covariance between k variables, with n observations each, is a $k \times k$ matrix with elements



cov(X,Y)<0



cov(X,Y)≈0



cov(X,Y)>0

$$q_{jk} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

```
> Xcov = cov(XX) # covariance
> Xcor = cor(XX) # correlation, entre -1 et 1
```

- quartiles, quantiles and inter-quartile distance: z is the k-th q-quantile, if

$$\Pr[X < z] \leq \frac{k}{q}$$

$\Rightarrow$ the median is the second quartile, Q_2

$\Rightarrow$ la inter-quartile distance

$$\mathrm{IQR} = Q_3 - Q_1$$

is a measure of dispersion

$\Rightarrow$ the 100-quantiles are called percentiles

```
> range(Xj) # max - min
> quantile(Xj) #  0, 25, 50, 75 et 100%
> IQR(Xj)
```

Summary Statistics

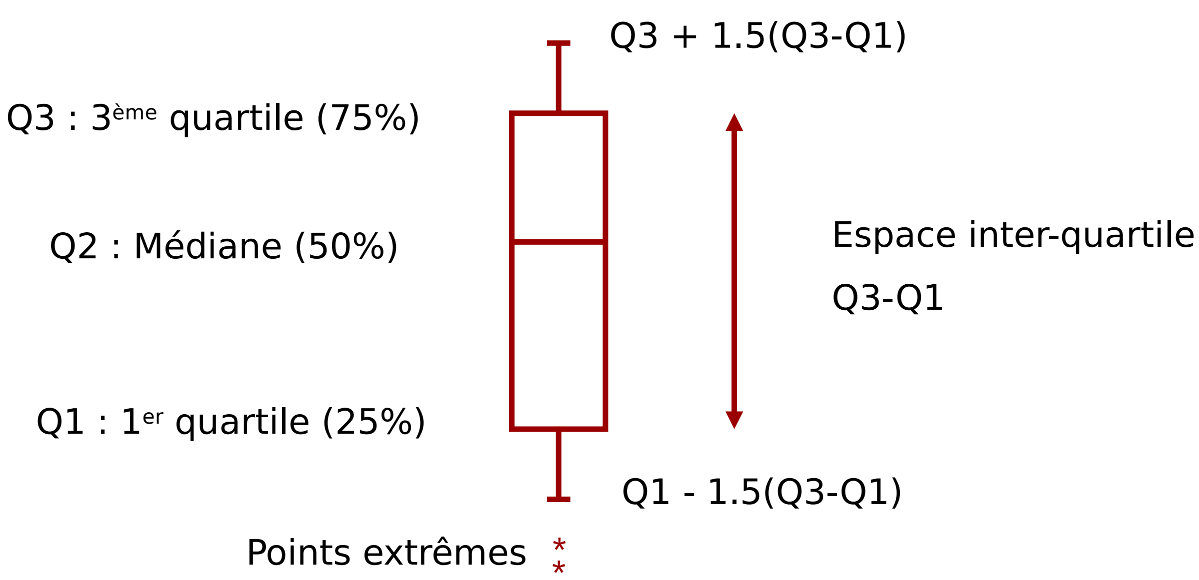
- the 5-number summary of Tukey is employed systematically for any data analysis

    1. minimum
    2. first quartile
    3. median
    4. third quartile
    5. maximum

```
> fivenum(Xj)
> summary(XX)
```

Plots and Graphics for EDA

- box-plots :

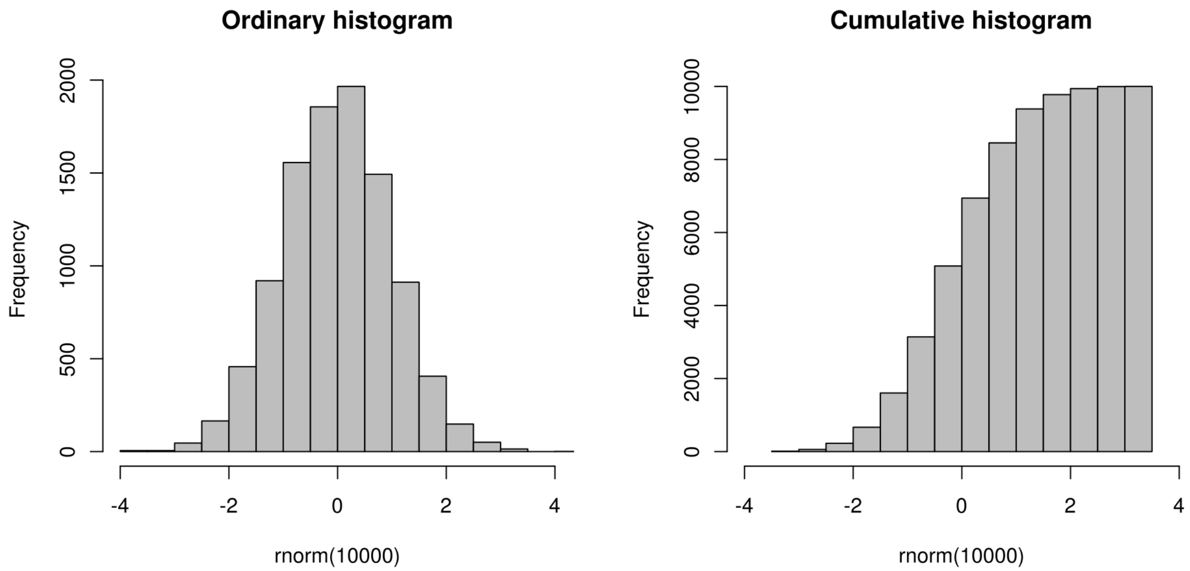


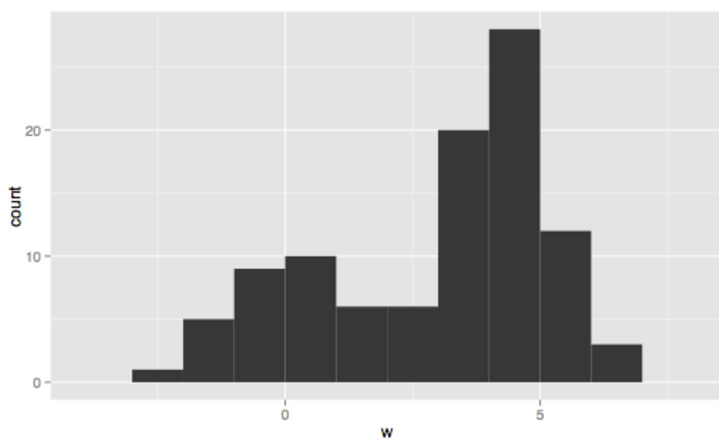

```
> boxplot(Xj)
```

- histograms :

⇒ approximates the probability density function
⇒ allows to detecti multi-modality...

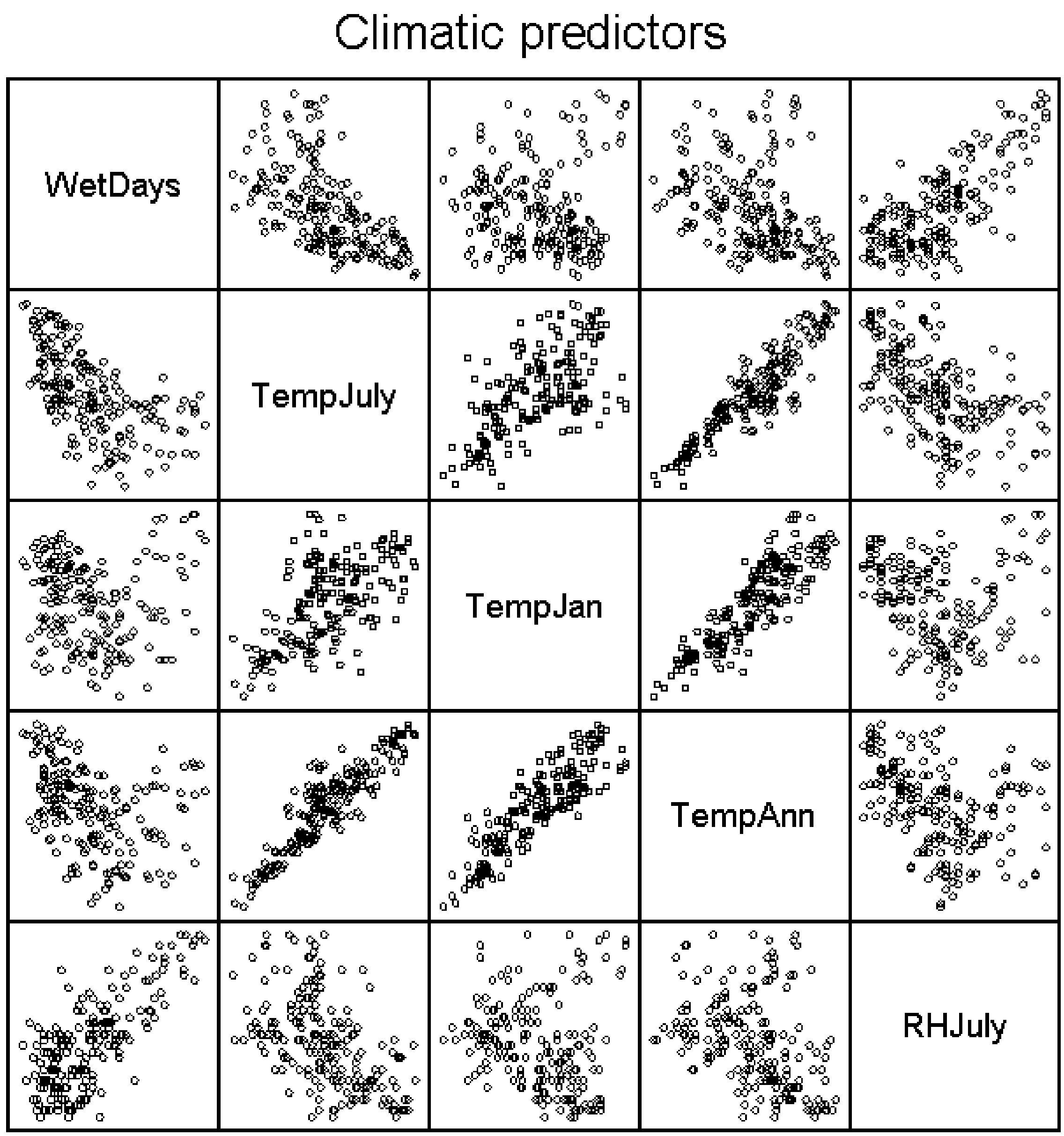
```
> hist(Xj)
```





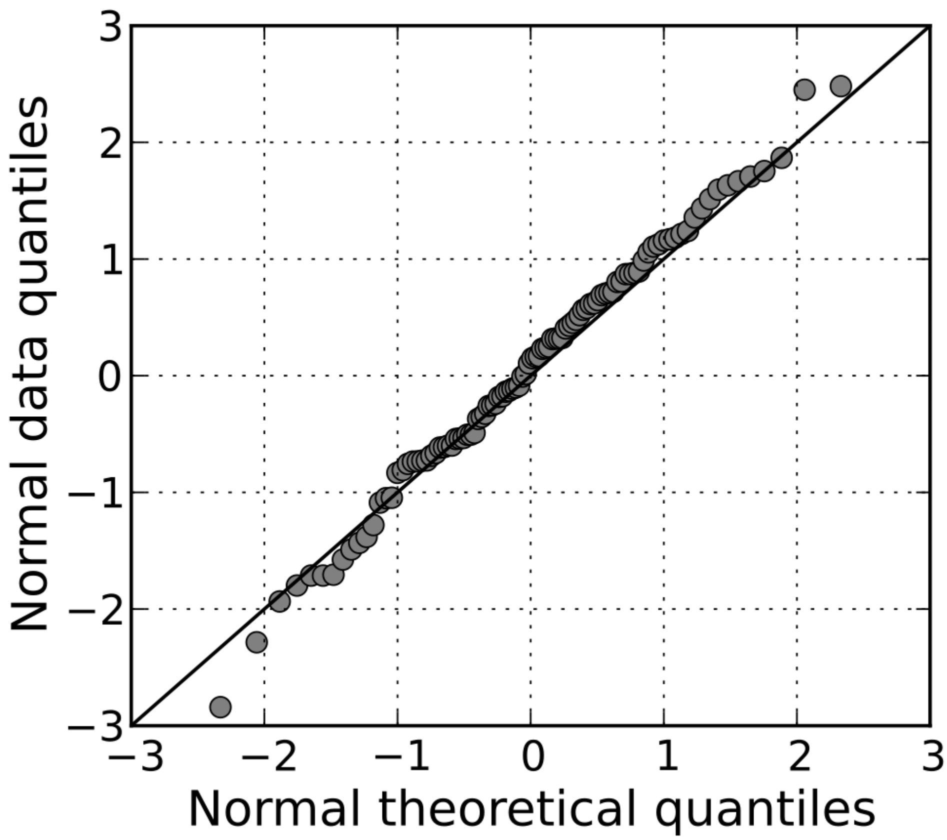
- scatter-plots : in the multi-variable case, allows to display all the correlations, 2-by-2

```
> plot(XX)
```



- q-q plots : graphic of quantiles to verify the hypothese of normality (Gaussian)

```
> qqnorm(Xj); qqline(Xj)
```



Significance and Covariates

- the 2 fundamental notions for UNDERSTANDING any statistical model
  - $\Rightarrow$ significance (bad!) and confidence intervals (better)
  - $\Rightarrow$ covariates need to be chosen judiciously (can produce false significance)

Significance Tests

Example. Compare a new and an old treatment against hypertension.

- suppose the data **seem** to indicate that the new treatment is better
- can we exclude a sampling «accident», where the new treatment was given almost exclusively to subjects in good health???
- the significance test would state that this result is very unlikely (small value of p) under the null hypothesis (= no effect)
- Conclusion (dangerous!) : the two treatments have a significant diffference at level α ($> p$).

Significance Tests : conclusion

- significance tests should be avoided (official recommendation of the ASA in 2016)
  - ⇒ at worst, they are misleading
  - ⇒ at best, they are uninformative
- producing a confidence interval (point estimate +/- error margin) is much better
  - ⇒ usually, at a 95% level
  - ⇒ "in 95% of all possible samples, the empirical estimate will lie within the error margin of the true value of the population"
  - ⇒ however, we will not repeat the sampling numerous times—this is ususally impossible... hence the interest of Bayesian approaches... (TBC)

Explanatory Variables

- we study the relationship between a variable Y and a variable X

Example. Evaluation in 4 hospitals of survival rates after a heart attack

- let the response $Y = 1$ if the patient survives, $Y = 0$ if not.

- let $X = 1, \ldots, 4$ be the identifier of the hospital

- measuring the relationship between Y and X implies here to compare the 4 hospitals in terms of the survival rate...

  $\Rightarrow$ but 1 of the 4 hospitals serves a zone with a large proportion of old patients
  $\Rightarrow$ so a direct comparison would be unfair, and inexact...

- we need to introduce a new explanatory variable , $Z = \mathrm{age}$ and measure the relationship between Y and X keeping Z constant (or by age intervals)

- a correlation can pass from positive to negative (change of sign) once the covariate Z is taken into account

$\Rightarrow$ Simpson's paradox..
$\Rightarrow$ related to causality! (TBC)

Cross Validation

- an ensemble of techniques for testing the predictive power of a statistical leanrning model

- indispensable step for validating the robustness of a model

  $\Rightarrow$ avoids the «good luck» effect

- also possible to propose confidence intervals

  $\Rightarrow$ the « bootstrap »