em lyon business school

Introduction to Machine Learning

Franck JAOTOMBO

Class rules

em lyon business school

- **Things you should know**
  - I am a **strict** Professor
    - I **will not accept you in class** after I complete attendance (5 minutes after the hour)
    - I **expect you to be focused** in class : no texting, no social networking, no distraction
    - I **expect high integrity** from each of you : no cheating, no plagiarism, submit your work and acknowledge what you have not done
  - I am very **supportive** of my students
    - I will make myself available to support you, as much as I can, including on evenings and week-ends
    - However, if you behave like a client, rather than a student, then you will get from me only that for which you have paid for : not much !

- **More things you should know**
  - Regarding your homework assignments
    - Each notebook should run from beginning to end without my having to tamper with it
    - DO NOT USE absolute paths
    - Install and or import all packages and modules required to run your notebook
  - On group assignments
    - Each contribution must be clearly and explicitly mentioned
    - A group leader must be identified for each group assignment
  - On using Generative AI
    - I have nothing against it, on the contrary
    - You must acknowledge its use
    - I will look for discrepancies between your pen and paper evaluations and your notebooks

Evaluations

- **A small project (30%)**
  - An individual report based on the aggregated Flourishing dataset
    - See the Flourishing case on Brightspace
  - Should include
    - A nicely written report
    - A python notebook
    - Due date : Oct 13, 23h00

- **A final exam (70%)**
  - On pen and paper
    - Includes concepts & codes
    - Exam date : Oct 9, 14h30

em lyon business school

Session 0 – A Broad Introduction

A high level overview of the main concepts used in Machine & Statistical Learning

Reference :

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning : With Applications in Python* (1st ed. 2023 edition). Springer.

Introduction to Data Analysis

em lyon business school

- **In Data Science, before you do anything, you first start by exploring your dataset.**
- **The basic steps in data exploration are following:**
  1. Univariate Data Analysis
  2. Bivariate Data Analysis
  3. (Modeling)

Step 1 - Univariate Data Analysis

1. If the variables are categorical.
   1. Generate the summary table for each variable.
   2. Plot their Pie Chart
   3. Plot their Bar Chart
2. If the variables are quantitative.
   1. Generate the frequency table for each variable
   2. Plot their histogram
   3. Plot their boxplot

Step 2 – Bivariate Data Analysis

1. If the variables are both categorical.
   1. Generate the contingency table
   2. Check the significance of their relationship with the chi-square test & provide Cramer's V (or Tschuprow's T)
   3. Plot their side-by-side bar charts
   4. Plot their stacked bar charts
2. If the variables are both quantitative.
   1. Compute the correlation (table)
   2. Check the significance of their relationship with the correlation test & provide the r value
   3. Plot their scatter plot (matrix)
3. If the variables are mixed categorical & quantitative.
   1. Compute the Anova table
   2. Check the significance of the difference in values between groups
   3. Plot the grouped boxplots

Modeling : supervised

- Which variables should be selected as the variable to be explained (***outcome***, target, response, dependent variable) from the others (***predictors***, features, independent variables)?
- The answer should be justified with theoretical, managerial or statistical arguments

- The goal is to ***find a function*** that captures in the best possible way the relationship between the outcome and the predictors
- This process is called “modeling” and statistical learning is one way of addressing it

- One goal of modeling is thus to explain the variability (or variance) in the outcome from the predictors.
- This approach to modeling is associated with “**supervised learning**” in statistical learning.

Modeling : unsupervised

- The variance may also be explained by the existence of subgroups of observations or subgroup heterogeneity.
- The process to account for these subgroups is associated with "**unsupervised learning**" in statistical learning.

When the number of variables is too numerous, the relationship between the outcome and the predictors can become too complex.

- To reduce this complexity and to simplify interpretability, dimension reduction is recommended.
- The set of tools to reduce dimensions is also associated with "**unsupervised learning**" in statistical learning.

Hypothesis Testing : A review

Correlation Test

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$$Statistic\ of\ the\ test \sim Student(n-2) : t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Chi-square Test

$H_0: \chi^2 = 0$

$H_1: \chi^2 > 0$

$$Statistic\ of\ the\ test \sim Chi\ square[(r-1)(c-1)] : X^2 = \sum_{i,j} \frac{\left[n_{ij} - \frac{r_i c_j}{n}\right]^2}{\frac{r_i c_j}{n}}$$

Anova Test

$H_0: \mu_1 = \mu_2 = \cdots = \mu_c$

$H_1: All\ the\ \mu_j\ are\ not\ equal$

$$Statistic\ of\ the\ test \sim Fisher\ (c-1, n-c) : F_{stat} = \frac{MSB}{MSW}$$

Hypothèses Testing 1

em lyon business school

Critical Value approach

1. State the null hypothesis, H_0 and the alternative hypothesis, H_1.
2. Choose the level of significance, α, and the sample. The level of significance is based on the relative importance of Type I error in the situation.
3. Determine the appropriate test statistic and sampling distribution.
4. Determine the critical values that divide the rejection and nonrejection regions.
5. Compute the value of the test statistic.
6. Make the statistical decision and state the managerial conclusion in the context of the theory, claim, or assertion being tested.
   - If the test statistic falls into the nonrejection region, you do not reject the null hypothesis H_0.
   - If the test statistic falls into the rejection region, reject the null hypothesis.

Example : test of correlation

1. $H_0: cor(X,Y) = \rho = 0$
   $H_1: cor(X,Y) = \rho \neq 0$

This is a case of a two-tailed Hypothesis Testing

2. $\alpha = 0.05$
3. $t_{stat} = r\sqrt{\frac{n-2}{1-r^2}} \sim Student(n-2)$

where $r = cor(\hat{X},\hat{Y})$ is estimated from the data

4. Obtain $t_{\alpha/2}^{n-2}$ from Excel
5. Obtain t_{stat} from the data
   - Use Excel / XLSTAT to compute r
6. Make a statistical decision

Linear Correlation Test (Pearson)

- We want to test if there is a significant association between two continuous variables

- The covariance between two variables X and Y indicates if there is an association between the variation of the two variables around their respective means

$$cov(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

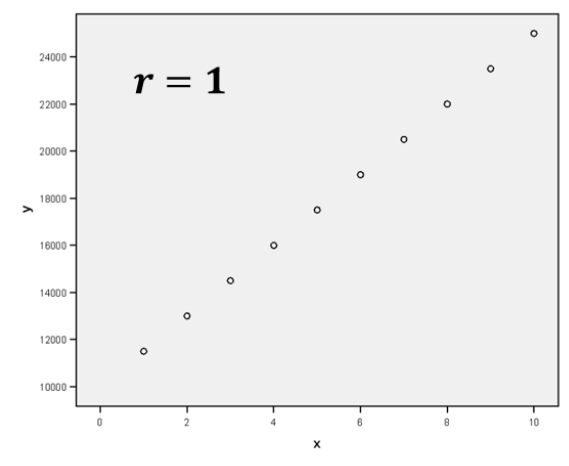
- Correlation is a standardized measure of the covariance

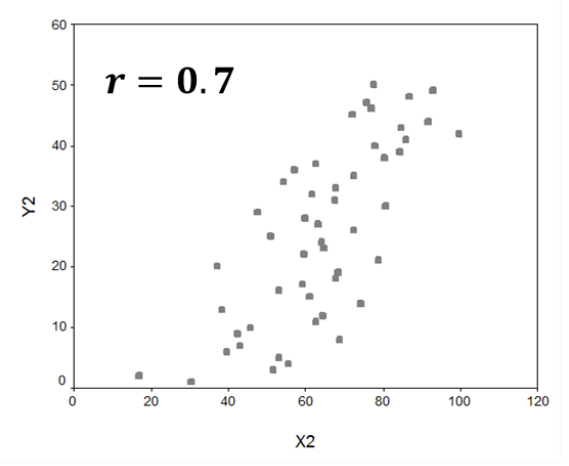
$$r = \frac{cov(X,Y)}{s_X s_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

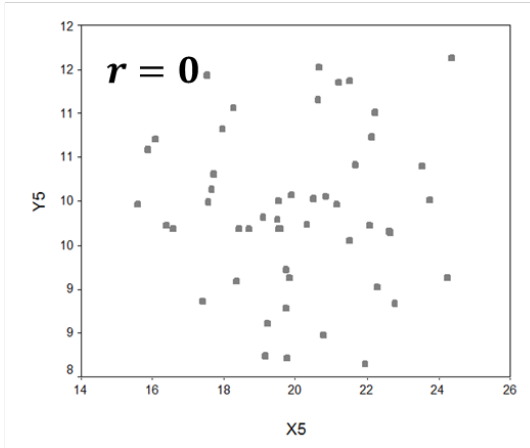
s_X and s_Y are the respective standard deviations of X and Y

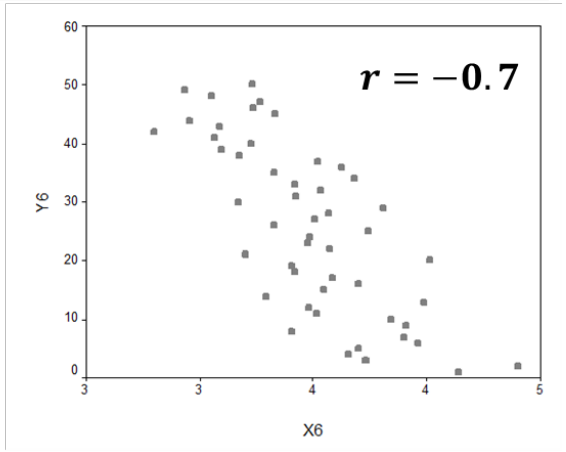
Correlation : size effect

- **The population coefficient of correlation is referred as ρ.**
- **The sample coefficient of correlation is referred to as r.**
- **Either ρ or r have the following features:**
  - Unit free.
  - Range between –1 and 1.
  - The closer to –1, the stronger the negative linear relationship.
  - The closer to 1, the stronger the positive linear relationship.
  - The closer to 0, the weaker the linear relationship.
- **Rule of Thumb (does not replace a test)**
  - ρ is significant if $|r| \geq \frac{2}{\sqrt{n}}$









Hypothèses Testing 2

em lyon business school

P-value approach

1. State the null hypothesis, H_0 and the alternative hypothesis, H_1.
2. Choose the level of significance, α. The level of significance is based on the relative importance of the risk of a type I error.
3. Determine the appropriate test statistic and sampling distribution.
4. Compute the value of the test statistic
5. Compute the p-value.
6. Make the statistical decision and state the managerial conclusion in the context of the theory, claim, or assertion being tested. If the p-value is < α reject H_0.

Example : Chi-square test of independence

1. $H_0: \chi^2 = 0 \;\; H_1: \chi^2 > 0$

This is a case of a one-tailed Hypothesis Testing

2. $\alpha = 0.05$

3. $X^2_{stat} = \sum_{i,j} \frac{\left[n_{ij} - \frac{r_i c_j}{n}\right]^2}{\frac{r_i c_j}{n}} \sim \chi^2 (R-1)(C-1)$

where

R is the number of rows and C number of columns

4. Obtain $\chi_\alpha^{(R-1)(C-1)}$ from Excel
5. Obtain *p value* from Excel
6. Make a statistical decision

Chi square

- **If the row and column variables are independent:** $E(n_{ij}) = \frac{r_i c_j}{n}$

$$X^2 = \sum_{i,j} \frac{\left[n_{ij} - \frac{r_i c_j}{n}\right]^2}{\frac{r_i c_j}{n}}$$

	X_1	X_j	X_c	$Total$
Y_1	n_{11}		n_{1c}	r_1
Y_i		n_{ij}		r_i
Y_l	n_{r1}		n_{rc}	r_l
$Total$	c_1	c_j	c_c	n

- r_i and c_j indicate respectively the total (marginal) frequency of row i and column j
- X^2 follows a Chi Square distribution with a degree of freedom = $(r-1) \times (c-1)$
  - where ***r* = number of modalities on rows** and ***c*= number of modalities on columns**
  - We need only to compare X^2 with the threshold values of the Chi Square (χ^2) distribution

- **Effect size :** $\phi = \sqrt{\frac{X^2}{n}}$; $V_{Cramer} = \sqrt{\frac{X^2}{n.\min[(r-1),(c-1)]}}$; $\mathrm{T}_{\mathrm{Tschuprow}} = \sqrt{\frac{X^2}{n\sqrt{(r-1)\times(c-1)}}}$
  - Use ϕ for 2×2 tables

Contigency Table : Interpretation

- **The residuals contigency table**
  - Provides information on the nature of the relationship between the contingency tables
  - $e_{ij} = \frac{\left[n_{ij} - \frac{r_i c_j}{n}\right]}{\sqrt{\frac{r_i c_j}{n}}} \sim \mathcal{N}(0,1)$
  - Look for cells where $|e_{ij}| \geq 1.96$
  - Condition : $E(n_{ij}) > 5$
- **Effect size**
  - $\phi = \sqrt{\frac{X^2}{n}}$ and $V_{Cramer} = \sqrt{\frac{X^2}{n.\min[(l-1),(c-1)]}}$

Example :

Relationship between smoking and having a lung cancer

Observed Values		Lung Cancer		
		Yes	No	Total
Smoking	Yes	75	8	83
	No	13	52	65
	Total	88	60	148

Expected Values		Lung Cancer		
		Yes	No	Total
Smoking	Yes	49.35	33.65	83
	No	38.65	26.35	65
	Total	88	60	148

Residuals		Lung Cancer		
		Yes	No	Total
Smoking	Yes	3.65	-4.42	
	No	-4.13	5.00	
	Total			
X^2 =	74.8667105			
p-valeur =	5.0359E-18			

Analysis of Variance

$$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(X_i^j - \bar{\bar{X}}\right)^2$$

$$SSB = \sum_{j=1}^{c} n_j\left(\bar{X}_j - \bar{\bar{X}}\right)^2$$

$$SSW = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(X_i^j - \bar{X}_j\right)^2$$

$$MST = \frac{SST}{n-1}$$

$$MSB = \frac{SSB}{c-1}$$

$$MSW = \frac{SSW}{n-c}$$

$$Grand\ mean:\ \bar{\bar{X}} = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\frac{X_i^j}{n}$$

$$Group\ mean: \bar{X}_j = \sum_{i=1}^{n_j}\frac{X_i^j}{n_j}$$

In Practice :

1. Test the existence of a difference in group mean.
2. If H_0 is rejected i.e. there is at least two significantly different mean values between two categories of the grouping variable. Then apply Post Hoc Tests.
3. Post Hoc Tests are two by two comparison of the means of each category
4. It requires some adjustments in p-values and there are several ways of addressing this
5. If in doubt apply Bonferroni correction

Exercise

- **Explore the <Flourishing> dataset**
  - Read the instructions in the <Flourishing_Case.docx> document