Introduction to Machine Learning

Franck JAOTOMBO

Introduction to Data Analysis

- 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

1. 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
    3. Plot their side-by-side bar charts
    4. Plot their stacked bar charts

1. 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)

1. 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 ***outcome*** (target, response, dependent variable) or variable to be explained 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 individual instances 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

$H_0\colon\ \rho = 0$

$H_1\colon\ \rho \neq 0$

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

Chi-square Test

$H_0\colon\ \chi^2 = 0$

$H_1\colon \chi^2 > 0$

$$Statistic\ of\ the\ test\ \ - Chi\ square[(r-1)(c-1)]\colon t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Anova Test

$H_0\colon\ \mu_1 = \mu_2 = \cdots \mu_k$

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

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

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

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 = $(r-1)*(c-1)$
  - 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 :** $\phi = \sqrt{\frac{X^2}{n}}$ **and** $V_{Cramer} = \sqrt{\frac{X^2}{n.\min[(r-1),(c-1)]}}$

Analysis of Variance

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

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

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

Exercise

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