MACHINE LEARNING FROM DATA

Lab Session 0 – Exploratory data analysis

1. Goal

The goal of this session is to

- learn how to do basic data exploration
- become familiar with Python functions for exploratory data analysis from NumPy, SciPy, Pandas, Matplotlib and Seaborn libraries
- analyze the gaussianity of synthetic data
- explore a simple dataset

2. Instructions

- Download and uncompress the file **Mlearn_lab0.zip**
- Follow instructions in this guide **Mlearn_lab0_guide_python.pdf**
- Answer the questions in a document **Mlearn_lab0_report_team_surnames.docx**
- Write the new code in a Colab Notebook **Mlearn_lab0_3_team_surnames.ipynb**. Make sure that you include the code to import all necessary libraries to run the Notework.

3. Introduction and previous study

Read the document **Mlearn_EDA_Python.pdf** to understand the following concepts and methods:

- Histograms
- Box-plots
- Measures of central tendency
- Measures of dispersion: skewness and kurtosis
- Distribution plots: normal probability plots and cumulative distribution plots
- Scatter Plots
- Confidence interval for the mean
- Hypothesis tests: testing goodness of fitness of a distribution

4. Gaussianity analysis of synthetic distributions

In this section we will use EDA tools to analyze the gaussianity of four data samples.

If you are not familiar with Google Colab, first follow this tutorial:
https://colab.research.google.com/notebooks/welcome.ipynb?hl=en

Load **Mlearn_lab0_1.ipynb** in a Google Drive folder and open it in Google Colaboratory

Read the code in the notebook **Mlearn_lab0_1.ipynb**, run, analyze and compare the measures obtained for the four distributions: normal, raleigh, laplacian and uniform.

Q1. Briefly describe the conclusions of your analysis (you can insert plots)

5. Observation of the iris dataset

The Iris dataset is available from the UCI Machine Learning Repository
http://archive.ics.uci.edu/ml/datasets/Iris

The dataset contains 3 classes of 50 instances each, wheare each class refers to a type of iris plant: Iris Setosa, Iris Versicolour and Iris Virginica.
Each sample is represented with 4 features:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm

Load **Mlearn_lab0_2.ipynb** in a Google Drive folder and open it in Google Colaboratory.
Read the code in the notebook. Run, analyze the plots and answer the following questions.

Q2. For each class and each feature, analyze histograms, cdfs and normal plots. Can we assume a Gaussian distribution for any of the features?

Q3. Analyze kurtosis and skewness values for each feature and class.

Q4. Edit the notebook to create the boxplots for all the features. Analyze boxplots by feature. Are there significant differences between the classes?

Q5. Analyze the scatter plot. Are features related in any way? What can you say about the separability of the classes?

Create a new Colab Notebook: **Mlearn_lab0_3_yourname.ipynb**.

Q6. Choose one feature (among the four available), write the code to compute the feature mean and confidence intervals at confidence levels 95%, 99% and 99.9% for the three classes.

Compute the confidence values:

(a) 'Manualy', computing the t-critical values, sample mean and sample standard deviation.

Notes:
- Use Numpy *mean* and **std**, to compute sample mean and sample standard deviation
- The t-distribution is available in **scipy.stats** with the nickname *'t'*, so we can get t-critical values with *stats.t.ppf()* (the Percent Point Function or inverse of CDF).
- When using the t-distribution, you have to supply the degrees of freedom (df). For this type of test, the degrees of freedom df is the sample size minus 1.

(b) using the Python function *stats.t.interval( )*.

Fill the following table:

	Mean	CI at 95%	CI at 99%	CI at 99,9%
Class 1				

Class 2				
Class 3				

Q7. Write the code to conduct the following hypothesis tests, using the Shapiro-Wilk test and the Anderson Darling test, for **all** the features K and classes J.

- Null hypothesis H_0: Feature K from class J comes from a Gaussian distribution at the significance level α

Note: use *shapiro( )* and *anderson( )* functions from *SciPy.stats*.

For each test complete the corresponding table with the decisions (acceptance/rejection) for the null hypothesis H_0 (feature Gaussianity), and the p-value or the critical and statistic values, respectively, for α = 0,05 and α = 0,01

Explain the meaning of the p-value / critical value and interpret the results accordingly.

Table for Shapiro-Wilk test

Feature #	Acceptance / rejection of H_0	*p*-value
class 1		
class 2		
class 3		

Table for Anderson Darling test

Feature #	Acceptance / rejection of H_0	critical value	stat
class 1			
class 2			
class 3			

6. Feature Engineering

Load **Mlearn_lab0b_feature_engineering.ipynb** in a Google Drive folder and open it in Google Colaboratory. Read and run the code.