MACHINE LEARNING FROM DATA
Fall 2018

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 Matlab functions for exploratory data analysis
- analyze the gaussianity of synthetic data
- explore a simple dataset

2. Instructions

Getting the material:

- Download and uncompress the file **ML_Lab0.zip**

Handling your work:

- Answer the questions in a document **Lab0_report_yourname.docx**
- Save the report, convert to pdf and upload the **pdf** file

3. Introduction and previous study

Read the document ML_MET_EDA.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.

Read the code in **lab0_synthetic.m**, run the script, analyze and compare the measures obtained for the four distributions: normal, raleigh, laplacian and uniform.

Note: you could also use a user interface 'randtool' to generate random samples from many distributions, change their parameters and export the sample to a variable in the workspace.

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. Petel length in cm
4. Petal width in cm

Read the code in **ML_Lab0_irisdataset.m**
Run the script, 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. 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?

Q6. Edit the script ML_Lab0_irisdataset.m. Choose one feature (among the four available) and compute the feature mean and confidence intervals at confidence levels 95%, 99% and 99.9% for the three classes.

Hint: use Matlab functions `tinv` and `var`

	Mean	CI at 95%	CI at 99%	CI at 99,9%
Class 1				
Class 2				
Class 3				

Q7. Copy the code used to answer Q6.

Q8. Choose one feature K (among the four available). Edit the script ML_Lab0_irisdataset.m to conduct the following hypothesis tests, using a chi-squared test

- Null hypothesis H_0: Feature K from class 1 comes from a Gaussian distribution at the significance level 0.001
- Null hypothesis H_0: Feature K from class 2 comes from a Gaussian distribution at the significance level 0.001
- Null hypothesis H_0: Feature K from class 3 comes from a Gaussian distribution at the significance level 0.001

Complete the following table with the decisions (acceptance/rejection) for the null hypothesis H_0 (feature Gaussianity), p-value and degrees of freedom for $\alpha = 0{,}001$.

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

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

Q9. Copy the code used to answer Q8.