Departamento de Electrónica, Telecomunicações e Informática

EXPLORAÇÂO de DADOS

Homework I: Exploratory Data Analysis

1 Background

Tasks that can be considered in a data mining project are feature selection and dimension reduction. Both have the goal of decreasing the number of inputs to the predictive (classification) models.

2 HOME WORK

1. Answer to the questions published in Exercises (Notes). Choose one of the following possibilities
   - Exercises 2.1 and 2.3
   - Exercises 2.2 and 2.4
2. Programming task. Choose one of the following tasks.
   - Task A: Consider the data set `https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope`. The data set has 10 numerical features and one categorical feature.
     (a) Select Features using one the following strategies
       – With the Pearson's correlation coefficient select is possible to identify the two best numerical features?
       – Rank the numerical features using ANOVA. The best two are discriminative?
     (b) Reduce the dimension of the feature vector from 10 to 2 using the PCA model and KPCA model (and RBF kernel). The projected data allows to discriminate between the two classes?
   - Task B: Load the digit data set `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html#sklearn.datasets.load_digits` and select a subset with two digits and 10 examples of each. The 20 examples data set should be centered and then
     – calculate the dual form of the PCA model. Suggestion: use the linear algebra manipulations.
     – with scikit-learn estimate the PCA model.
     – how can you compare the two models? Illustrate this issue commenting your notebook.

  - Calculate the values of the projections onto the two principal directions. The two projections allow to discriminate the two classes?

Note that questions are tips to comment the simulations.

The Jupyter Notebook and the Exercises should be download in Elearning.

3 Useful Links

- Pandas. Data structures and data analysis tools for the Python programming language.
  https://pandas.pydata.org/pandas-docs/stable/getting_started/index.html
- Numpy. Multi-dimensional arrays and matrices.
  https://www.tutorialkart.com/numpy-tutorial/
- Dimension reduction
  https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html https://scikit-learn.org/stable/auto_examples/decomposition/plot_kernel_pca.html

REMARK: Whenever justified the group can consider to include the contribution of each group member to the homework.