CS659, Spring 2015
Homework # 1. Due 2/17/15, before the start of the lecture (hardcopy).

This is an individual assignment.

1) Given the dataset found in file z
   a) Use the first 99 records to compute the covariance matrix Σ
   b) The centroid of those 99 records, which is computed as a record whose attributes values are the average of the other record values
   c) Using the result of a), compute
      i) The Mahalanobis distance of records 100, 101 and 102 to the centroid computed in b)
      ii) Repeat i) but compute the Euclidian distance
      iii) Compare and explain the differences between the results in i) and ii) Submit a graph and your brief comments.
   d) Using the first 99 records, produce a dataset with nominal (categorical) attributes, where each of the two attributes has 3 values. Use equi-width, and equi-depth for this task.

2) This assignment will be using ***Weka*** data mining tool. ***Weka*** is an open source Java development environment for data mining from the ***University of Waikato in New Zealand***. It can be downloaded freely from ***http://www.cs.waikato.ac.nz/ml/weka/***
Heart disease datasets. The dataset studied is the ***heart disease*** dataset from ***UCI repository***. Two different datasets are provided: ***heart-h.arff*** (Hungarian data), and ***heart-c.arff*** (Cleveland data). These datasets describe factors of heart disease. You can also find the data and information about it here.

   (The ***question*** on which this machine learning study concentrates is whether it is possible to predict heart disease from the other known data about a patient. The ***data mining*** task of choice to answer this question will be classification/prediction, and several different algorithms will be used to find which one provides the best predictive power. However this exercise focuses on the various aspects of the KDD process, **not on the prediction problem.**)

   a) Data Preparation and integration

   We want to merge the two datasets into one, in a step called data integration. Revise the ***arff*** notation from the tutorial, which is ***Weka*** data representation language. Once you understand the data, merge the two datasets into a single one, solving all the problems you encounter. Document what you have done.

   b) Descriptive data summarization

   Before preprocessing the data, an important step is to get acquainted with the data – also called ***data understanding***.

   i) Stay in the ***Preprocess*** tab for now. Study for example the ***age*** attribute. What is its ***mean***? Its ***standard deviation***? Its ***min*** and ***max***?

ii) Provide the ***five-number summary*** of this attribute. Is this figure provided in Weka?

iii) Switch to the ***Visualize*** tab. What is the term used in the textbook to name the series of boxplots represented? By selecting the maximum jitter, and looking at the ***num*** column – the last one – can you determine which attributes seem to be the most linked to heart disease? Paste the ***boxplot*** representing the attribute you find the most predictive of heart disease (Y) as a function of ***num*** (X). Does any pair of different attributes seem correlated?

c) **Data preparation and cleaning**

Data cleaning deals with such defaults of real-world data as incompleteness, noise, and inconsistencies. In ***Weka***, data cleaning can be accomplished by applying ***filters*** to the data in the ***Preprocess*** tab.

i) Remove the missing values with the method of your choice, explaining which filter you are using and why you make this choice. Document whatever you did in your homework solution. **Extra credit:** If a filter is not available for your method of choice, develop a new one that you add to the available filters as a Java class. (that should be exciting and fun … send me an email if you plan to do this)
ii) Save the clean dataset into heart-cleaned.arff. Paste a screenshot of the first 10 rows of this set in your homework solution.

d) **Data preparation: transformation**

Use the Weka filters to do the following:

i) ***Attribute construction*** – for example adding an attribute representing the sum of two other ones. Which ***Weka filter*** permits to do this?

ii) ***Normalize*** an attribute. Which ***Weka filter*** permits to do this? Can this filter perform Min-max normalization? Z-score normalization? Decimal normalization? Provide detailed information about how to perform these in ***Weka***.

iii) ***Normalize*** all real attributes in the dataset using the method of your choice – state which one you choose.

iv) Save the normalized dataset into ***heart-normal.arff***, and paste here a screenshot showing at least the first 10 rows of this dataset – with all the columns.

e) **Data preparation: reduction**

Beside attribute selection, a reduction method is to select rows from a dataset. This is called sampling. How to perform sampling with ***Weka filters***? Can it perform the two main methods: ***Simple Random Sample Without Replacement***, and ***Simple Random Sample With Replacement***? Report your findings

3) For the Cleveland heart data, use Weka to perform Principal Component Analysis
   a) Keep only attributes (in the new space) that represent up to 70% of the total variance

b) Remove the rest of the attributes and save the resulting data set.
c) Reconstruct the approximate original dataset

d) Compute the distance (Euclidian) between every vector in the original Cleveland set and the approximate one. Report the average and standard deviation of the set of distances. Report the three records with highest distance between the approximation and the original.