COMPUTATIONAL BIOLOGY

Instituto Superior Técnico

2017/2018

BC_LAB# 5 – UNSUPERVISED/SUPERVISED LEARNING

This laboratory will use the data from the article "*Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*", by Alizade et. Al., NATURE, VOL 403, n. 3, 503-511, 2000, available at http://eps.upo.es/bigs/datasets.html.

From the diverse datasets available, select the reduced database, with 45 instances of 4026 genes each, in format ARFF (Reduced database (45 instances x 4026 genes) in ARFF format, with two labelled classes (Germinal Centre, GCL, and Activated, ACL) [1Mb]).

Group I (8 points)

The K-means algorithm

1.1 Using the programming language of your choice, write code for the K-means algorithm. Your program should accept at least four input parameters: three integers and a file in txt format.

Input parameters:

Number of clusters (G)
Number of rows (R)
Number of columns (C)

Input file example:

```
@RELATION fff
@ATTRIBUTE GENE1835X REAL
@ATTRIBUTE GENE1836X REAL
@ATTRIBUTE class    { P, N }
@DATA
0.25, 0.30, P
0.34, 0.33, P
-0.33, -0.44, N
-0.3, ?, N
-0.44, -0.48, N
```

In this example the input parameters R and C are equal to 5 and 2, respectively.
Each row R has C real numbers separated by a comma, followed by a label that is a string. Example: 0.25, 0.30, P

The question mark character (?) identifies missing values. Missing values should be removed before running the algorithm, using a technique of your choice. For the K-means algorithm ignore the classification P or N at the end of each line.

After running the K-means algorithm using the provided dataset example, use the P or N classification to compute the precision and recall of the algorithm.

Group II (8 points)

Using the Weka package (http://www.cs.waikato.ac.nz/~ml/weka/downloading.html)

Use the file that is made available in ARFF format, which corresponds to the data in the reduced database.

2.1 Clustering using K-means

a) Use the K-means available in the Weka package to cluster the data into two classes. Report the centroids of the clusters.

b) Run your K-means algorithm in this dataset (arff.txt) and compare the results with the results obtained by Weka.

2.2 Supervised classification

Use the J48 classifier and the Naïve Bayes classifier within the Weka package to build classifiers for this dataset.

a) Report the precision of each classifier

b) Describe succinctly each of the obtained classifiers

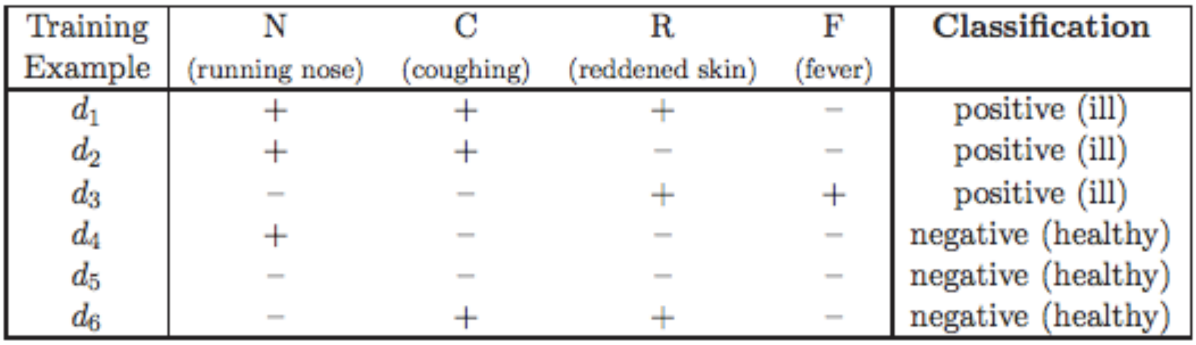
c) Compare the precision of the obtained classifiers with the precision of the clustering method obtained in 2.1, assuming that to each cluster is assigned the most frequent class.

Group III (4 points)

Consider the problem where the task is to describe whether a person is ill. We use a representation based on three features per subject to describe an individual person. These features are “running nose”, “coughing”, and “reddened skin”, each of which can take the value true (‘+’) or false (‘–’), see Table 1.

(a) Given the data set in Table 1, determine all probabilities required to apply the naive Bayes classifier for predicting whether a new person is ill or not.

(b) Apply the naive Bayes classifier to the test patterns corresponding to the following subjects: a person who is coughing and has fever, a person whose nose is running and who suffers from fever, and a person with a running nose and reddened skin (d7 = (N(-), C, R(-), F), d8 = (N, C(-), R(-), F), and d9 = (N, C(-), R, F(-))).



Training Example	N (running nose)	C (coughing)	R (reddened skin)	F (fever)	**Classification**
d_1	+	+	+	−	positive (ill)
d_2	+	+	−	−	positive (ill)
d_3	−	−	+	+	positive (ill)
d_4	+	−	−	−	negative (healthy)
d_5	−	−	−	−	negative (healthy)
d_6	−	+	+	−	negative (healthy)

Table 1: List of training instances.