Math 6365 first homework HW1

due date = Thursday feb 8th at midnight

This HW1 is a preliminary study of a concrete classification task of your choice, to prepare for "solving" this specific classification task in HW2 by implementation of MLPs

0)

select a publicly available data set associated to a clearly described CLASSICATION TASK,

1)

begin your report by writing your own description of the CLASSIFICATION TASK involved :

Outline a word description of "cases"; give the number N of cases

Give the number p of "descriptors" for each case. Outline the concrete meaning of each descriptor

Describe concretely each "class" C1 C2 … Ck involved in the classification task. Give the size |Cj| of each class Cj and the number k of classes. If k is larger than 5 , eliminate the classes C6 C7 … Ck as well a sall the cases corresponding to classes C6 C7 … Ck

Indicate how the correct classification of each case was originally obtained and how it is encoded in the data base

Indicate if there are any publication analyzing the classification task associated to this data set

2) A descriptor is called "discrete" or "categorical" if it can take only a "small" finite set of distinct values, which may be indexed by numerical values, or by distinct "names" in various formats. Of course one can always re-index this finite set of values by arbitrarily selected numbers. Examples : job type, accident severity, car type, etc.

A descriptor is called "continuous" if its set of potential values includes a full segment of real numbers. Examples: weight, length, width, volume, etc.

For each descriptor D1…Dp, indicate if it is categorical or continuous.

For each continuous descriptor compute and display its range , its min and max values, its mean and standard deviation. Compute and display its histogram

For each categorical descriptor indicate the finite list of distinct values , and the corresponding frequencies of these values in the whole data set

3)

For each descriptor D , and for each class Cj, compute the histogram HISTj of D values for the cases belonging to class Cj. If graphically feasible, display the k histograms HIST1 … HISTk on a single graph. Interpret these informations to evaluate intuitively if the descriptor D will be helpful to discriminate between some pairs of classes.

One can use “Kolmogorov-Smirnov tests” to compare pairs of histograms HISTi and HISTj in order to decide if they are strongly distinct or not.

For each continuous descriptor D compute and display the mean Mj and standard deviation Sj of D values for the cases belonging to class Cj. One can then use classical “t-tests” to compare pairs of means Mi and Mj to decide if |Mi –Mj| is “strongly “different from 0 or not ,

Both the t-tests and the Kolmogorov-Smirnov test are easy to use functions available in Matlab as well as in R

4)

For each case CASn with n = 1,2, ..., N regroup all the continuous descriptors of CASn in a single vector Vn of fixed dimension r. Implement a PCA analysis for the cloud of N vectors V1,V2,… VN. Explain the PCA method in your report. Compute and display graphically the r eigenvalues $S1 \geq S2 \geq \ldots \geq Sr$ generated by the PCA. Compute and display graphically the ratios $RATj = (S1 + S2 + \ldots + Sj) / (S1 + S2 + \ldots + SN)$. Determine the smallest j such that $RATj \geq 0.95$.

5)

Call W1 … Wr the eigenvectors associated to S1 , S2 , … , Sr. Compute and display graphically the orthogonal projections of the N vectors V1,V2,… VN on the 3 dimensional vector space generated by the othonormal basis {W1 , W2, W3}. Use different colors for these projected points namely 1 color per class. Interpret the results in your report

Do the same operation for the 3 separate projections onto the three planes {W1 , W2}, {W1 , W3}, and {W2, W3}. Interpret the results in your reports.