MA 3022/4022/7022 Data Mining and Neural Networks

Computational Task 1

Due till 17.02.2021

100 marks available

Breast cancer diagnosis

Using Breast Cancer Wisconsin (Diagnostic) Data Set, kNN method for k=1 and 3, and Fisher's linear discriminant

Go to the machine learning depository and read the data description

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

1. Have a look on the original paper

W.H. Wolberg, W.N. Street, D.M. Heisey, and O.L. Mangasarian. Computer-derived nuclear features distinguish malignant from benign breast cytology. Human Pathology, 26:792--796, 1995
http://dollar.biz.uiowa.edu/~nstreet/research/hu_path95/hp95.pdf

There is no need to study it in detail but you should answer the questions:

a) What is the problem authors aimed to solve?
b) Which methods did they use?
c) How did they test the accuracy of classification?

(**10 marks**)

2. Take the data table
http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data
How many attributed does it include? How many malignant cases (M) and benign cases (B) are there? For each real attribute find its mean, variance, standard deviation and mean, variance and standard deviation in for each attribute in each class. Are the attributes normalised? If not, then normalise to unit variance.
(**10 marks**)

3. Try to create predictors by one attribute. For this purpose, create histograms for each attribute and each class and select the best threshold a for each attribute x for the decision rule: if $x>a$ then one class (B or M) and if $x<a$ then another class (M or B) (**the optimal cut**). Find the classification error for each attribute. Which attribute gives the best prediction? Arrange the attributes in their prediction ability
(**20 marks**)

4. Test 1NN and 3NN classification rules. Present the classification errors. Which rule is better? (Use the normalised attributes)
(**20 marks**)

5. Find in the literature description and explanation of Fisher's linear discriminant. Read, understand and write a comprehensive description of the algorithm with main formulas and explanation (not more than 1 page!)
(**10 marks**)

6. Apply Fisher's linear discriminant to the Breast Cancer Wisconsin (Diagnostic) data set. Analyse the quality of classification. Compare to 1NN and 3NN methods.
(**20 marks**)

7. Prepare and submit a report with pictures and tables.
(**10 marks for the quality of the report**)

Expected volume of the account is ~7 pages.
Attach your programs, Excel spreadsheets, tables. Use histograms and other tools for visualisation. Keep your intermediate results, you will need them for the next task.