COMP3308 Artificial Intelligence
Semester 1, 2010

Assignment 2. Classification of Biomedical Data

Submission deadline: Friday, week 11, 5pm (21 May).
Late policy submission: A penalty of -1 mark will apply for each day late and the assignment will not be accepted if it is submitted more than 7 days after the due date.
This assignment can be completed individually or in pairs. Working in pairs is encouraged. Both students will receive the same mark.
Submission instructions: You need to submit:

1) Hard copy (report + code + plagiarism cover sheet) in the locker COMP3308/3608 (School of IT building, level 1, in the undergraduate labs wing, close to the room where the AI tutorials are held) and

2) Electronic copy (report + code + the two data files: `pima1.csv` and `pima2.csv`) via eLearning. All files should be zipped together in a single file. The zip file should be named 0123456.zip, where 0123456 is your SID. In case of a pair submission, put both SIDs separated by an underscore: 0123456_0789123.zip. Only one of the two students needs to submit.

Programming language: You can write the program in a language of your choice (e.g. Java, C, C++, Python, Matlab) but we need to be able to test your code on the University machines. You need to include instructions how to run your code.
Weight: This assignment is worth 18 marks = 18% of your final mark

The goal of this assignment is to 1) implement the Naïve Bayes algorithm and the cross validation method, 3) evaluate the classification performance of the implemented Naïve Bayes and other classifiers from Weka on a real dataset, 4) investigate the effect of feature selection, in particular the Correlation-based feature selection method (CFS) from Weka.

1. Download the dataset from the UCI Machine Learning Repository
Go to the UCI Machine Learning Repository http://archive.ics.uci.edu/ml/ and download the *Pima Indian Diabetes* dataset. There are 2 files associated with each dataset: *.names describing the data (e.g. the number and values of the attributes, number and names of the classes, and their meanings) and *.data containing the data. **Your task is to predict the class (class 0 or class 1).**

2. Pre-process the dataset
--Read the *.names file and learn more about the meaning of the attributes and the class.
--Convert the data into csv format. Add a header line based on the information from the *.names file. Convert the class variable from numeric to nominal, e.g. 0->class0, 1->class1.
--Take a look at the features. Are there missing values? If so, apply a method to handle this. You can do this in Weka; it has an in-build filter to do this. Weka also can read csv files.
- Normalize the values for each attribute. You can do this in Weka, it has an in-build filter for this.
- Save the preprocessed file as `pima1.csv`.

3. Implement the Naïve Bayes algorithm. As the features are numeric, you need to implement the version for numeric attributes using probability density function. Assume normal distribution, i.e. use the probability density function for normal distribution. Your program should be able to read the csv file.

4. Implement 10-fold stratified cross validation for evaluating the performance of the Naïve Bayes classifier. For each 10-fold cross validation run, your program should print the accuracy of the Naïve Bayes on the test set (the 1 fold not used for training), and at the end the average accuracy over the 10 runs.

5. Apply feature selection using CFS

CFS [1] is a method for selecting a subset of the original attributes It searches for the "best" subset of features where "best" is defined by a heuristic which takes into consideration two criteria: 1) how good the individual features are at predicting the class and 2) how much they correlate with the other features. Good subsets of features contain features that are highly correlated with the class and uncorrelated with each other.

[1] Hall, M.: Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. 17th Int. Conf. on Machine Learning (ICML). Morgan Kaufmann (2000) 359-366. http://waikato.researchgateway.ac.nz/handle/10289/1024

Apply the CFS feature selector to reduce the number of features. It is available from the "Select attributes" tab in Weka. Use "Exhaustive Search" as a search method. Save the csv file with the reduced number of attributes (can be done in Weka) and name it `pima2.csv`.

6. In WEKA select 10-fold cross validation (it is actually 10-fiold *stratified* cross validation) and run ZeroR, 1R, k-Nearest Neighbor (k-NN), Decision Tree (DT) and Multi-Layer Perceptron (MLP) and Naïve Bayes from Weka (NB). Compare their performance with your Naïve Bayes. Do this for the case without feature selection and with CFS feature selection.

7. Write a report (similar to a research paper) describing your analysis and findings. It should include the following sections

1) Aim – briefly state the aim of your study (e.g. predicting X based on Y etc.) and write a paragraph why the problem is important.

2) Data
   - Data set used – Briefly describe the dataset. Mention the number of attributes and classes.
   - Data preparation – Provide a brief summary of the preprocessing applied.
   - Attribute selection – Briefly describe the CFS method. List the selected attributes by CFS.

3) Results and discussion
   --Results – Present the accuracy results (in %, using 10-fold cross validation) in the following table where MyNB is the Naïve Bayes you implemented

Accuracy on test set [%]

	ZeroR	1R	1-NN	3-NN	DT	MLP	NB	MyNB
no feature selection								
CFS								

   --Discussion – compare the performance of the classifiers, with and without feature selection, compare your NB with the Weka's NB, discuss the effect of the feature selection – does CFS select a subset of the original features or was not able to do this, does the selected subset of features intuitively make sense to you, was the feature selection beneficial (i.e. resulted in the same or improved accuracy or other advantages)? Include anything else that you consider important.

4) Conclusions – summarize your main findings and, if possible, suggest future work.

5) Reflection – what was the most important thing you learned from this assignment? [1-2 paragraphs]