CSE471: Statistical Methods in AI -- Spring 2016
Assignment 4: NAÏVE BAYES (NB) CLASSIFIER

DUE: Before 12:00 midnight on 09 Mar 2016

INSTRUCTIONS:

1. You may do the assignment in Matlab/Octave, R, Python, C/C++ or Java.
2. You need to upload pdf files in the Course Portal. One file should contain your answers, results and analysis. A separate file should contain code you have written and its sample output.
3. At the top-right of the first page of your submission, include the assignment number, your name and roll number.
4. *IMPORTANT:* Make sure that the assignment that you submit is your own work. *Do not copy any part from any source* including your friends, seniors or the internet. Any breach of this rule could result in serious actions including an **F grade** in the course.
5. Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code.

Preamble:

This assignment requires you to implement the *Naïve Bayes (NB) classifier* and test it on two different datasets from the UCI Machine learning repository (*http://archive.ics.uci.edu/ml/*), namely the Bank Marketing Dataset (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing) and Breast Cancer Wisconsin Dataset (http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29). Given below are the instructions for carrying out the experiments. Once you conduct the experiments, calculate the metrics described and answer the questions below. The psuedocode for NB Classifier can be found on page 183 (Chapter 6.9, Machine Learning by Tom Mitchell) and many other resources on web (http://en.wikipedia.org/wiki/Naive_Bayes_classifier) (http://www.csee.wvu.edu/~timm/cs591o/old/BasicMethods.html). Although the pseudocode is for discrete valued features and the example is for text document classification, it can be modified appropriately for any classification problem with continuous feature values (using Gaussian Distribution) as well as for classification problems from other application domains.

Experimental Procedure:

1. Download the two datasets mentioned above from the UCI ML repository. Both require binary classification (*Bank Marketing Data: has the client subscribed a term deposit? ('yes','no'); Cancer Data: Malignant or Benign*). Please identify and remove any records with missing values. Please make a note of these and indicate in your report.
2. In the first dataset (Bank Marketing Dataset), work with only Discrete/Categorical Data and do not consider the continuous valued features for this assignment. So, make a subset of the Bank Marketing dataset consisting of discrete valued features only as BankMarketing_Discrete dataset and work with this.

3. For the second dataset (Cancer Dataset), all the features (attributes) are continuous. So, you need to use *Gaussian Naïve Bayes (GNB)* approach wherein you need to estimate Mean and Variance of each attribute and use Gaussian Model for estimating the Likelihoods.
4. Since multiplication of probabilities leads to very small numbers, it is advised that you use log-probabilities in the calculations to avoid numerical errors.
5. Randomly partition each dataset into two parts of equal size named *Training set* and *Testing set*.
6. Use the training set as labelled examples and do NB classification of the Testing set samples. Deal with any ties appropriately but mention this in observations part of the report.
7. Compute the accuracy (percentage of correct classification) using the labels of the Test samples and the output from your classifier.
8. Repeat steps 2 - 4 for ten times, each time dividing the dataset differently. Report the mean and standard deviation of the accuracy over the ten test runs.
9. Report the *best Naïve Bayes Classification Model* you obtained in the 10 test runs. Specification of the Model is in terms of all the estimated probabilities for each class.
10. Report the *confusion matrix* of each dataset, which depicts the classes that are most confused.

Questions to be Answered:

1. Take three example records that are misclassified in each dataset and explain why these were misclassified.
2. What is the role of the following *Laplacian smoothing* used in the pseudocode for estimating posterior probabilities?

$$P(w_k|v_j) = \frac{n_k+1}{n+|vocabulary|}$$

3. Briefly explain what modifications you would suggest in order to build an NB classifier dealing with mixed data (consisting of both continuous and discrete features) in the first dataset (Adult Dataset).
4. What procedure would you suggest for considering missing values (and not discarding them!)?