CS 481, Homework 3

Out: Mar 23, 2015, Due: Apr 01, 2015 (11:55 pm), Total: 70

Note:

- This homework will carry 7 points towards your final score
- Please answer all the questions below. Please type or write legibly.
- If a question asks you to write a code, you need to submit a working code through oncourse submission site.
- Homeworks are individual work, please do not collaborate with others inside or outside of the class.
- Please email the instructor or use the office hours for any questions.

Questions

In this assignment you will evaluate the performance of the Full Bayes and Naive Bayes classifier on the `iris.txt.shuffled` dataset (uploaded in piazza's resource section). The dataset has 150 instances. Each instance has four numeric dimensions, and in addition to that, it has a categorical dimension, which is the class label. Over the entire datasets there are three distinct classes.

Solve the following questions. You must use Python and the NumPy scientific computing package for the programming part of a question.

1. Implement a python program that accepts a dataset as a command line parameter and generates a model file in the current directory. The model file contains: (i) the prior probabilities of each of the classes; (ii) the mean and the covariance matrix of each of the classes. Our objective is to use this model file to perform classification using full Bayes classification method. To ensure readability of the model file, please write all the numeric values using 2 digits after the decimal point. You can use build-in functions in the NumPy package for computing the mean and the covariance.

2. Implement a python program that accepts a model file (output of Q1) and a test file as command line parameter. The test file has identical format of the train file. For each instance of the test file, the program outputs the predicted label. The program also prints a confusion matrix by comparing the true labels and predicted labels of all the instances.

3. Now, we will use 3-fold cross-validation for assessing the performance of the classifier. For this, make 3-folds of the file `iris.txt.shuffled` by considering 50 consecutive instances as one fold (do not reorder the instances in the files). Use the program from Q1 for training purpose using instances from two of the folds and use the program from Q2 for testing on the instances of the remaining fold. Print the confusion matrix for each of the three folds (when they were used as test). Also, for each class, print the accuracy, precision, recall, and F-score, averaged over 3-folds.

4. Solve the Question 1-3, but instead of full Bayes, use Naive Bayes

1 Deliverables

Submit 4 python scripts: one for Q1, one for Q2, and two more for the case of Naive Bayes. Submit a pdf file that contains the answers of question 3 and 4. Do not submit printout of source code in the pdf.