Theory

1. (5 points) What are the three types of learning schemes in the pattern recognition literature? Give a practical example for each one of them.
2. (10 points) section 2.1, question 1
3. (10 points) section 2.2, question 2

Programming

4. (10 points) Using the following code snippet (written in MATLAB), generate a 2-dimensional data distribution containing 100 samples.

```
1 mu = [2,3]; % mean of 2 random variables
2 sigma = [1,1.5;1.5,30]; % range of std of 2 random variables
3 rng default % For reproducibility
4 r = mvnrnd(mu,sigma,100); % randomly generated 100 points
5 figure;plot(r(:,1),r(:,2),'+')
6 % corr(r(:,1),r(:,2))
7 corr(r(:,1).*r(:,1),r(:,2).*r(:,2))
```

From the 2-dimensional data distribution, perform the following tasks:

1. Obtain the correlation of the 2 dimensions of the data.
2. Let the first dimension be X (r(:,1)) and the second dimension be Y (r(:,2)). If we fix the sigma parameter for X (line 2 in snippet) and vary sigma parameter for Y, what effect can be observed in correlation (X,Y). Plot the correlation scores by increasing and decreasing sigma parameter for Y.
3. If data is distributed diagonally with a correlation of 0.9, can we decorrelate data by applying some transformations?
4. On the original data distribution (obtained after running the code snippet), if every value of dimension X (represented by r(:,1)) and dimension Y (represented by r(:,2)) is squared, what can be said about the new correlation. [we are looking for correlation of X^2 and Y^2]
5. If we change line 1 of the snippet to mu = [-2,3], what is the correlation of X^2 and Y^2? Compare this correlation with the correlation of X and Y. Explain the difference visually using the scatter plot.

5. (20 points) Download teaching assistant evaluation data set from UCI machine learning repository (`http://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation`). There are three-classes, low, medium and high score. Implement a Bayes classifier to classify low scored TAs and high scored TAs. For this exercise, all attributes have nominal (i.e., discrete) values.

Randomly split your data set into 70% training and 30% testing instances. Now train your classifier on the training set and perform following tasks:

1. Once the training is complete, you should classify the test examples (i.e., remaining 30% instances). The output should include the number of true test examples classified correctly (TPR), the number of negative examples classified correctly (FAR), and the percent accuracy on the test set.

2. Evaluate the performance of your classifier using a confusion matrix and ROC curve on the same split of the data set.
3. Effect of train/test split: Split data randomly into 50% training, 50% testing. What happens to the training and testing set accuracy? Why?
4. Perform 5 fold cross-validation on the complete data set. Report mean and standard deviation of percent accuracy. Is there any difference in the accuracy reported in part 1 and this part of the question. Why?

Submission format: Please submit a report for all your analysis and observations only and only in the PDF format. Other format will not be evaluated. You are allowed to use only **PYTHON** or **MATLAB** for the programming assignment. All the graphs should have labels (on the axis), legends, title. You should also try to combine the graphs and plots for comparison and better representation.

Submission files: Along with the report, submit following files:

1. main.py: To read, to make partitions of the data and to call the training and testing functions.

2. train_bayes.py: To train bayes classifier on the input data set. This script has following arguments:
a) Input: trainset (training set of your split)
b) Output: model (a structure having required probabilities for both the classes, example. model.low represents the posterior probability for low scored TA class)

3. test_bayes.py: To perform bayes classification on the input data set. The script has following arguments:
a) Input: testset, model (testing set and trained model)
b) Output: TPR, FAR and percent accuracy over multiple thresholds so that you can plot ROC from the values.

4. Please submit your trained model file for 70-30 train-test split.