Assignment Two
ECE 4950

- Problems 1 will be self-graded by students. Teaching staff will grade problem 2 and 3.
- For each sub-part, assign full points if answer is correct (or you think is approximately correct), assign zero otherwise.
- Write the total, your name, and net-id on top of the solutions (that you scanned and uploaded). Hand in the graded solutions on or before **Friday, 24th March**. This can be done after the class or during office hours on any day before 24th March.
- Please include the plots from Problems 2 and 3 in your self-graded solutions.

Each part of the first problem is worth 5 points.

Problem 1 (20 points). Consider following data set:

Gender	Height(in)	Foot size
M	72	12
M	68	9
M	75	11
M	64	10.5
F	65	8
F	67	7.5
F	62	6
F	70	8.5
F	64	8

1. What are the class probabilities Pr(M) and Pr(F) according to empirical rule covered in the class?
   Solution: (1 Point each, 2 Points total) Pr(M)=4/9, and Pr(F)=5/9.

2. Assuming Gaussian naive Bayes estimate following:
   Solution: (1 Point each, 8 Points total) means: $\mu_{H|M} = 69.75$, $\mu_{S|M} = 10.625$, $\mu_{H|F} = 65.6$, $\mu_{S|F} = 7.6$,
   variances: $\sigma^2_{H|M} = 17.1875$, $\sigma^2_{S|M} = 1.172$, $\sigma^2_{H|F} = 7.44$, $\sigma^2_{S|F} = 0.74$.

Height(in)	Foot size
68	9.5

3. Suppose you were given above example:

   Using means and variances from above part compute:
   Solution: (2 Points each, 8 Points total)
   $p(H|M) = 0.0880$, $p(S|M) = 0.2148$, $p(H|F) = 0.0993$, $p(S|F) = 0.0405$.

4. Using Gaussian naive Bayes assumption, what would you classify above sample, male or female?
   Solution:(2 Points) Since

   $$0.084 = Pr(M)p(H|M)p(S|M) > Pr(F)p(H|F)p(S|F) = 0.0022,$$

   above sample will be classified as male.

Problem 2 (20 points). Recall the WDBC dataset, and the data-sets we used in the last assignment. In particular, we had two datasets, the first consisted of `X-trn-200.csv`,`Y-trn-200.csv`, `X-tst-200.csv`, and `Y-tst-200.csv`. The second dataset contained four files ending with `-400.csv`. This time, in addition, we are providing four files that end with `-50.csv` that has 50 training examples and 519 test examples obtained as before.

1. Write a code that implements Gaussian Naive Bayes over the new dataset, and the two previous datasets (three datasets in total `-50.csv, -200.csv, -400.csv`).

2. What is the accuracy of the the algorithm over the *three* test sets?

   Solution: The accuracy over the three datasets are 0.10982, 0.03523, 0.03550 respectively.

3. Use the code you wrote for decision tree in the first assignment on the dataset ending with `-50.csv`, again plotting the test and train accuracy as a function of the maximum depth. Which one has a better test error: Naive Bayes or the decision tree with the best depth?

   Solution: The error for NB is about 10% and decision trees around 25%. This is perhaps because Naive Bayes takes into account all the features, unlike decision tree which only considers a few features, and with 50 training examples, underfits.

4. Instead of using maximum depth to prevent overfitting, modify the code to obtain decision trees with minimum leaf size i. For $i = 2, \ldots, 15$, plot the test and training accuracy. Does this achieve a higher accuracy compared to the maximum depth criterion? Is it better than Naive Bayes?

   Solution: The minimum samples criterion provides an improvement over maximum depth, which is prominent on the set with only 50 examples. In general maximum leaf size is a better criterion (think why).

Please submit all the codes, and plots.

Problem 3 (20 points). In this problem we will implement the Multinomial Naive Bayes (add-β) algorithm for the MNIST dataset. Each image is a 28x28 gray-scale image. Therefore each image is a 784 dimensional feature vector. There are 60,000 training examples, and 10,000 test examples. The training data is `mnist_train.csv`, that contains 785 entries in each row, where the first entry is the digit label, and the remaining 784 are the values of the 784 pixels. The test data is `mnist_test.csv` which has the same structure.

1. Implement a Multinomial Naive Bayes for smoothness parameter β= 0.0000001, 0.0001, 0.1, 1, 1000, 1000000, 1000000000, 1000000000000. Plot the error rate as a function of β.

2. What does the classifier do as $\beta \to \infty$? Please explain the error rate obtained for $\beta \to \infty$.

   Solution: As $\beta \to \infty$, the features are completely smoothed out. Comparing $P(C)P(X|C)$ reduces to comparing $P(C)$. The NB classifier simply assigns each test to the class with the highest prior probability.

   As $\beta \to \infty$, the error rate obtained is 0.8865. In the training examples, the digit appearing the highest number of times is 1. Hence all test examples are labeled 1. The number of 1's in the test set is 1135. Therefore, the error rate is 1-1135/10000 = 0.8865.

Notes regarding the programming assignment:

- It is perhaps easiest to implement these problems using Numpy, and Scipy, and in particular `sci-kit-learn`. Look for naive Bayes implementations in sci-kit-learn.