**CS 4375 Fall 2019**

**Introduction to Machine Learning**

**Assignment #4**

**Due Date: December 2nd, 2019 (by 11:59 p.m.)**

**Section A**

**Problem Statement:**

Imagine that you are given the following set of training examples. Each feature can take on one of three nominal values: a, b, or c.

Test example: F1 = a, F2 = c, F3 = b

|  |  |  |  |
| --- | --- | --- | --- |
| F1 | F2 | F3 | Category |
| a | c | a | + |
| c | a | c | + |
| a | a | c | - |
| b | c | a | - |
| c | c | b | - |

1. Using the Bayes rules, which class has highest posterior probability given the test example?
2. Plot and show how a 3-nearest-neighbor algorithm would classify the test example given above.

Hint: Since the data is nominal, we will use hamming distance as the distance metric. Hamming distance is the number of attributes on which the two examples agree on.

**Section B**

Please solve the following problem by coding Python programs:

***Data:***

You will be working on MNIST data. To reduce time of running the program, you will only work with 1,000 randomly selected samples out of these, although make sure you have equal number of samples belonging to each label (i.e.,100 samples of label ‘0’, 100 samples of label ‘1’ and so on).

***Problem Statement:***

1. Subsample the data and divide them to training and testing set.
2. Use PyTorch from scratch to build CNN model, and show your testing result.
3. Perform Adaboost, Bagging on the subset.
4. Compare the precision and recall of Adaboost and Bagging.

***Task 1:***

You can implement the method from sklearn to implement classification. You will need to sample 100

instances out of each label (1,000 instances in total). Randomly divide the dataset with 80-20 ratio as

training and testing set.

***Task 2:***

Apply PyTorch package to implement the CNN algorithm, and train the model using the training set

generated in the last task and use your model to predict labels in the test set. The structure of CNN is defined as follows.

Structure:

Convolutional layer -> Max pooling layer -> Convolutional layer - > Max pooling layer -> Fully

connected layer x2 -> Softmax layer

***Task 3:***

You are going to perform Adaboost, Bagging and using two base classifiers respectively, including decision tree and logistic regression.

Adaboost:

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>

Bagging:

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html>

***Task 4:***

Compare the precision and recall of different base classifier.

To calculate the precision and recall for multi-label classes, you can refer those links:

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html>

***Deliverables:***

1. In this submission, please include your project report and source code.
2. There is no minimum length requirement, the maximum length is 5 pages.