ECE 6524 – Deep Learning
No. HW-02-6524

Homework Assignment #2: Naïve Bayes Classifiers

Part A

Build a Naïve Bayes classifier using the data shown in the following table to predict whether a text file will be saved (class label y = +1) or discarded (y = -1). The features (x(i), i = 1, 2, …, 5) are binary values (representing some attributes of a text file, e.g., "known author or not", "long or short", etc.).

features					class
x(1)	x(2)	x(3)	x(4)	x(5)	y
0	0	1	1	0	-1
1	1	0	1	0	-1
0	1	1	1	1	-1
1	1	1	1	0	-1
0	1	0	0	0	-1
1	0	1	1	1	+1
0	0	1	0	0	+1
1	0	0	0	0	+1
1	0	1	1	0	+1
1	1	1	1	1	-1

The classifier computes the probability in the following steps:

- Step 1: Compute the prior probability for each class (p(y)), y = -1 or +1;
- Step 2: Compute the likelihood probability with each class for each feature (p(x(i) | y));
- Step 3: Calculate the posterior probability for each class given each feature (p(y | x(i));
- Step 4: Predict a class for a given text file based on the posterior probability. Specifically, what is the posterior probability that *y* = +1 given the feature vector x = (1 1 0 1 0). Which class would be predicted for x = (0 0 0 0 0)? What about for x = (1 1 0 1 0)?

For Part A, you need to give the detailed, intermediate results for each step listed in the assignment.

Part B

1) **Describe** the Gaussian Naïve Bayes algorithm in detail.

2) Using *GaussianNB*() as implemented in the *scikit-learn* library to classify the samples in the breast cancer Wisconsin dataset.

You can know more about the dataset at https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html. A brief summary of the dataset is listed in Table 1:

Table 1: A summary of the breast cancer dataset

Classes	2
Samples per class	212(M),357(B)
Samples total	569
Dimensionality	30
Features	real, positive

(More details of the dataset can be found at UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).)

(a) **Implement** a simple program using *GaussianNB*() to classify the data samples (including train and test data samples). **Report the prediction accuracy** and **confusion matrix** of the classifier.

(b) Add (30-dimensional) zero-mean Gaussian noises with different variances (at least 5, e.g., 50, 100, 200, 400, 800) to the features of the dataset; **train** Gaussian Naïve Bayes classifiers to classify the noisy datasets and **report the prediction accuracies** and **confusion matrices** of the classifiers.

3) Discuss the advantages and disadvantages of the Naïve Bayes classifiers. Additionally, discuss possible ways to overcome the "**zero probability" (or "zero frequency")** problem.

For Part B, you need to prepare **a written report** (in the pdf format) including the following sections: (1) Approach(es), (2) Experimental Results, and (3) Discussion. In addition, you need to attach your **implementation codes** as separate files to the report.