EE6407 Assignment 1

Question:

The training data is given in TrainingData.xlsx, where columns A-D are the features, and column E gives the class label. The test data is given in TestData.xlsx (30 samples, no label is provided).

• Q1: Are there any missing values and outliers in the training data? If there are, describe two methods that can address each of the two problems.

• A1:

Yes, there are both missing values and outliers in the training data. Specifically, missing values were identified in Feature1 and Feature4, while an outlier was found in Feature3 with a value far exceeding the typical range. One approach to address missing values is removing any samples that contain them, especially if the number of such samples is relatively small. Alternatively, missing values can be imputed using statistical techniques, such as replacing them with the mean or median of the corresponding feature. As for outliers, a common method to handle them is by using the Interquartile Range (IQR) to detect and remove values that fall significantly outside the normal range. Another option is to apply data transformations or cap the extreme values to reduce their influence on the model. Both strategies help ensure that the training data is cleaner and more suitable for building reliable classifiers.

• Q2: Assume the samples with missing values and outliers are simply removed from the training data. Design a Naïve Bayes classifier using this modified training data, show the parameters and the decision rules.

• A2:

After removing the samples with missing values and outliers from the training data, a Gaussian Naïve Bayes classifier was constructed using the remaining clean data. The classifier assumes that each feature follows a normal distribution within each class. For each class, the model calculates the mean and variance of each feature, as well as the prior probability of the class, based on its frequency in the training set. These parameters are then used to compute the likelihood of a new sample belonging to each class. The decision rule is based on selecting the class with the highest posterior probability, which is proportional to the product of the prior and the likelihood of the features. Mathematically, for a test sample, the model predicts the class c that maximizes $P(c)\prod_{i=1}^{4}\mathcal{N}(x_i \mid \mu_{c,i}, \sigma_{c,i}^2)$, where

$\mu_{c,i}$ and $\sigma^2_{c,i}$ are the mean and variance of feature i for class c. This results in a simple yet effective classifier that relies on feature independence assumption within each class.

Q3: Predict the class label of the test data.

A3:

Using the trained Gaussian Naïve Bayes classifier based on the cleaned training data, the class labels of the test dataset were predicted. Each of the 30 test samples was evaluated by computing the posterior probability for each class, and the class with the highest probability was assigned as the predicted label. As a result, each test sample received a corresponding class label based on the learned statistical patterns from the training data. The full list of predicted labels has been presented in the table below.

Sample Index	Predicted Label
0	1
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1

16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1
29	1