EE6407: Assignment Report

Name: Zhang Zezheng
Matriculation number: G2402715A

1 Assignment 1

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).

(1) 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.

(2) 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.

(3) Predict the class label of the test data.

1.1 Question (1): Address Missing and Outliers

We examined the training dataset and found missing values or outliers.

- **Missing Values:**We found that columns A and D contain entries marked with ?, which are treated as missing values. As shown in Table 1.
- **Outliers:**We also checked for outliers in the cleaned dataset using the Z-score method, where values with $|Z| > 3$ are considered outliers. Several outliers were identified in the numerical feature columns. As shown in Table 2.

Table 1: Removed Rows with Missing Values

A	B	C	D	Label	Original_Excel_Row
?	3.3	5.7	2.5	2	40
?	3.6	1.0	0.2	1	62
5.9	3.0	4.2	?	3	84

Table 2: Removed Rows with Outliers (Z-score > 3)

A	B	C	D	Label	Original_Excel_Row
6.3	3.3	16.0	2.5	2	95
10.6	3.2	1.4	0.2	1	100
5.7	4.4	1.5	0.4	1	102
5.7	2.9	4.2	5.3	3	121

Approaches to address Missing Values

1. **Row Removal:** Eliminate any sample (row) that contains at least one missing value.
2. **Value Imputation:** Estimate and fill missing values using strategies such as the column's mean, median, or via K-Nearest Neighbors (KNN) imputation.

Approaches to address Outliers

1. **Z-score Technique:** Identify data points as outliers if their standardized score (Z-score) is greater than 3 or less than -3.
2. **IQR Technique:** Detect outliers by calculating the interquartile range (IQR) and flagging values falling outside the interval $[Q1 - 1.5IQR, Q3 + 1.5IQR]$.

1.2 Question (2): Classify

After removing the rows with missing values and outliers, a Naïve Bayes classifier is designed using these modified training data and shows the parameters and the decision rules.

1.2.1 Methodology: Naïve Bayes Classifier

The Naive Bayes classifier relies on the assumption that features are independent from each other, which is different from the Bayes Decision Rule. The classifier computes the class conditional probability density function for each feature independently. These probabilities are multiplied together. The class associated with the highest posterior probability is selected. For a Gaussian Naïve Bayes, the class conditional probability density function is shown below.

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right)$$

In this equation, μ_k and σ_k^2 are the mean and variance of feature i for class C_k.

1.2.2 Bayes Decision Rule Classifier Parameters

For class 1, the mean vector μ_1 and the covariance matrix Σ_1 are given below.

$$\mu_1 = [4.9865,\ 3.3811,\ 1.4784,\ 0.2514]$$

$$\Sigma_1 = \begin{bmatrix} 0.0968 & 0.0758 & 0.0183 & 0.0096 \\ 0.0758 & 0.1232 & 0.0218 & 0.0074 \\ 0.0183 & 0.0218 & 0.0301 & 0.0067 \\ 0.0096 & 0.0074 & 0.0067 & 0.0131 \end{bmatrix}$$

For class 2, the mean vector μ_2 and the covariance matrix Σ_2 are given below.

$$\mu_2 = [6.5541,\ 2.9459,\ 5.5514,\ 2.0108]$$

$$\Sigma_2 = \begin{bmatrix} 0.4414 & 0.0874 & 0.3455 & 0.0502 \\ 0.0874 & 0.0842 & 0.0637 & 0.0362 \\ 0.3455 & 0.0637 & 0.3437 & 0.0430 \\ 0.0502 & 0.0362 & 0.0430 & 0.0715 \end{bmatrix}$$

For class 3, the mean vector μ_3 and the covariance matrix Σ_3 are given below.

$$\mu_3 = [5.9615,\ 2.7821,\ 4.3077,\ 1.3436]$$

$$\Sigma_3 = \begin{bmatrix} 0.2535 & 0.0709 & 0.1790 & 0.0541 \\ 0.0709 & 0.0968 & 0.0780 & 0.0390 \\ 0.1790 & 0.0780 & 0.2328 & 0.0752 \\ 0.0541 & 0.0390 & 0.0752 & 0.0383 \end{bmatrix}$$

1.2.3 Naïve Bayes Classifier Parameters

For class 1, the mean vector μ_1 and the variance vector σ_1^2 are given below.

$$\mu_1 = [4.9865,\ 3.3811,\ 1.4784,\ 0.2514]$$

$$\sigma_1^2 = [0.0941,\ 0.1199,\ 0.0293,\ 0.0128]$$

For class 2, the mean vector μ_2 and the variance vector σ_2^2 are given below.

$$\mu_2 = [6.5541,\ 2.9459,\ 5.5514,\ 2.0108]$$

$$\sigma_2^2 = [0.4295,\ 0.0819,\ 0.3344,\ 0.0696]$$

For class 3, the mean vector μ_3 and the variance vector σ_3^2 are given below.

$$\mu_3 = [5.9615,\ 2.7821,\ 4.3077,\ 1.3436]$$

$$\sigma_3^2 = [0.2470,\ 0.0943,\ 0.2269,\ 0.0373]$$

1.3 Question (3): Predict

We applied the trained classifier to the 30 samples in TestData.xlsx.

Table 3: Naïve Bayes Prediction on Test Samples

Sample #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Prediction	1	1	1	1	1	1	1	1	1	1	3	2	2	2	2
Sample #	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Prediction	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3

2 Assignment 2

Given a 2-class pattern classification problem, where the training data is given in *data_train*, and the class label is given in the *label_train*. The test data is given in *data_test* (23 samples).

(1) Train a Fisher linear discriminant classifier: detail the learning process, and give the values of weight vector, bias term and the decision rule.

(2) Use the trained classifier in part (1) to predict the class label of the test data: give the class labels of all the testing samples.

2.1 Question (1)

2.2 Fisher Linear Discriminant: Step-by-Step Implementation

1. **Compute class mean vectors**

$$\mu_+ = \frac{1}{N_+}\sum_{x\in X_+} x, \qquad \mu_- = \frac{1}{N_-}\sum_{x\in X_-} x$$

2. **Compute within-class scatter matrix**

$$S_+ = \sum_{x\in X_+}(x-\mu_+)(x-\mu_+)^T$$
$$S_- = \sum_{x\in X_-}(x-\mu_-)(x-\mu_-)^T$$
$$S_W = S_+ + S_-$$

3. **Compute optimal projection vector w**

$$\mathbf{w} = S_W^{-1}(\mu_+ - \mu_-)$$

4. **Compute the bias term** w_0

   The decision boundary is placed halfway between the two class means:

$$w_0 = -\frac{1}{2}\mathbf{w}^T(\mu_+ + \mu_-)$$

Decision Rule

Define the decision rule:

$$f(x) = \mathbf{w}^T x + w_0$$

Fisher Linear Discriminant Training Results

The weight vector **w** and bias term w_0 are computed as:

$$\mathbf{w} = [-0.00178602,\ -0.0034729,\ -0.00194673,\ -0.00168257,\ -0.00265808]$$

$$w_0 = 0.0087$$

Classification:

- If $y \geq 0$, predict class $+1$
- If $y < 0$, predict class -1

2.3 Question (2): Predict

Using the trained classifier, we compute prediction scores for each of the 23 test samples.

Table 4: Predicted Class Labels for 23 Test Samples

Sample #	1	2	3	4	5	6	7	8	9	10	11	12
Prediction	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
Sample #	13	14	15	16	17	18	19	20	21	22	23	
Prediction	1	1	1	1	1	1	1	1	1	1	1	