Assignment 1:

(1) Parameters for Classifiers:

The entire Covariances matrices are used for Bayes decision rule, and only the diagonals of the matrices are used for Naïve Bayes.

Class -1:

Means:

1.05170834	1.79957542	0.85956476	1.06420754	1.00855478

Covariances:

1.04775072	9.74443630e-02	4.00084748e-02	-4.42894806e-02	6.51321191e-02
9.74443630e-02	8.39017214e-01	-9.01956614e-02	9.91459313e-04	-7.83496197e-02
-4.00084748e-02	-9.01956614e-02	1.09720991	5.13726225e-02	1.12383358e-02
-4.42894806e-02	9.91459313e-04	5.13726225e-02	1.03777085	4.58946940e-02
6.51321191e-02	-7.83496197e-02	1.12383358e-02	4.58946940e-02	1.02044233

Class 1:

Means:

0.01379514	-0.08755727	0.08461257	-0.15482884	0.02275036

Covariances:

1.12042246	0.05647016	-0.00456629	0.0285	0.05122562
0.05647016	0.93843837	0.02756376	-0.04174351	0.07543512
-0.00456629	0.02756376	0.85220569	-0.09399161	0.04922219
0.0285	-0.04174351	-0.09399161	0.9464923	0.07759107
0.05122562	0.07543512	0.04922219	0.07759107	1.16957147

(2) Prediction Results:

Two classifiers give the same prediction result:

-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
(15* -1, 11* 1)

(3) Discussion:

Both classifiers are based on the Bayes Theorem:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

However, due to different assumptions, Classifiers calculate P(x|c) (or called **posterior probability or likelihood**) in different ways.

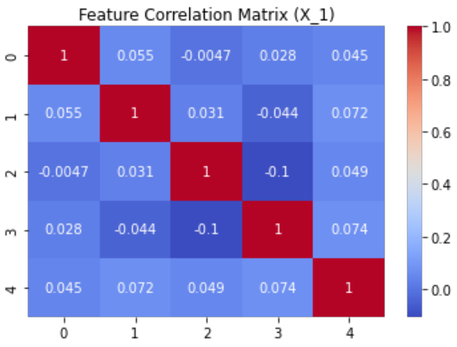
Bayes Decision Rule (BDR) does not assume independence between features. Consequently, the entire covariance matrix is used to capture the relationships between the features. Typically, a Multivariate Gaussian distribution is employed to model the distribution of the features for each specific class, which allows for the proper estimation of the likelihood (P(x|c)).

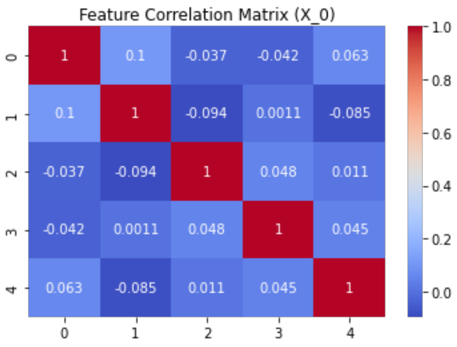
The Naïve Bayes classifier assumes that features are independent of each other. As a result, only the mean and variance of each individual feature are used to compute the likelihood. The overall likelihood is calculated as:

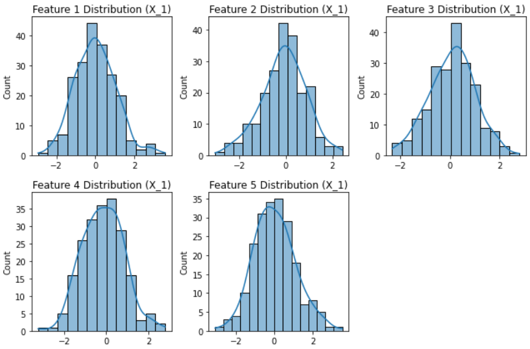
$$P(x|c) = \prod_{i=1}^{n} P(x_i|c) = P(x_1|c) \cdot P(x_2|c) \cdot \ldots \cdot P(x_n|c)$$

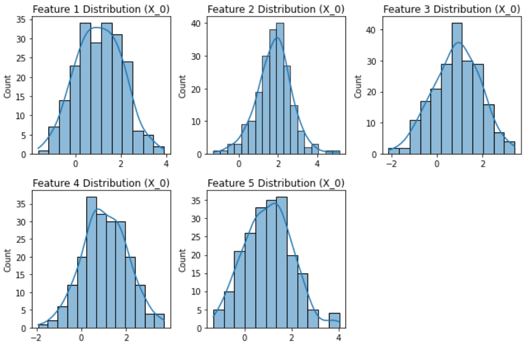
Typically, the likelihood of each feature $P(x_i|c)$ is assumed to follow a Gaussian distribution.

Upon observing the data, we noticed that there is no significant correlation between the features. The features within each class approximately follow a normal distribution. As shown below (X_1 for class 1, X_0 for class -1):









This observation suggests that the data fits well with Naïve Bayes' independence assumption. Therefore, we can reasonably expect Naïve Bayes to perform well on this dataset, and its classification accuracy should be high.

However, experimental results show that both Naive Bayes and BDR produced **identical predictions** on this dataset. This can be explained by the characteristics of the data.

Although BDR employs a **multivariate Gaussian distribution** to model $P(x|c)$, its effectiveness is contingent on the **covariance matrix** Σ_c. In this dataset, we observed that the diagonal elements of the covariance matrix (the variances of the features) are close to 1, while the off-diagonal elements (the covariances between features) are close to 0. This indicates that the features are almost independent of each other.

In this scenario, the multivariate Gaussian distribution in BDR approximates the product of independent univariate Gaussian distributions. Therefore, the likelihood $P(x|c)$ in BDR is effectively the same as the likelihood calculated by Naïve Bayes, where each feature is treated independently. As a result, both classifiers produce identical predictions.

Assignment 2:

(1) Handling Missing Values and Outliers

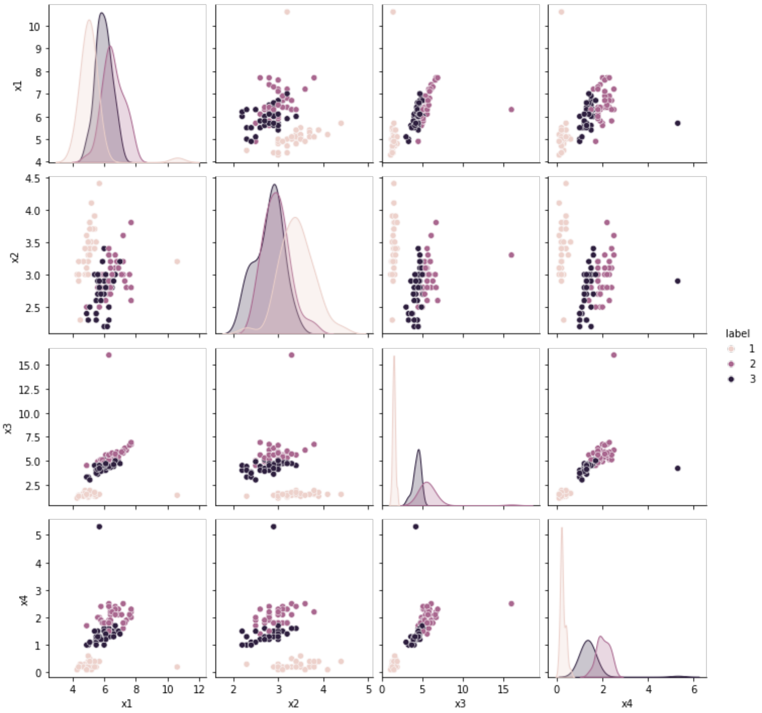
Upon analyzing the training data, no empty cells were found, but a few non-numerical missing values were detected. Specifically, one missing value was found in each of the three classes, as shown below:

?	3.3	5.7	2.5	2
?	3.6	1.0	0.2	1
5.9	3.0	4.2	?	3

There are 120 samples in total, and the class distribution is relatively balanced (41, 40, and 39 samples per class). Since removing the rows with missing values does not significantly affect the overall data balance, these rows were deleted, reducing the dataset to 117 samples (40, 39, and 38 per class).

To identify outliers in the dataset, a descriptive analysis was first conducted. The sns.pairplot library was used to visualize the distribution of samples for each class across different feature dimensions. The results of the analysis are as follows:

1. Feature x2 shows significant overlap between the samples from class 2 and class 3, making it difficult to distinguish between the two classes in this dimension.
2. In other feature dimensions, there is a clear distinction between class samples.
3. Most combinations of feature dimensions show that the three classes of samples are linearly separable.



Due to the presence of a few outliers, some overlap exists between class clusters in the feature space. IQR (Interquartile Range), Z-score, and boxplot methods were used to analyze and detect the outliers. After a comprehensive analysis of the entire dataset, the results from the three methods identified five outliers. The values marked in red below represent the outliers:

10.6	3.2	1.4	0.2	1
5.2	4.1	1.5	0.1	1
5.7	4.4	1.5	0.4	1
6.3	3.3	16.0	2.5	2
5.7	2.9	4.2	5.3	3

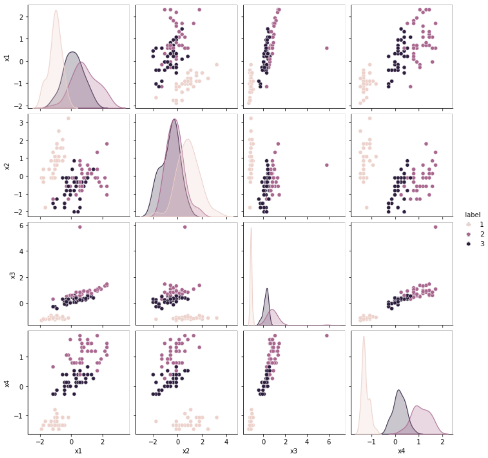
SVM aims to find a hyperplane that maximizes the margin between different classes in feature space, making it particularly sensitive to noise near the margin. However, not all detected outliers are located near the margin or contribute to overlap between class clusters. Therefore, we conducted a more detailed outlier analysis for each class separately, supplemented by visual inspection, to specifically identify outliers that may impact SVM's classification performance. This process allows us to focus only on outliers that might affect the model’s decision boundary.

Aside from the overlap between class 2 and class 3 in the x2 feature space, only two outliers in other feature spaces cause overlap between class distributions:

10.6	3.2	1.4	0.2	1
5.7	2.9	4.2	5.3	3

Removing these two samples does not significantly affect the dataset size. Therefore, they were removed, and the dataset was reduced to 115 samples (39, 39, 37).

Next, the data was standardized. The visualization results showed that there was no overlap between different clusters in other feature spaces due to outliers, except in the x2 feature space.



(2) Training SVM

The data was split into 80% for training and 20% for validation. Due to the overlapping distributions in the x2 feature space, multiple values of the regularization parameter C (less than 1) were tested to find the optimal soft margin for classification. Since the three classes of samples were linearly separable in most feature dimensions, a linear kernel was chosen for the SVM classifier.

The best accuracy of 95.7% was achieved with C=0.9.

(3) Prediction

The trained SVM model was used to predict the class labels of the test data (30 samples, no labels provided). The predicted labels were as follows:

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3

The final predicted distribution was: 10 samples for class 1, 11 samples for class 2, and 9 samples for class 3.