Assignment 5
DSL253 - Statistical Programming

Amay Dixit - 12340220

Submitted to Dr. Anil Kumar Sao

Links

- Notebook Link:
  https://colab.research.google.com/drive/1Dt-RE_
  pgcAek6BIsJN8iltmUtlyTknIa?usp=sharing

- Github Link:
  https://github.com/amaydixit11/Academics/tree/main/DSL253/
  assignment_5

1 Question 1: Bivariate Normal Distribution Probabilities

1.1 Introduction

This report addresses the calculation and visualization of various probabilities within a bivariate normal distribution. The bivariate normal distribution is a key statistical concept that extends the normal distribution to two dimensions, allowing for the modeling of correlated random variables.

1.2 Data

We are given a bivariate normal distribution for random variables X and Y with the following parameters:

$$\mu_X = 3$$
$$\mu_Y = 1$$
$$\sigma_X^2 = 16$$
$$\sigma_Y^2 = 25$$
$$\rho_{XY} = \frac{3}{5}$$

1.3 Methodology

To calculate the requested probabilities, we use properties of the bivariate normal distribution:

1. For marginal probabilities: We use the fact that the marginal distributions of X and Y are univariate normal distributions.

2. For conditional probabilities: We use the fact that the conditional distribution of one variable given the other is also normally distributed with adjusted parameters.

Specifically, for a bivariate normal distribution:

- The marginal distribution of X is $N(\mu_X, \sigma_X^2)$

- The marginal distribution of Y is $N(\mu_Y, \sigma_Y^2)$

- The conditional distribution of Y given $X = x$ is $N(\mu_{Y|X}, \sigma_{Y|X}^2)$ where:

$$\mu_{Y|X} = \mu_Y + \rho_{XY} \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$
$$\sigma_{Y|X}^2 = \sigma_Y^2(1 - \rho_{XY}^2)$$

- Similarly, the conditional distribution of X given $Y = y$ is $N(\mu_{X|Y}, \sigma_{X|Y}^2)$ where:

$$\mu_{X|Y} = \mu_X + \rho_{XY} \frac{\sigma_X}{\sigma_Y}(y - \mu_Y)$$
$$\sigma_{X|Y}^2 = \sigma_X^2(1 - \rho_{XY}^2)$$

1.4 Results and Discussion

Using the given parameters, the calculated probabilities are:

(a) $P(3 < Y < 8) = 0.263822$

This represents the marginal probability that the random variable Y falls between 3 and 8. The calculation uses the univariate normal CDF with $\mu_Y = 1$ and $\sigma_Y = 5$.

(b) $P(3 < Y < 8 | X = 7) = 0.440051$

This is the conditional probability that Y falls between 3 and 8 given that $X = 7$. When we condition on $X = 7$, the distribution of Y shifts toward higher values since X is above its mean and the correlation is positive, resulting in a higher probability than part (a).

(c) $P(-3 < X < 3) = 0.433193$

This represents the marginal probability that the random variable X falls between -3 and 3. The calculation uses the univariate normal CDF with $\mu_X = 3$ and $\sigma_X = 4$.

(d) $P(-3 < X < 3 | Y = -4) = 0.643078$

This is the conditional probability that X falls between -3 and 3 given that $Y = -4$. When we condition on $Y = -4$, which is below its mean, the distribution of X shifts toward lower values due to the positive correlation. This makes the region centered at $X = 0$ more likely, increasing the probability compared to part (c).
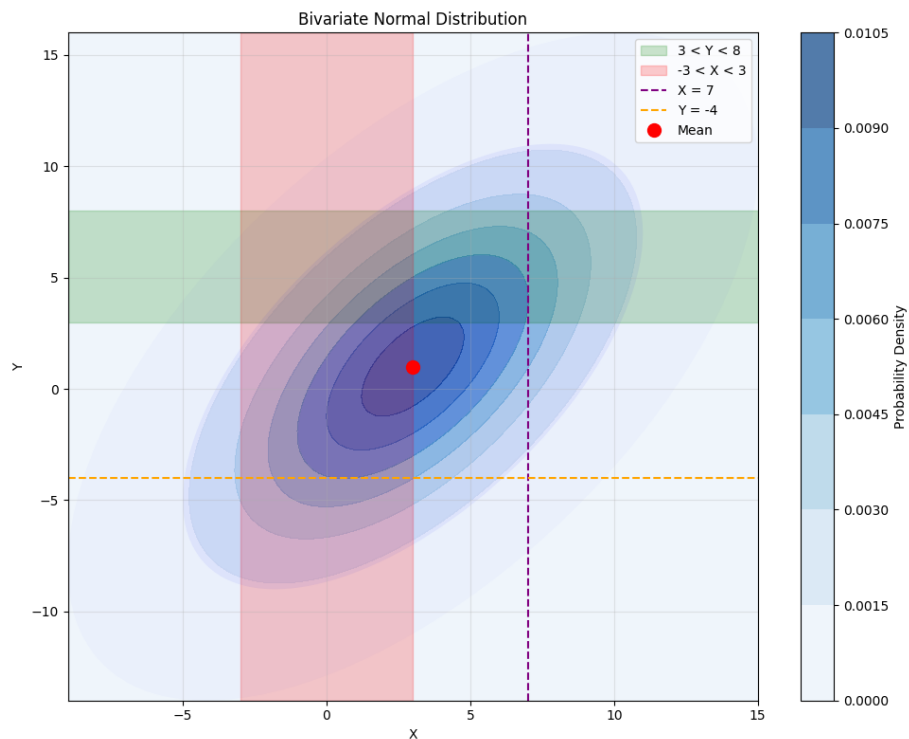
1.5   Visualization

Figure 1: Question 1

1.6   Conclusion

The calculation of probabilities in a bivariate normal distribution requires understanding both marginal and conditional distributions. The results show how conditioning on one variable affects the distribution of the other, particularly in the presence of correlation.

The implementation provided is flexible and can accommodate any set of parameters, making it useful for analyzing various bivariate normal scenarios. The visualization technique offers an intuitive understanding of how these probabilities relate to areas under the bivariate normal density function.

2 Question 2

2.1 Introduction

This section investigates the statistical properties of Mahalanobis distance and its relationship to the chi-square distribution. The Mahalanobis distance measures how many standard deviations away a point is from the mean of a distribution, taking into account the covariance structure. For multivariate normal data, the squared Mahalanobis distance follows a chi-square distribution.

2.2 Data

For this, we generate samples from multivariate normal distributions with varying dimensions (n) and sample sizes (P):

- Dimensions $n \in \{2, 5, 10\}$

- Sample sizes $P \in \{1000, 10000\}$

- Mean vector $\mu = \mathbf{0}$ (vector of zeros)

- Covariance matrix $\Sigma = \mathbf{I}$ (identity matrix)

2.3 Methodology

Our approach consists of three main steps:

(a) Generate P samples from a multivariate normal distribution $\mathcal{N}_n(\mu, \Sigma)$.

(b) For each sample X, compute the squared Mahalanobis distance:

$$Y = (X - \mu)^T \Sigma^{-1} (X - \mu) \tag{1}$$

(c) Calculate the probability $\text{Prob}[(X - \mu)^T \Sigma^{-1}(X - \mu) \leq c^2]$ for different values of c.

According to statistical theory, for $X \sim \mathcal{N}_n(\mu, \Sigma)$, the squared Mahalanobis distance follows a chi-square distribution with n degrees of freedom:

$$(X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi_n^2 \tag{2}$$

2.4   Results and Discussion

2.4.1   Distribution of Squared Mahalanobis Distance

The histograms in Figure 1 show the distribution of the squared Mahalanobis distance Y for different dimensions and sample sizes. For each combination of n and P, the observed distribution (blue histogram) closely follows the theoretical chi-square distribution with n degrees of freedom (red line).

Key observations:

- As the dimension n increases, the distribution shifts to the right, with both the mode and mean increasing.

- The shape of the distribution changes with dimension:

  - For $n = 2$, the distribution is highly right-skewed with a peak near zero.
  - For $n = 5$, the distribution becomes more symmetric but still maintains some right skewness.
  - For $n = 10$, the distribution becomes more symmetric and approaches a normal-like shape.

- Increasing the sample size from $P = 1000$ to $P = 10000$ provides a smoother histogram that more closely follows the theoretical distribution, but doesn't change the fundamental shape.

These observations confirm that the squared Mahalanobis distance follows a chi-square distribution with degrees of freedom equal to the dimension of the data.
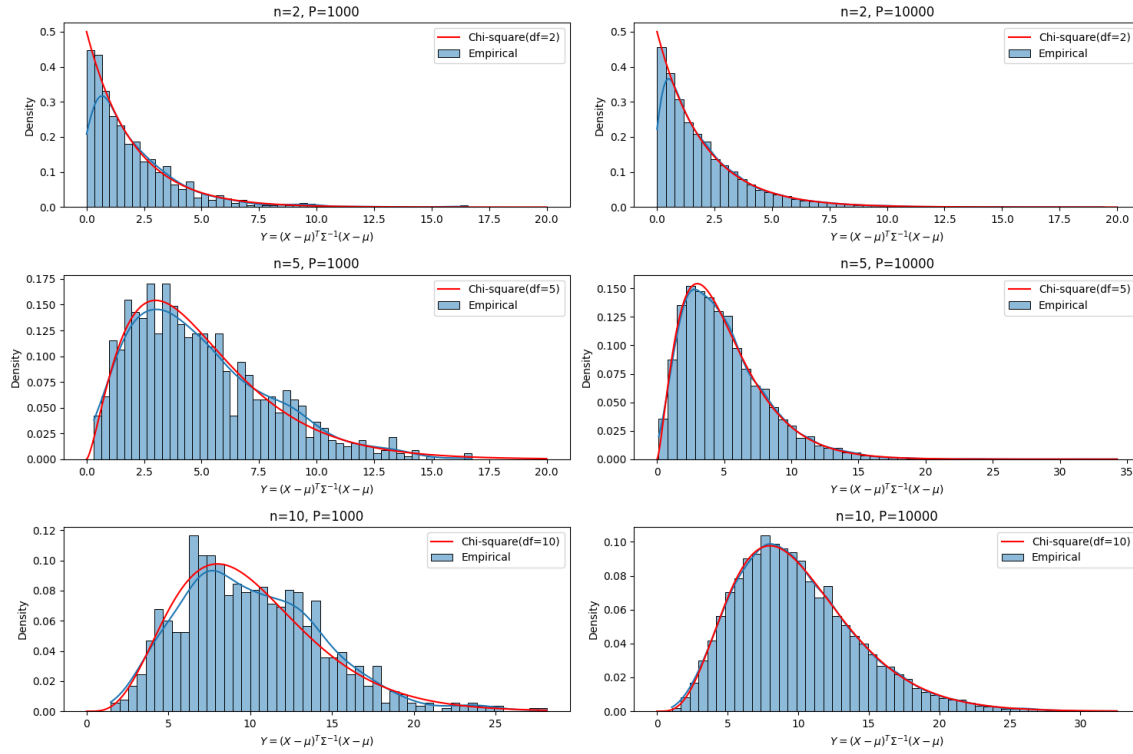
Figure 2: Distribution

2.4.2 Probability Analysis

Figure 2 shows the cumulative distribution function (CDF) for the squared Mahalanobis distance, representing the probability $\text{Prob}[(X - \mu)^T \Sigma^{-1}(X - \mu) \leq c^2]$ for different values of c and dimensions n.

Key observations:

- For a fixed value of c, the probability decreases as dimension n increases.

- For example, at $c = 2$:

    - For $n = 2$, approximately 86% of points fall within a Mahalanobis distance of 2.

    - For $n = 5$, approximately 45% of points fall within a Mahalanobis distance of 2.

    - For $n = 10$, only about 8% of points fall within a Mahalanobis distance of 2.

- The specific calculation shows that $\text{Prob}[(X - \mu)^T \Sigma^{-1}(X - \mu) \leq 2^2]$ for $n = 5$ is approximately 0.450584.

This has important implications for multivariate outlier detection and confidence region estimation: as the dimension increases, a larger Mahalanobis distance threshold is needed to capture the same proportion of the data.
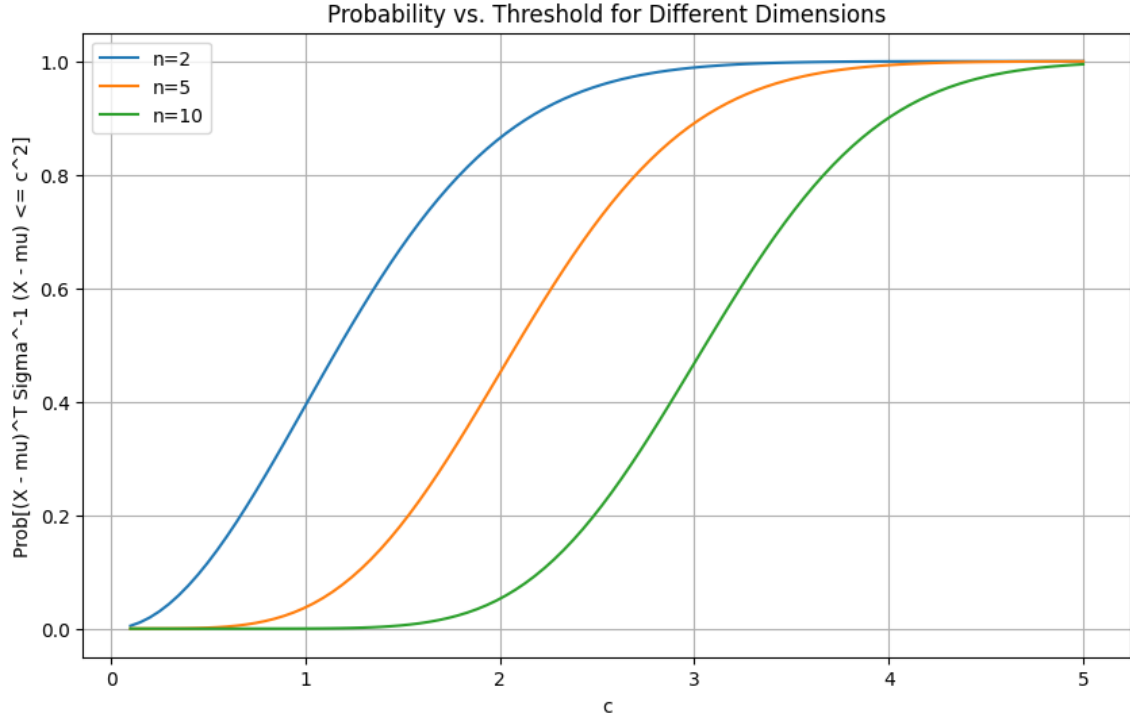

Figure 3: Probability vs. Threshold

2.5 Conclusion

The experimental results confirm that for multivariate normal data, the squared Mahalanobis distance follows a chi-square distribution with degrees of freedom equal to the dimension of the data. This relationship holds regardless of sample size, though larger samples provide better approximations to the theoretical distribution.

The probability analysis demonstrates that higher-dimensional spaces require larger distance thresholds to capture the same proportion of data points. This illustrates the curse of dimensionality: as the number of dimensions increases, data

becomes more sparse and points tend to be farther from the center in terms of Mahalanobis distance.

These findings have important applications in:

- Multivariate outlier detection

- Constructing confidence regions for multivariate normal distributions

- Statistical process control for multivariate processes

- Classification and pattern recognition

3 Question 3: Bayes Classification with Multivariate Normal Distributions

3.1 Introduction

This section explores Bayesian classification for data drawn from two different classes, each following a multivariate normal distribution with distinct parameters. Bayes' theorem provides a principled approach to classification by calculating posterior probabilities based on class-conditional densities and prior probabilities.

3.2 Problem Formulation

We are given two classes with probability distributions that follow multivariate normal distributions with the following parameters:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \qquad\qquad \boldsymbol{\mu}_2 = \begin{bmatrix} -2 \\ -3 \end{bmatrix}$$
$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}, \qquad\qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & -0.3 \\ -0.3 & 1 \end{bmatrix}$$

The goal is to classify a set of data points using Bayes' theorem and illustrate the decision boundary between the two classes.

9

3.3 Theoretical Background

For a classification problem with K classes, Bayes' theorem gives the posterior probability of class ω_i given the observation $\mathbf{x}$ as:

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{K} p(\mathbf{x}|\omega_j)P(\omega_j)} \tag{3}$$

where:

- $P(\omega_i|\mathbf{x})$ is the posterior probability of class ω_i given observation $\mathbf{x}$

- $p(\mathbf{x}|\omega_i)$ is the class-conditional probability density function for class ω_i

- $P(\omega_i)$ is the prior probability of class ω_i

- The denominator is the evidence, which serves as a normalization constant

For multivariate normal distributions, the class-conditional density for class ω_i is given by:

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right) \tag{4}$$

where d is the dimensionality of the feature space (in our case, $d = 2$).

3.4 Decision Boundary Derivation

The Bayes decision rule assigns an observation $\mathbf{x}$ to the class with the highest posterior probability. For a two-class problem with equal priors $P(\omega_1) = P(\omega_2) = 0.5$, the decision boundary is determined by the points where:

$$P(\omega_1|\mathbf{x}) = P(\omega_2|\mathbf{x}) \tag{5}$$

Which, using Bayes' theorem, is equivalent to:

$$p(\mathbf{x}|\omega_1)P(\omega_1) = p(\mathbf{x}|\omega_2)P(\omega_2) \tag{6}$$

Taking the natural logarithm of both sides and substituting the multivariate normal densities:

$$\ln\left(\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)}\right) + \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = 0 \tag{7}$$

$$\ln\left(\frac{|\mathbf{\Sigma}_2|^{1/2}}{|\mathbf{\Sigma}_1|^{1/2}}\right) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) + \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = 0 \tag{8}$$

This can be simplified to:

$$\frac{1}{2}\ln\left(\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}\right) - \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_1^{-1}\mathbf{x} + \frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}_1^{-1}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \tag{9}$$

$$+ \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_2^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}_2^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = 0 \tag{10}$$

Further simplification leads to a quadratic decision boundary in the form:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c = 0 \tag{11}$$

where:

$$\mathbf{A} = \frac{1}{2}(\mathbf{\Sigma}_2^{-1} - \mathbf{\Sigma}_1^{-1}) \tag{12}$$

$$\mathbf{b} = \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2 \tag{13}$$

$$c = \frac{1}{2}\ln\left(\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}\right) - \frac{1}{2}\boldsymbol{\mu}_1^T\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2 + \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right) \tag{14}$$

This quadratic equation describes the decision boundary between the two classes.

3.5 Classification Implementation

To classify each data point $\mathbf{x}$, we compute the posterior probabilities using Bayes' theorem:

$$P(\omega_1|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_1)P(\omega_1)}{p(\mathbf{x}|\omega_1)P(\omega_1) + p(\mathbf{x}|\omega_2)P(\omega_2)} \tag{15}$$

$$P(\omega_2|\mathbf{x}) = 1 - P(\omega_1|\mathbf{x}) \tag{16}$$

The classification rule is then:

$$\text{Predicted class} = \begin{cases} \omega_1, & \text{if } P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}) \\ \omega_2, & \text{otherwise} \end{cases} \tag{17}$$

3.6   Results and Analysis

Based on the classification of the given data points, the following results were obtained:

- Number of points classified as Class 1: 398

- Number of points classified as Class 2: 602

The Mahalanobis distance was also computed for each point with respect to both class means, providing insight into how distant points are from each class center in terms of the class covariance structure:

- Class 1 Mahalanobis distances:

    - Mean: 7.444
    - Minimum: 0.218
    - Maximum: 15.100

- Class 2 Mahalanobis distances:

    - Mean: 7.353
    - Minimum: 0.417
    - Maximum: 17.122

Figure below shows the visualization of the classification results. The points are colored according to their assigned class (blue for Class 1, red for Class 2), with the intensity of the color representing the posterior probability of the assigned class. The decision boundary between the two classes is shown as a dashed black line.
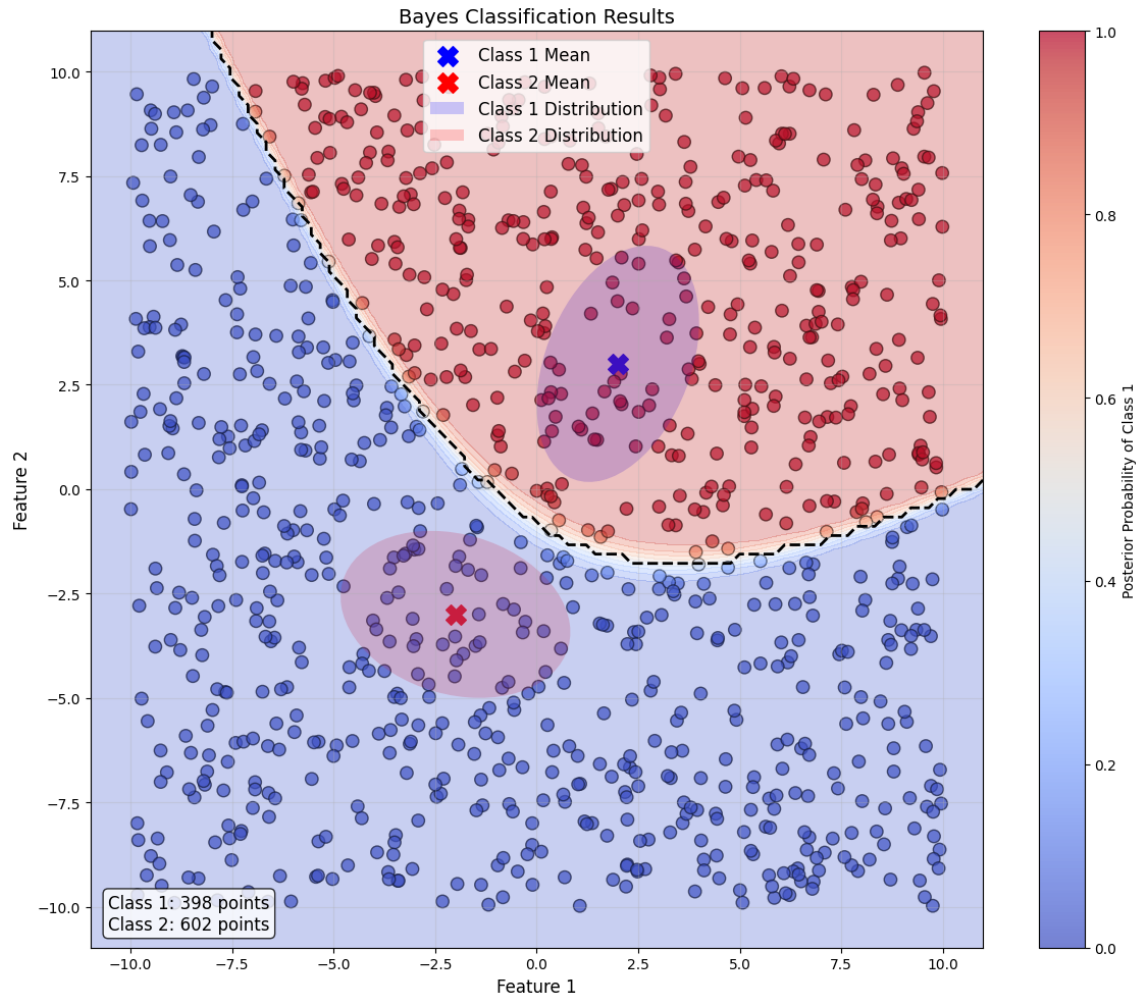
Figure 4: Bayes classification results showing the distribution of points classified into two classes. The class means are shown as blue and red crosses, and the background color represents the posterior probability of Class 1.

3.7 Discussion

Several observations can be made from the results:

1. The decision boundary between the two classes is not linear but quadratic. This is because the covariance matrices of the two classes are different, leading to a quadratic discriminant function.

2. The contours of equal posterior probability around each class mean form ellipses that are oriented according to the covariance matrices of the respective classes. For Class 1, the ellipses are stretched more along the vertical axis due to the higher variance in that direction ($\Sigma_{1,(2,2)} = 2$), while for Class 2, the ellipses are stretched more horizontally ($\Sigma_{2,(1,1)} = 2$).

3. The positive correlation in Class 1 (due to $\Sigma_{1,(1,2)} = 0.5$) causes the ellipses to be oriented in a direction with positive slope, while the negative correlation in Class 2 (due to $\Sigma_{2,(1,2)} = -0.3$) causes the ellipses to be oriented in a direction with negative slope.

4. Points that are close to the decision boundary have posterior probabilities close to 0.5 for both classes, indicating higher classification uncertainty.

3.8   Conclusion

Bayes classification with multivariate normal distributions provides a probabilistic framework for assigning data points to classes based on their posterior probabilities. The derived decision boundary depends on the means and covariance matrices of the class distributions, and can be linear or quadratic depending on whether the covariance matrices are equal across classes.

The approach not only provides a classification decision but also a measure of confidence through the posterior probabilities, which is valuable for understanding the reliability of the classifications. The Mahalanobis distance calculations further provide insight into how "typical" or "atypical" each point is with respect to its assigned class.

This implementation demonstrates the flexibility of Bayesian classification, which can accommodate any given values of means and covariance matrices for the class distributions, making it widely applicable in various domains.