EECE5644: Assignment 1

Dawei Wang
NUID: 002842604

October 15, 2025

1 Question 1

1.1 Data Generation

Consider a 3-dimensional random vector with class-conditional Gaussian distributions. The class priors are $P(L = 0) = 0.65$ and $P(L = 1) = 0.35$. The parameters are:

$$\boldsymbol{\mu}_0 = \begin{bmatrix} -0.5 \\ -0.5 \\ -0.5 \end{bmatrix}, \quad \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & -0.5 & 0.3 \\ -0.5 & 1 & -0.5 \\ 0.3 & -0.5 & 1 \end{bmatrix}, \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.3 & -0.2 \\ 0.3 & 1 & 0.3 \\ -0.2 & 0.3 & 1 \end{bmatrix} \tag{1}$$

10,000 samples generated according to this distribution. Which is showed in Figure 1.

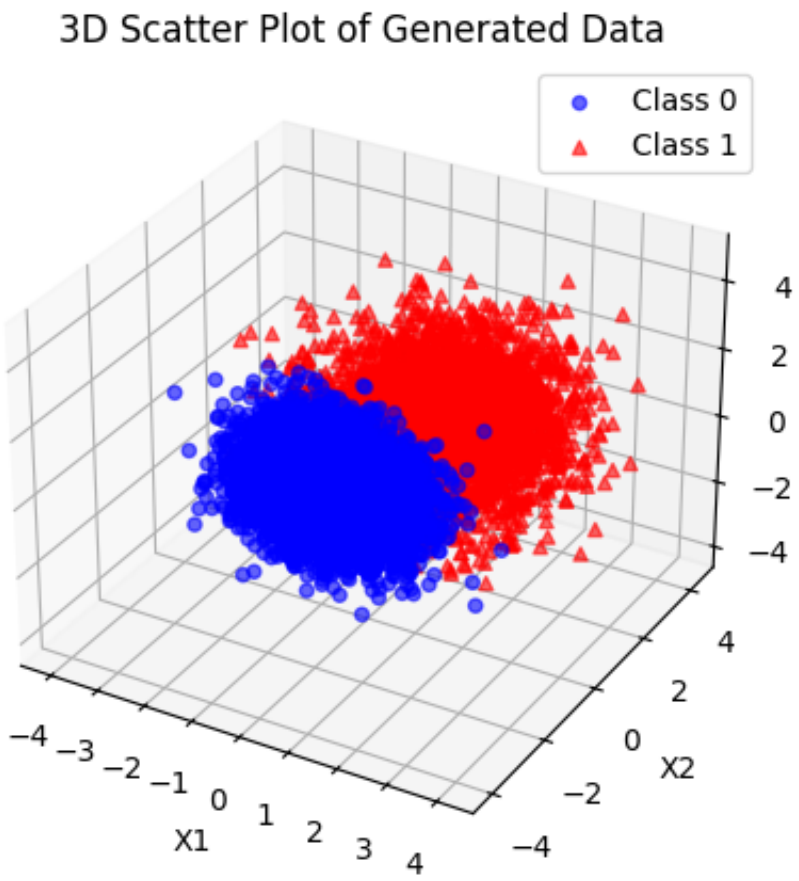

Figure 1: Generated Data Scatter

1.2 Part A: ERM Classification with True Distribution

1.2.1 Likelihood Ratio Test

The minimum expected risk classification rule is:

$$D(\mathbf{x}) = \begin{cases} 1, & \dfrac{p(\mathbf{x} \mid L=1)}{p(\mathbf{x} \mid L=0)} \geq \gamma, \\ 0, & \dfrac{p(\mathbf{x} \mid L=1)}{p(\mathbf{x} \mid L=0)} < \gamma. \end{cases} \tag{2}$$

For 0-1 loss, the threshold is:

$$\gamma = \frac{P(L=0)}{P(L=1)} = \frac{0.65}{0.35} \approx 1.857 \tag{3}$$

1.2.2 ROC Curve

Vary the threshold γ gradually from 0 to ∞ and computed:

- True Positive Rate: $P(D=1|L=1;\gamma)$
- False Positive Rate: $P(D=1|L=0;\gamma)$

The ROC curve is shown in Figure 2. At $\gamma = 0$, the curve is at $(1, 1)$, and as $\gamma \to \infty$, it reaches $(0, 0)$.

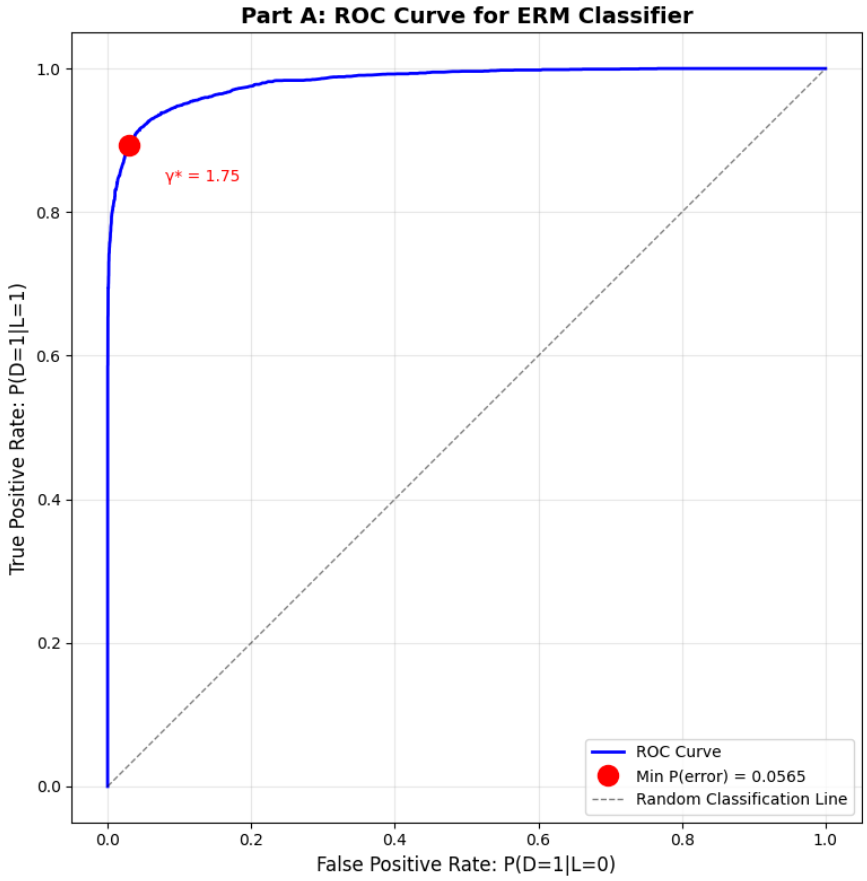

Figure 2: Part A: ROC Curve for ERM Classifier

1.2.3 Minimum Probability of Error

The probability of error is:

$$P(\text{error}; \gamma) = P(D = 1|L = 0; \gamma)P(L = 0) + P(D = 0|L = 1; \gamma)P(L = 1) \tag{4}$$

Results:

- Theoretical optimal threshold: $\gamma^* = 1.857$
- Empirical optimal threshold: $\gamma_{\text{emp}} = 1.7508$
- Minimum probability of error: $P(\text{error})_{\text{min}} = 0.0565$
- True Positive Rate at optimal point: 0.8931
- False Positive Rate at optimal point: 0.0294

The empirical threshold is close to the theoretical value.

1.3 Part B: Naive Bayes Classifier

Assume features are independent given each class, using identity covariance matrices $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \mathbf{I}$ while keeping the true means and priors. This represents incorrect knowledge of the data distribution.

Repeat the ROC curve analysis with this model. Results:

- NB minimum P(error): 0.0680
- Part A minimum P(error): 0.0565
- Performance degradation: 20.40%

Figure 3 compares the ROC curves. The model mismatch negatively impacts performance because the true covariance matrices have significant off-diagonal elements. Ignoring these correlations leads to incorrect probability estimates and suboptimal decision boundaries.

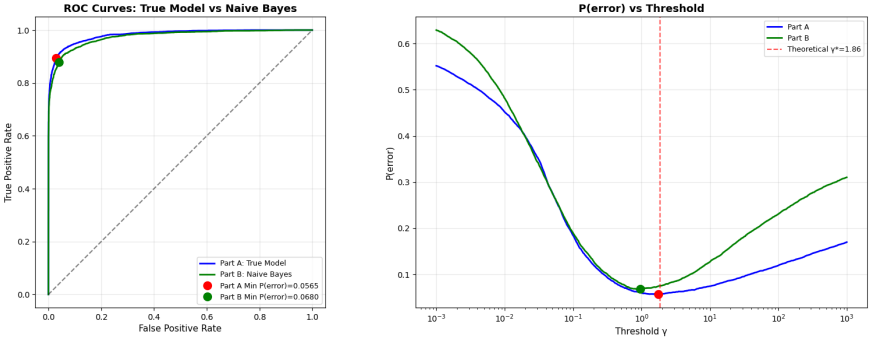

Figure 3: ROC Comparison: True Model vs Naive Bayes

1.4 Part C: Fisher LDA Classifier

1.4.1 Parameter Estimation

From the 10,000 samples, I estimated class means and covariances using sample averages:

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{N_j} \sum_{i:L_i=j} \mathbf{x}_i, \quad \hat{\boldsymbol{\Sigma}}_j = \frac{1}{N_j} \sum_{i:L_i=j} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)^T \tag{5}$$

1.4.2 Fisher LDA Projection

Using equal weights as specified, computing:

Within-class scatter: $\mathbf{S}_W = \frac{\hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1}{2}$

Between-class scatter: $\mathbf{S}_B = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)^T$

Fisher LDA projection: $\mathbf{w}_{\text{LDA}} = \mathbf{S}_W^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$

1.4.3 Classification and ROC Curve

Project data as $y = \mathbf{w}_{\text{LDA}}^T \mathbf{x}$ and classified based on threshold τ varying from $-\infty$ to ∞. The ROC curve is shown in Figure 4.

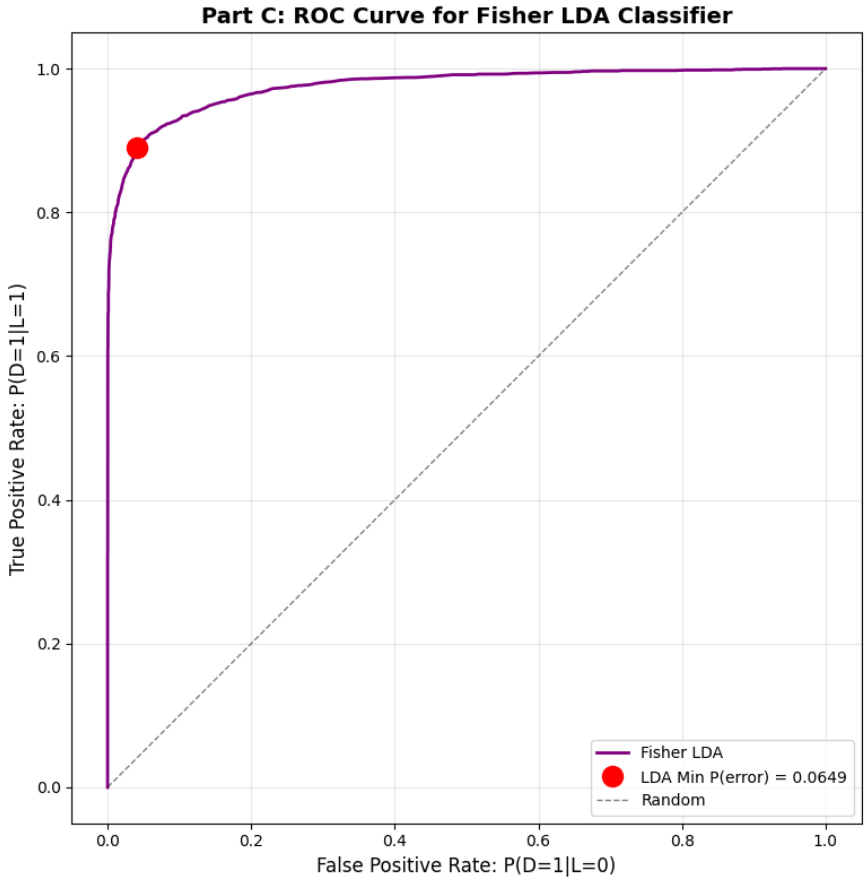

Figure 4: Part C: ROC Curve for Fisher LDA

Results:

- LDA minimum P(error): 0.0649
- Optimal threshold: $\tau^* = 0.3558$

Fisher LDA performs nearly as well as the true model classifier (Part A). This is because:

1. With 10,000 samples, parameter estimates are accurate
2. When classes are Gaussian with similar covariances, LDA finds the optimal linear boundary
3. The problem is only 3-dimensional, far from high-dimensional challenges

1.5 Final Comparison

Figure 5 compares all three methods.

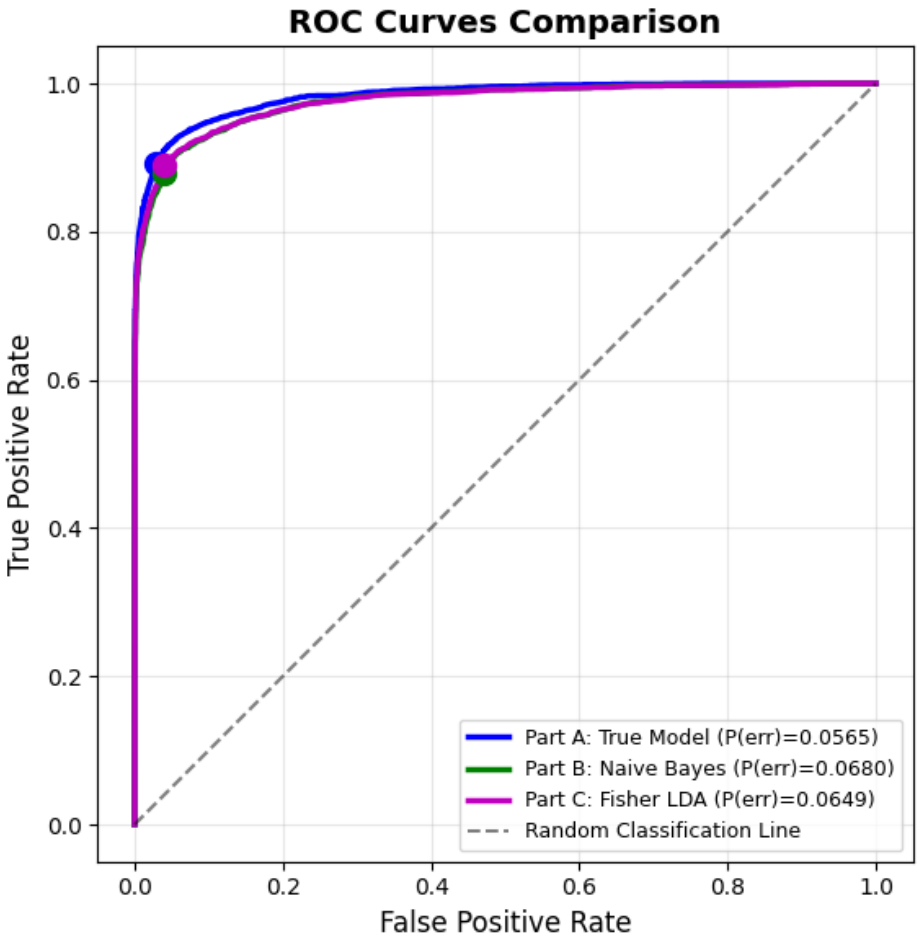

Figure 5: ROC Curves: All Three Methods

Method	Min P(error)	Accuracy
Part A: True Model	0.0565	94.35%
Part B: Naive Bayes	0.0680	93.20%
Part C: Fisher LDA	0.0649	93.51%

Table 1: Performance Comparison

Part A achieves the best performance. Part B shows degraded performance due to model mismatch. Part C performs nearly optimally despite using estimated parameters.

2 Question 2

2.1 Problem Setup

A 2-dimensional random vector $\mathbf{X}$ that takes values from a mixture of four Gaussians. Each Gaussian represents the class-conditional pdf for one of four class labels $L \in \{1, 2, 3, 4\}$.

2.1.1 Gaussian Parameters

The four Gaussian class-conditional pdfs are defined as follows:

$$\text{Class 1:} \quad \boldsymbol{\mu}_1 = \begin{bmatrix} -3 \\ 3 \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.0 & 0.3 \\ 0.3 & 1.0 \end{bmatrix}$$

$$\text{Class 2:} \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.2 & -0.2 \\ -0.2 & 0.8 \end{bmatrix}$$

$$\text{Class 3:} \quad \boldsymbol{\mu}_3 = \begin{bmatrix} 0 \\ -3 \end{bmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1.5 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$

$$\text{Class 4:} \quad \boldsymbol{\mu}_4 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Sigma}_4 = \begin{bmatrix} 2.0 & 0.5 \\ 0.5 & 2.0 \end{bmatrix}$$

All class priors are set to $P(L = j) = 0.25$ for $j \in \{1, 2, 3, 4\}$.

Class 4 is positioned at the center with the largest covariance, making it the class that overlaps most with the other 3 classes.

2.1.2 Data Generation

10000 samples generated from the mixture distribution:

1. For each sample, randomly select a class label L according to the prior distribution
2. Generate a sample $\mathbf{x}$ from the corresponding Gaussian $(\boldsymbol{\mu}_L, \boldsymbol{\Sigma}_L)$

One example:
Total samples: 10000
Class 1: 2483 samples (24.83%)
Class 2: 2508 samples (25.08%)
Class 3: 2523 samples (25.23%)
Class 4: 2486 samples (24.86%)

2.2 Part A: MAP Classification

2.2.1 Confusion Matrix

The empirical confusion matrix $P(D = i|L = j)$ estimated from 10000 samples is:

2.2.2 Results

- Overall Accuracy: 91.96%
- Error Rate: 8.04%

	$L = 1$	$L = 2$	$L = 3$	$L = 4$
$D = 1$	0.9863	0.0012	0.0000	0.0197
$D = 2$	0.0020	0.9705	0.0000	0.0571
$D = 3$	0.0000	0.0000	0.9239	0.1259
$D = 4$	0.0117	0.0283	0.0761	0.7973

Table 2: MAP Classifier Confusion Matrix

2.2.3 Visualization

Figure 6 shows the classified samples where:

- Different markers represent different true classes (circle, square, triangle, diamond)
- Green indicates correct classification
- Red indicates incorrect classification
- Contour lines show the Gaussian distributions

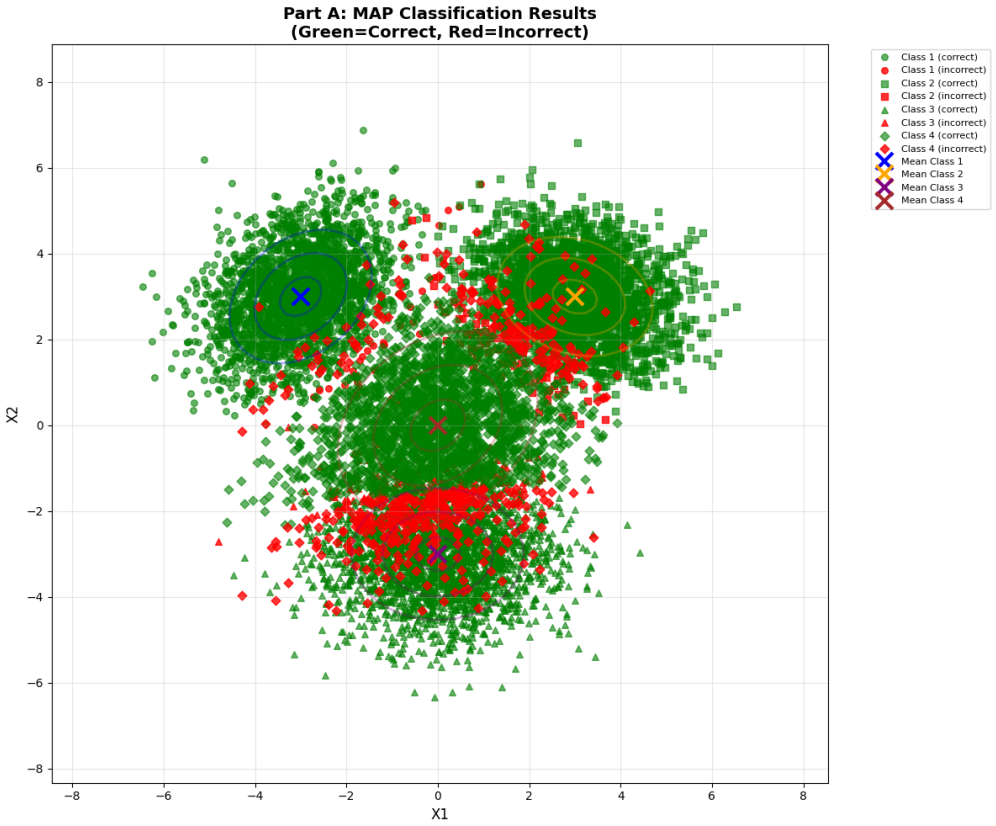

Figure 6: MAP Classification Results

2.3 Part B: ERM Classification (20 points)

2.3.1 Loss Matrix

The loss matrix $\mathbf{\Lambda}$ is defined as:

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & 10 & 10 & 100 \\ 1 & 0 & 10 & 100 \\ 1 & 1 & 0 & 100 \\ 1 & 1 & 1 & 0 \end{bmatrix} \tag{6}$$

where Λ_{ij} represents the loss for classifying a sample from class j as class i.

2.3.2 Confusion Matrix

The empirical confusion matrix for the ERM classifier is:

	$L=1$	$L=2$	$L=3$	$L=4$
$D=1$	0.7700	0.0000	0.0000	0.0012
$D=2$	0.0000	0.3764	0.0000	0.0032
$D=3$	0.0000	0.0000	0.0931	0.0008
$D=4$	0.2300	0.6236	0.9069	0.9948

Table 3: ERM Classifier Confusion Matrix

2.3.3 Minimum Expected Risk

The minimum expected risk is estimated as:

$$\hat{R}_{\min} = \frac{1}{N} \sum_{n=1}^{N} \Lambda_{D_n, L_n} \tag{7}$$

where $N = 10{,}000$, D_n is the decision for sample n, and L_n is the true label.
Results:

- Minimum Expected Risk: 0.5723
- Overall Accuracy: 55.64%

2.3.4 Classification Decision Distribution

Decision Class	Number of Samples
Class 1	1915
Class 2	952
Class 3	237
Class 4	6869

Table 4: ERM Classifier Decision Distribution

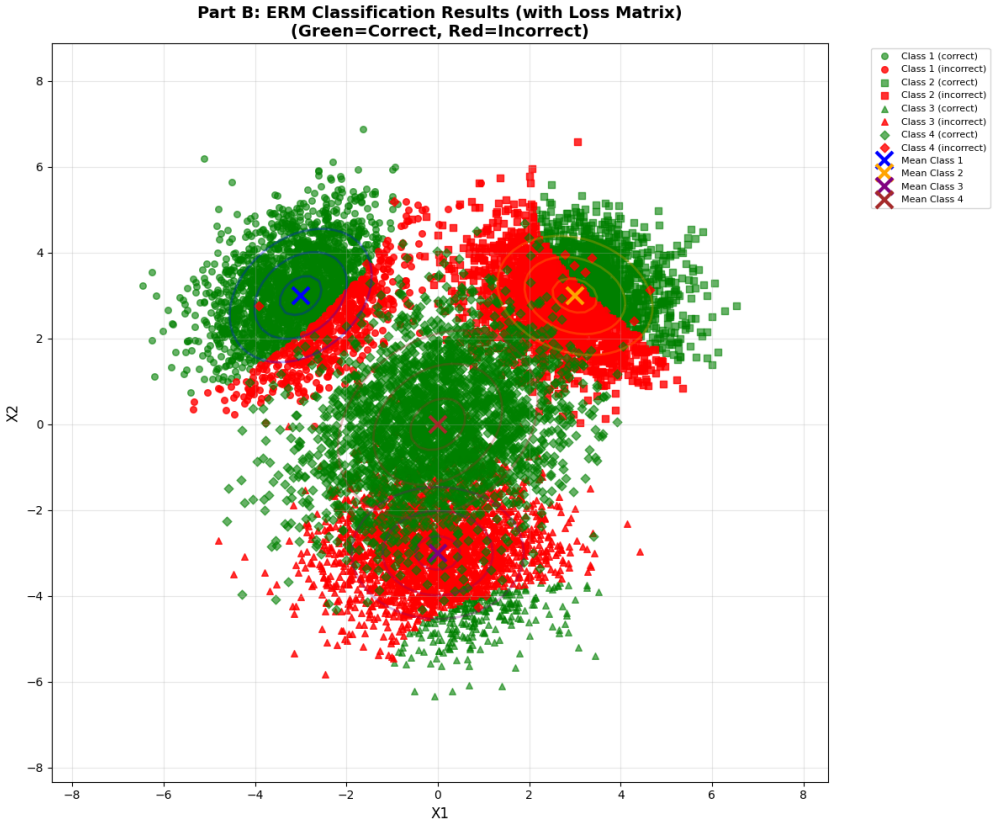

Figure 7: ERM Classification Results

2.3.5 Visualization

Figure 2 is the result of ERM Classification.

2.4 Comparison and Analysis

2.4.1 Performance Comparison

Metric	MAP Classifier	ERM Classifier
Accuracy	91.96%	55.64%
Expected Risk (under $\mathbf{\Lambda}$)	5.0727	0.5723

Table 5: Comparison of MAP and ERM Classifiers

1. **MAP Classifier:** Maximizes classification accuracy by choosing the class with the highest posterior probability. It treats all misclassification errors equally.
2. **ERM Classifier:** Minimizes expected risk by considering the different costs of misclassification. It strongly prefers classifying ambiguous samples as Class 4 to avoid the high penalty (loss = 100) of misclassifying Class 4 samples.

3. **Trade-off:** The ERM classifier achieves lower expected risk at the expense of lower accuracy. This demonstrates the fundamental trade-off between different loss functions.

4. **Decision Bias:** The ERM classifier shows a significant bias toward Class 4, classifying more samples as Class 4 than the true proportion. This is the optimal strategy given the asymmetric loss matrix.

3 Question 3

3.1 Datasets

3.1.1 Wine Quality Dataset

The Wine Quality dataset contains 4,898 white wine samples with 11 features. Each sample has a quality score from 3 to 9, giving us 7 classes. The dataset is imbalanced, with most samples having quality scores of 5, 6, or 7.

3.1.2 Human Activity Recognition (HAR) Dataset

The HAR dataset totally contains 10,299 samples with 561 features. There are 6 activity classes: Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing, and Laying. The dataset is relatively balanced across all classes, which contains:
Activity 1: 1722 samples (16.72%)
Activity 2: 1544 samples (14.99%)
Activity 3: 1406 samples (13.65%)
Activity 4: 1777 samples (17.25%)
Activity 5: 1906 samples (18.51%)
Activity 6: 1944 samples (18.88%)

3.2 Methodology

Assume each class follows a multivariate Gaussian distribution.

3.2.1 Parameter Estimation

For each class j, I estimate:

Class Prior:

$$\hat{P}(L = j) = \frac{n_j}{N} \tag{8}$$

Mean Vector:

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n_j} \sum_{i:y_i=j} \mathbf{x}_i \tag{9}$$

Covariance Matrix:

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{n_j} \sum_{i:y_i=j} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)^T \tag{10}$$

3.2.2 Covariance Regularization

To prevent numerical issues with ill-conditioned covariance matrices, especially in the high-dimensional HAR dataset, I apply regularization:

$$\boldsymbol{\Sigma}_{\text{regularized}} = \hat{\boldsymbol{\Sigma}} + \lambda \mathbf{I} \tag{11}$$

where:

$$\lambda = \alpha \cdot \frac{\text{trace}(\hat{\boldsymbol{\Sigma}})}{\text{rank}(\hat{\boldsymbol{\Sigma}})} \tag{12}$$

Use $\alpha = 0.01$. This ensures all eigenvalues are positive and the matrix is invertible.

3.2.3 MAP Classification

Use the MAP decision rule:

$$D^*(\mathbf{x}) = \arg\max_j P(L = j|\mathbf{x}) \tag{13}$$

where the posterior is computed using Bayes' theorem:

$$P(L = j|\mathbf{x}) = \frac{p(\mathbf{x}|L = j) \cdot P(L = j)}{\sum_k p(\mathbf{x}|L = k) \cdot P(L = k)} \tag{14}$$

3.3 Results

3.3.1 Wine Quality Dataset

Metric	Value
Overall Accuracy	30.67%
Error Rate	69.33%

Table 6: Wine Quality Classification Results

The confusion matrix is shown in Table 7 and Figure 8. The diagonal elements show correct classification rates, while off-diagonal elements show misclassification rates.

	$L = 3$	$L = 4$	$L = 5$	$L = 6$	$L = 7$	$L = 8$	$L = 9$
$D = 3$	0.200	0.012	0.001	0.000	0.000	0.000	0.000
$D = 4$	0.000	0.000	0.001	0.000	0.000	0.000	0.000
$D = 5$	0.150	0.067	0.051	0.020	0.002	0.000	0.000
$D = 6$	0.400	0.466	0.422	0.291	0.106	0.154	0.000
$D = 7$	0.200	0.454	0.524	0.686	0.885	0.817	1.000
$D = 8$	0.050	0.000	0.002	0.002	0.007	0.029	0.000
$D = 9$	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 7: Wine Quality Confusion Matrix $P(D = i|L = j)$

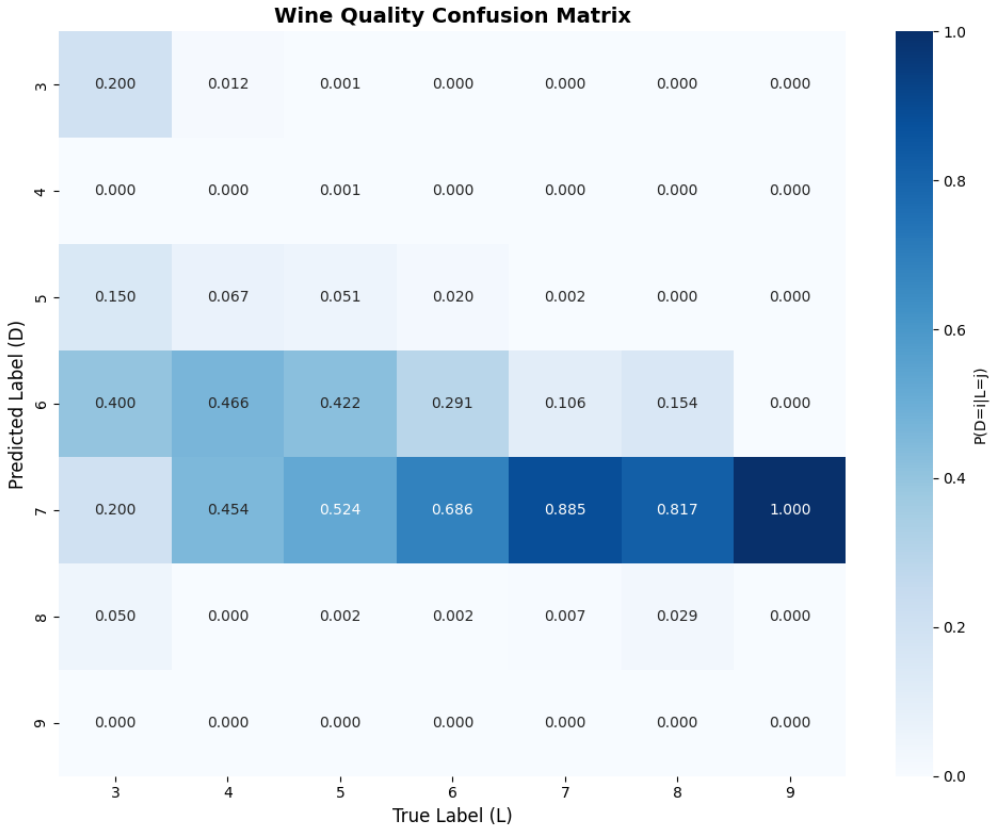

Figure 8: Wine Quality Confusion Matrix

The confusion matrix shows that the classifier has difficulty distinguishing between adjacent quality levels. This makes sense because wine quality is somewhat subjective and adjacent scores likely have similar chemical properties.

3.3.2 Human Activity Recognition Dataset

Metric	Value
Overall Accuracy	96.31%
Error Rate	3.69%

Table 8: HAR Classification Results

	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$
$D=1$	0.999	0.000	0.000	0.000	0.000	0.000
$D=2$	0.001	1.000	0.034	0.000	0.000	0.000
$D=3$	0.000	0.000	0.966	0.000	0.000	0.000
$D=4$	0.000	0.000	0.000	0.814	0.000	0.000
$D=5$	0.000	0.000	0.000	0.186	1.000	0.000
$D=6$	0.000	0.000	0.000	0.000	0.000	1.000

Table 9: HAR Confusion Matrix (1=Walking, 2=Upstairs, 3=Downstairs, 4=Sitting, 5=Standing, 6=Laying)

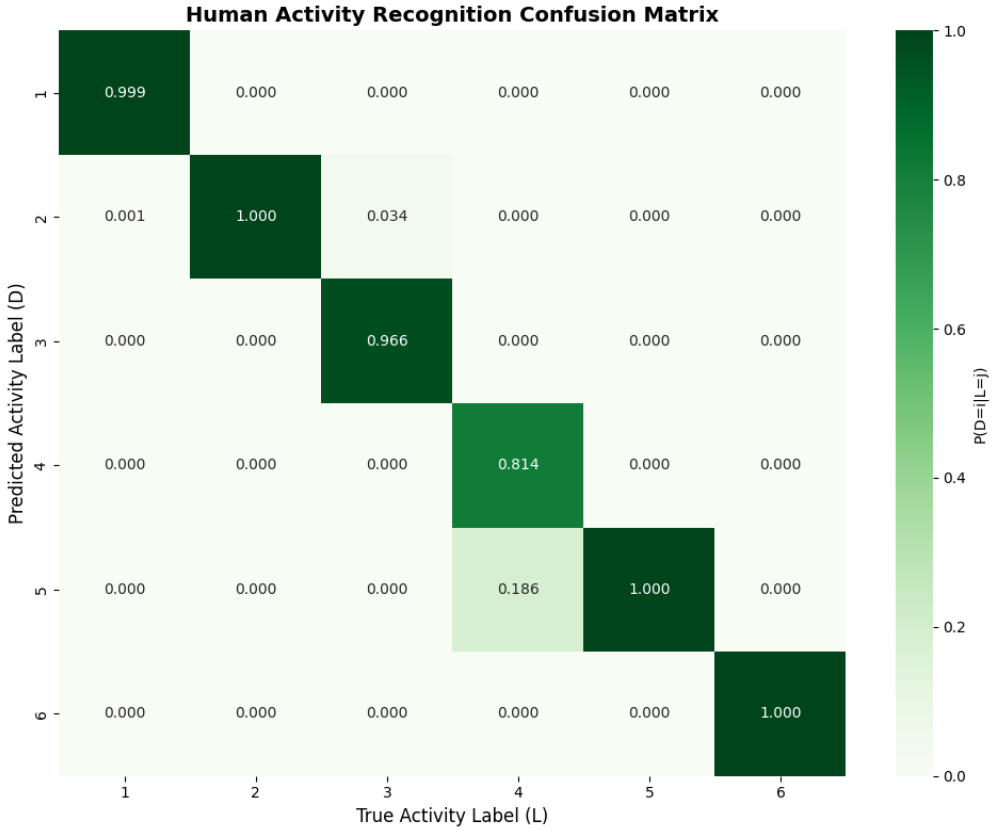

Figure 9: HAR Confusion Matrix

The HAR dataset achieves higher accuracy than Wine Quality. The confusion matrix shows good separation between dynamic activities (walking, upstairs, downstairs) and static activities (sitting, standing, laying), though some confusion exists within these groups.

3.4 Visualization

Use PCA to visualize the datasets by projecting features onto 2D and 3D spaces.

Wine Quality: The first 3 principal components explain approximately 99.91% of the variance. The visualization shows significant overlap between different quality classes, especially adjacent ones. There's no clear clustering, which explains the lower classification accuracy.

HAR: The first 3 principal components explain approximately 71.02% of the variance. The visualization shows much clearer separation between classes, especially between dynamic and static activities. This better structure explains the higher classification accuracy.

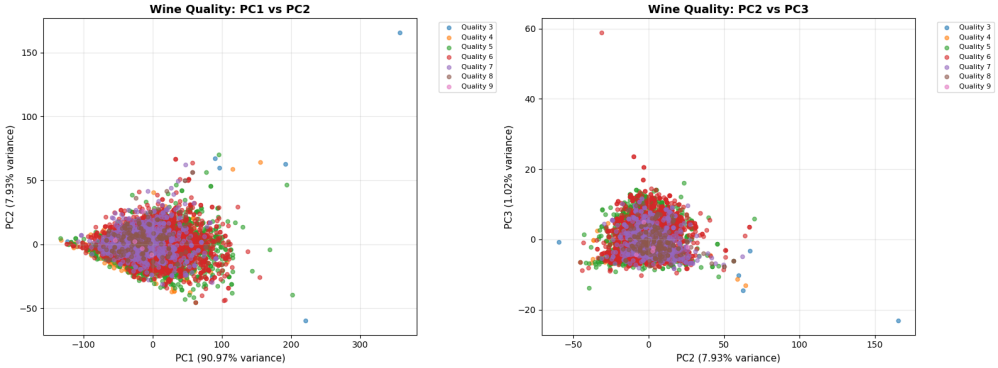

Figure 10: Wine PCA Visualization (2D)

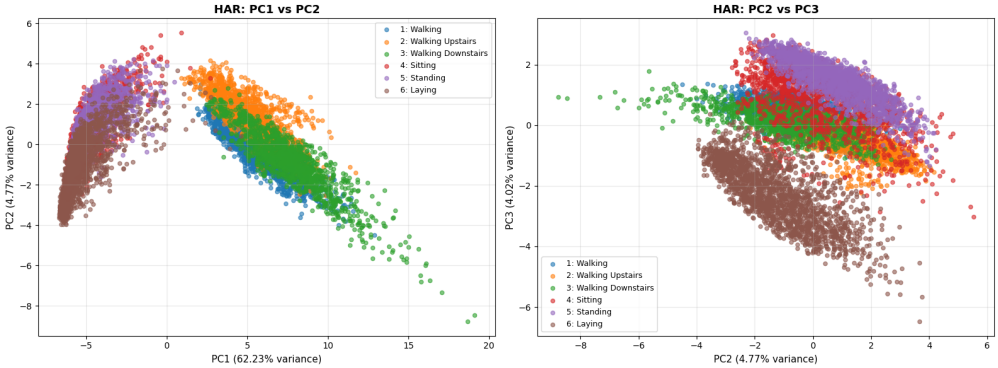

Figure 11: HAR PCA Visualization (2D)

3.5 Analysis and Discussion

3.5.1 Is the Gaussian Model Appropriate?

Wine Quality Dataset: The Gaussian assumption is not suitable for this dataset. Wine quality scores are ordinal variables with a natural ordering, but this model treats them as independent categories. This ignores the inherent structure in the data. Additionally, the PCA visualization reveals significant overlap between adjacent quality classes. The relationship between these 11 features and quality is likely nonlinear and complex, which simple Gaussian models cannot capture. Besides, the class imbalance leads to poor parameter estimates for the extreme classes.

These model limitations explain the low classification accuracy of approximately 30%. Looking at the confusion matrix, we see that most misclassifications occur between adjacent quality scores. This suggests that wine quality varies continuously rather than in discrete jumps, making it difficult to classify with Gaussian model.

Human Activity Recognition Dataset: The Gaussian assumption works much better for HAR.

The 561 features are statistical summaries, including means, standard deviations. According to the Central Limit Theorem, averages of many samples tend toward Gaussian distributions, so these engineered features are likely approximately Gaussian. Moreover, the PCA visualization shows clear separation between activity classes. The high dimensionality actually helps here because the model has enough parameters to capture the complex patterns that distinguish activities.

The higher classification accuracy (around 96%) confirms that the Gaussian model is appropriate. The confusion matrix shows strong diagonal elements, meaning most samples are correctly classified. Errors occur mainly between similar activities, which makes sense intuitively.

Related codes: https://github.com/wang-dawei1/EECE5644