Q9 - HW2: Pattern Recognition

Vahid Maleki
Student ID: 40313004

October 18, 2025

1 Question 9

Suppose in an n-dimensional space we have N training samples belonging to M different classes. Let N_1 samples belong to class ω_1, N_2 samples belong to class ω_2, and so on, up to N_M samples belonging to class ω_M, such that:

$$N = \sum_{j=1}^{M} N_j$$

The **overall mean** (centroid) of all training samples is defined as:

$$\mathbf{m} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j$$

The **mean of class** ω_i is defined as:

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$$

where $\mathbf{x}_{ij}$ denotes the j-th sample belonging to class i.

1.1 (a)

Express the overall mean $\mathbf{m}$ in terms of the class means $\mathbf{m}_i$.

1.1.1 Solution

The overall mean (centroid) of all training samples is:

$$\mathbf{m} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j$$

But the samples are grouped by class, so we can rewrite the sum over all samples as a sum over classes:

$$\mathbf{m} = \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$$

Recall that the mean of class ω_i is:

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$$

So, $\sum_{j=1}^{N_i} \mathbf{x}_{ij} = N_i \mathbf{m}_i$. Plug this into the overall mean:

$$\mathbf{m} = \frac{1}{N} \sum_{i=1}^{M} N_i \mathbf{m}_i$$

1.2 (b)

Let Σ_B, Σ_W, and Σ denote the **between-class**, **within-class**, and **total covariance matrices**, respectively, defined as follows:

$$\Sigma_B = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$$

$$\Sigma_W = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(\mathbf{x}_{ij} - \mathbf{m}_i)(\mathbf{x}_{ij} - \mathbf{m}_i)^T$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(\mathbf{x}_{ij} - \mathbf{m})(\mathbf{x}_{ij} - \mathbf{m})^T$$

Show that:

$$\Sigma = \Sigma_B + \Sigma_W$$

1.2.1 Solution

Let's recall the definitions: - **Between-class covariance:**

$$\Sigma_B = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$$

- **Within-class covariance:**

$$\Sigma_W = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(\mathbf{x}_{ij} - \mathbf{m}_i)(\mathbf{x}_{ij} - \mathbf{m}_i)^T$$

- **Total covariance:**

$$\Sigma = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N_i}(\mathbf{x}_{ij} - \mathbf{m})(\mathbf{x}_{ij} - \mathbf{m})^T$$

Let's expand $\mathbf{x}_{ij} - \mathbf{m}$:

$$\mathbf{x}_{ij} - \mathbf{m} = (\mathbf{x}_{ij} - \mathbf{m}_i) + (\mathbf{m}_i - \mathbf{m})$$

So,

$$(\mathbf{x}_{ij}-\mathbf{m})(\mathbf{x}_{ij}-\mathbf{m})^T = (\mathbf{x}_{ij}-\mathbf{m}_i)(\mathbf{x}_{ij}-\mathbf{m}_i)^T+(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T+(\mathbf{x}_{ij}-\mathbf{m}_i)(\mathbf{m}_i-\mathbf{m})^T+(\mathbf{m}_i-\mathbf{m})(\mathbf{x}_{ij}-\mathbf{m}_i)^T$$

When you sum over all samples in class i, the cross terms (the last two terms) vanish because:

$$\sum_{j=1}^{N_i}(\mathbf{x}_{ij} - \mathbf{m}_i) = \mathbf{0}$$

So, summing over all classes and samples:

$$\Sigma = \Sigma_W + \Sigma_B$$

1.3 (c)

Define a new variable using an n-dimensional vector $\mathbf{a}$ as:

$$Z_i = \mathbf{a}^T(\mathbf{x}_i - \mathbf{m})$$

Compute the **variance** of Z_i and express it in terms of Σ.

1.3.1 Solution

Define:

$$Z_i = \mathbf{a}^T(\mathbf{x}_i - \mathbf{m})$$

The variance of Z_i is:

$$\text{Var}(Z_i) = E\left[(Z_i)^2\right] = E\left[(\mathbf{a}^T(\mathbf{x}_i - \mathbf{m}))^2\right]$$

This can be rewritten as:

$$\text{Var}(Z_i) = \mathbf{a}^T E\left[(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T\right]\mathbf{a} = \mathbf{a}^T \Sigma \mathbf{a}$$

1.4 (d)

We wish to find a vector $\mathbf{a}$ that maximizes the following quantity:

$$\frac{\mathbf{a}^T \Sigma_B \mathbf{a}}{\mathbf{a}^T \Sigma_W \mathbf{a}}$$

Explain what maximizing this quantity means and why it is useful in classification.

1.4.1 Solution

We want to find $\mathbf{a}$ that maximizes:

$$J(\mathbf{a}) = \frac{\mathbf{a}^T \Sigma_B \mathbf{a}}{\mathbf{a}^T \Sigma_W \mathbf{a}}$$

Interpretation: - The numerator measures how far apart the class means are (projected onto $\mathbf{a}$). - The denominator measures the spread of samples within each class (projected onto $\mathbf{a}$).

Why is this useful? Maximizing this ratio finds a direction $\mathbf{a}$ that best separates the classes: it makes the projected class means as far apart as possible, while keeping the projected within-class scatter as small as possible. This is the principle behind **Fisher's Linear Discriminant Analysis (LDA)**.

1.5 (e)

Show that maximizing

$$\frac{\mathbf{a}^T \Sigma_B \mathbf{a}}{\mathbf{a}^T \Sigma_W \mathbf{a}}$$

is equivalent to maximizing

$$\frac{\mathbf{a}^T \Sigma_B \mathbf{a}}{\mathbf{a}^T \Sigma \mathbf{a}}$$

1.5.1 Solution

Recall from part (b):

$$\Sigma = \Sigma_B + \Sigma_W$$

So,

$$\mathbf{a}^T \Sigma \mathbf{a} = \mathbf{a}^T \Sigma_B \mathbf{a} + \mathbf{a}^T \Sigma_W \mathbf{a}$$

If you maximize:

$$\frac{\mathbf{a}^T \Sigma_B \mathbf{a}}{\mathbf{a}^T \Sigma_W \mathbf{a}}$$

Or:

$$\frac{\mathbf{a}^T \Sigma_B \mathbf{a}}{\mathbf{a}^T \Sigma \mathbf{a}}$$

The maximizing $\mathbf{a}$ will be the same, because maximizing one is equivalent to maximizing the other (since $\mathbf{a}^T \Sigma_B \mathbf{a}$ is always less than or equal to $\mathbf{a}^T \Sigma \mathbf{a}$).

1.6 (f)

Maximizing

$$\frac{\mathbf{a}^T \Sigma_B \mathbf{a}}{\mathbf{a}^T \Sigma \mathbf{a}}$$

is equivalent to maximizing $\mathbf{a}^T \Sigma_B \mathbf{a}$ subject to the constraint $\mathbf{a}^T \Sigma \mathbf{a} = 1$. Using the **Lagrange multiplier method**, maximize the above ratio under this constraint, and derive the relationship between the vector $\mathbf{a}$ and the matrices Σ_B and Σ.

What conclusion can be drawn from this result?

1.6.1 Solution

We want to maximize:

$$\mathbf{a}^T \Sigma_B \mathbf{a}$$

subject to:

$$\mathbf{a}^T \Sigma \mathbf{a} = 1$$

Set up the Lagrangian:

$$L(\mathbf{a}, \lambda) = \mathbf{a}^T \Sigma_B \mathbf{a} - \lambda(\mathbf{a}^T \Sigma \mathbf{a} - 1)$$

Take the derivative with respect to $\mathbf{a}$ and set to zero:

$$\frac{\partial L}{\partial \mathbf{a}} = 2\Sigma_B \mathbf{a} - 2\lambda \Sigma \mathbf{a} = 0$$

$$\Sigma_B \mathbf{a} = \lambda \Sigma \mathbf{a}$$

This is a **generalized eigenvalue problem**:

$$\Sigma_B \mathbf{a} = \lambda \Sigma \mathbf{a}$$

The solution $\mathbf{a}$ is the eigenvector of $\Sigma^{-1}\Sigma_B$ corresponding to the largest eigenvalue λ.

Conclusion: - The optimal direction $\mathbf{a}$ for class separation is the eigenvector of $\Sigma^{-1}\Sigma_B$ with the largest eigenvalue. - This is the basis of Fisher's Linear Discriminant Analysis (LDA): it finds the direction that best separates classes by maximizing the ratio of between-class to total (or within-class) variance.