Student ID	
Name	

Sumission Guidelines

1. **Due date:** October 30, 2023, at 18:00 (6:00 PM)
2. **Problem types:** This assignment consists of theoretical derivations (Problem 1-3) and programming assignment ('assignment1.ipynb').
3. **Submission methods: Soft copy**
   (a) Please save the hand-written answer as a PDF, then create a zip file with the ipynb file and submit it to Blackboard under the name '(assignment1)_student_id_name.zip'.
   (b) The submission file consists of
   - '(assignment1)_student_id_name_thcprob.pdf'
   - '(assignment1)_student_id_name_progprob.ipynb'

   (c) i.e., '(assignment1)_20221111_SaeromPark.zip'
4. **Requirements:**
   (a) Problems 1-2 should be written by hand.
   (b) Please write clearly so that it is easy to understand.
   (c) For Problems 1-2, include clear and step-by-step derivations.
   (d) For Problems 1-2, provide explanations for your derivations and any assumptions made.
   (e) The detailed guidelines for the coding assignment will be provided in 'assignment1.ipynb' file.
   (f) Late submission will not be accepted.
   (g) Plagiarism and collaboration are prohibited. Even if you get some help from another student, you must write your own answers and code by yourself.

1. Suppose we have trained a linear regression model $\hat{f}(\mathbf{x}) = \hat{\boldsymbol{\beta}}^T\mathbf{x}$.

- **Notations**: We have a data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$, and a label vector $\mathbf{y} = [y_1, \dots, y_n]^T \in \mathbb{R}^n$.
- **IID Assumption**: We assume that the true relationship between the input features and the response variable is given by $Y_i = f(\mathbf{X}_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and ϵ_i's are independent.
- **Least Squares (LS) Solution**: The LS solution for estimating the regression coefficients is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

(a) Derive the mean and variance of the parameter vector $\hat{\boldsymbol{\beta}}$

Answer: Distribution of $\boldsymbol{\beta}$: $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, $Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$

Mean of $\hat{\beta}$:

The mean of $\hat{\beta}$ can be found by taking the expected value of $\hat{\beta}$:

$$E(\hat{\boldsymbol{\beta}}) = E((\boldsymbol{X}^T\boldsymbol{X})^{-1}X^T\boldsymbol{y})$$

Given that $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ ($\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ and $E(\boldsymbol{\epsilon}) = 0$), we can simplify the above equation:

$$E(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^TE(\boldsymbol{y})$$

Now, using properties of the expected value and assuming that $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are not stochastic (i.e., they are fixed), we can further simplify:

$$E(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X\beta}$$

Since $\boldsymbol{X}^T\boldsymbol{X}$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ are non-stochastic matrices, they can be factored out of the expectation operator:

$$E(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X\beta} = \boldsymbol{\beta}$$

Variance of $\hat{\boldsymbol{\beta}}$:

$$Var(\hat{\boldsymbol{\beta}}) = E((\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))^T)$$

Substituting the expression for $E(\hat{\boldsymbol{\beta}})$ that we derived earlier, we get:

$$Var(\hat{\boldsymbol{\beta}}) = E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T)$$

Now, we need to consider the properties of the residuals $\boldsymbol{\epsilon}$: $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X\beta}$ and $E(\boldsymbol{\epsilon}) = 0$

Therefore, the variance simplifies to:

$$Var(\hat{\boldsymbol{\beta}}) = E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T) = E((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

Again, we assume that $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are not stochastic, so they can be factored out of the expectation:

$$Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Var(\boldsymbol{\epsilon})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Because $Var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$, we can further simplify:

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

(b) Derive the prediction interval for the predicted value $\hat{f}(\boldsymbol{x}^*)$ for a new test data $\boldsymbol{x}^*$. (Hint: calculate the variance of forecasting error: $e^* = Y^* - \hat{Y}^*$)

Answer: To derive the prediction interval for the predicted value $f(\boldsymbol{x}^*)$ for a new test data point $\boldsymbol{x}^*$, we can use the variance of the forecasting error $e^* = Y^* - \hat{Y}^*$.

The prediction interval is typically constructed by considering the distribution of the forecasting error. We assume the noise distribution ϵ^* follows a normal distribution with mean 0 and variance σ^2 where $\hat{Y}^* = \boldsymbol{\beta}^T\boldsymbol{x}^* + \epsilon^*$.

First, calculate the forecasting error variance ($Var(e^*)$):

$$Var(e^*) = Var(Y^* - \hat{Y}^*) = Var(Y^*) + Var(\hat{Y}^*) - 2Cov(Y^*, \hat{Y}^*),$$

where $Cov(Y^*, \hat{Y}^*) = 0$.

Note that $Y^* = (\boldsymbol{x}^*)^T\boldsymbol{\beta} + \epsilon^*$, $\hat{Y}^* = (\boldsymbol{x}^*)^T\hat{\boldsymbol{\beta}}$, and

$$\begin{aligned}
Cov(Y^*, \hat{Y}^*) =& E[(Y^* - E[Y^*])(\hat{Y^*} - E[\hat{Y}^*]] \\
=& E[((x^*)^T \beta + \epsilon^* - (\boldsymbol{x}^*)^T \boldsymbol{\beta})((\boldsymbol{x}^*)^T \hat{\boldsymbol{\beta}} - E[(\boldsymbol{x}^*)^T \hat{\boldsymbol{\beta}}]] \\
=& E[\epsilon^*((\boldsymbol{x}^*)^T \hat{\boldsymbol{\beta}} - (\boldsymbol{x}^*)^T \boldsymbol{\beta})] \\
=& E[\epsilon^*(x^*)^T \hat{\beta} - \epsilon^*(x^*)^T \boldsymbol{\beta}] \\
=& E[\epsilon * (\boldsymbol{x}^*)^T \hat{\boldsymbol{\beta}}] - E[\epsilon^*(\boldsymbol{x}^*)^T \boldsymbol{\beta}] \\
=& E[\epsilon^*] * E[\boldsymbol{x}^*)^T \hat{\boldsymbol{\beta}}] - E[\epsilon^*](\boldsymbol{x}^*)^T \boldsymbol{\beta} = 0
\end{aligned}$$

because of $E[\epsilon^*] = 0$
Therefore, the forecasting error variance can be represented as:

$$Var(e^*) = \sigma^2 + \sigma^2 (\boldsymbol{x}^*)^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}^*$$

Then, we can calculate the prediction interval using this forecasting error variance. The prediction interval for a new data point $\boldsymbol{x}^*$ is given by:
$f(\boldsymbol{x}^*) \pm t_{n-p-1,\alpha/2} * \sqrt{Var(e^*)}$
The value of α depends on the desired confidence level for the prediction interval.
Finally, this prediction interval provides a range in which we can expect the true value Y* to lie with a specified confidence level.

2. Fisher's Linear Discriminant Analysis (Fisher's LDA) aims to find a linear transformation of the data that maximizes the between-class variance while minimizing the within-class variance. On the other hand, Gaussian Discriminant Analysis (GDA) models the likelihood of data in each class as a Gaussian distribution.

   Suppose we have the following dataset in Table 1.

   - **(Notations)** discriminant function: $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$ with $\boldsymbol{w} \in \mathbb{R}^2$, training data: $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^n$, the index set for class k: $\mathcal{C}_k$ ($|\mathcal{C}_k| = n_k$ and $\sum_k n_k = n$), the class mean of input data for class k: $\boldsymbol{\mu}^{(k)} \in \mathbb{R}^2$
   - **Fisher's LDA**: maximize $J(\boldsymbol{w})$

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}} \tag{1}$$

Feature 1	Feature 2	Label
3.05	4.24	class 1
2.58	5.34	class 1
3.12	3.41	class 1
2.57	3.9	class 1
1.93	4.24	class 1
4.58	4.37	class 2
4.95	3.57	class 2
4.76	3.7	class 2
5.39	3.37	class 2
3.68	4.98	class 3
1.96	5.89	class 3
4.01	5.05	class 3
4.86	4.62	class 3
3.52	5.38	class 3
4.41	6.38	class 3

Table 1: Example Data

(a) For Fisher's LDA, calculate $\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \boldsymbol{\mu}^{(3)}$, the between-class covariance $\boldsymbol{S}_B$, and the within-class covariance $\boldsymbol{S}_W$ for the example data in Table 1.

Answer: For Class 1,

$$\begin{aligned}\boldsymbol{\mu}_1 =&((3.05+2.58+3.12+2.57+1.93)/5,\\ &(4.24+5.34+3.41+3.9+4.24)/5) = (2.85, 4.426)\end{aligned}$$

For Class 2,

$$\begin{aligned}\boldsymbol{\mu}_2 =&((4.58+4.95+4.76+5.39)/4, (4.37+3.57+3.7+3.37)/4)\\ =&(4.67, 3.775)\end{aligned}$$

For Class 3,

$$\boldsymbol{\mu}_3 = ((3.68+1.96+4.01+4.86+3.52+4.41)/6,$$
$$(4.98+5.89+5.05+4.62+5.38+6.38)/6) = (3.80, 5.21)$$

And then we need to calculate $\boldsymbol{\mu}_{total} = (5/15) * \boldsymbol{\mu}_1 + (4/15) * \boldsymbol{\mu}_2 + (6/15) * \boldsymbol{\mu}_3$

The between-class covariance (S_B) for 3 classes is defined as $\sum_{i=1}^{3} n_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{total})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{total})^T$

$$\begin{aligned}S_B =&5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{total})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{total})^T +\\ &4(\boldsymbol{\mu}_2\boldsymbol{\mu}_{total})(\boldsymbol{\mu}_2\boldsymbol{\mu}_{total})^T + 6(\boldsymbol{\mu}_3\boldsymbol{\mu}_{total})(\boldsymbol{\mu}_3\boldsymbol{\mu}_{total})^T\end{aligned}$$

Now, let's calculate the within-class covariance (S_W) which is defined as $\sum_{i=1}^{3}\sum_{j=1}^{n_i}(x_{ij}-\boldsymbol{\mu}_i)(x_{ij}-\boldsymbol{\mu}_i)^T$

$$S_{W1} = \sum_{i=1}^{5}(X_i - \boldsymbol{\mu}^1)(X_i - \boldsymbol{\mu}^1)^T$$
$$S_{W2} = \sum_{i=1}^{4}(X_i - \boldsymbol{\mu}^2)(X_i - \boldsymbol{\mu}^2)^T$$
$$S_{W3} = \sum_{i=1}^{6}(X_i - \boldsymbol{\mu}^3)(X_i - \boldsymbol{\mu}^3)^T$$

As a result, $S_W = S_{W1} + S_{W2} + S_{W3}$

(b) Find a projection $\boldsymbol{w}$ that maximizes $J(\boldsymbol{w})$ in (1).

Answer: We then use the determinant of the scatter matrices to obtain a scalar objective function:

$$J(W) = \frac{|S_B|}{|S_W|} = \frac{W^T S_B W}{W^T S_W W} = trace((S_W)^{-1}S_B)$$

We will seek the projection matrix W^* that maximizes this ratio.
It can be shown that the optimal projection matrix W^* is the one whose columns are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem.

$$W^* = [w_1^*, w_2^*] = argmax_W\{\frac{|W^T S_B W|}{|W^T S_W W|}\} => (S_B - \lambda_i S_W)w_i^* = 0$$

(where i=1,2)

(c) For the above example in Table 1, we can obtain at most 2 projection directions. Explain why. What if we have 3 features? Or what if we have 4 classes?

Answer: In general, you can obtain at most $C-1$ projection directions for LDA, where C is the number of classes. This is because $rank(\boldsymbol{S}_B) \leq min(C-1, d)$ where we can find a non-zero linear combination for the columns of $\boldsymbol{S}_B$ (refer to Classification II page 9). In addition, for this example, $d = 2$, $rank(\boldsymbol{S}_B) = 2$ and $rank(\boldsymbol{S}_W) = 2$. In our problem, we have 3 classes so we can obtain at most 2 projection directions.
If we have 3 features, this doesn't change the maximum number of

projection directions, which is C-1. So we can still obtain at most 2 projection directions if we have 3 features.
Similarly, if we have 4 classes, we can obtain at most 2 projection directions because $d = 2$.

(d) Assume that the above example has the shared covariance matrix $\Sigma = (0.6)\boldsymbol{I}_2$ ($\boldsymbol{I}_2$: 2×2 identity matrix). Calculate the decision boundary for the linear GDA classifier between classes 1 and 3.

Answer: In the GDA, the decision boundary between classes is determined by the quadratic discriminant function.
The discriminant function for GDA is as follows:

$$\log P(X|Y = i) = g_i(x) = -\frac{1}{2}(x - \boldsymbol{\mu}_i)^T \Sigma^{-1}(x - \boldsymbol{\mu}_i) - \frac{1}{2}log|\Sigma| + log(\pi_i)$$

Given that $\sum = 0.6I_2$ is a diagonal matrix, its inverse is simply $\frac{1}{0.6}I^2 = 5/3I_2$.
The prior probabilities are given as $\pi_1 = 5/15, \pi_3 = 6/15$ We can obtain $\boldsymbol{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$ and $w_{i0} = -1/2\boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \log \pi_i$, where $g_i(x) = \boldsymbol{w}_i^T\boldsymbol{x} + w_{i0}$
Now, we can calculate the decision boundary by setting $g_1(x) = g_3(x)$ and solving for x:

$$-\frac{1}{2}(x - \boldsymbol{\mu}_1)^T \frac{5}{3}I_2(x - \boldsymbol{\mu}_1) - \frac{1}{2}log|0.6I_2| + log(5/15)$$
$$= -\frac{1}{2}(x - \boldsymbol{\mu}_3)^T \frac{5}{3}I_2(x - \boldsymbol{\mu}_3) - \frac{1}{2}log|0.6I_2| + log(6/15)$$

Then, you can simplify this equation and get the linear decision boundary.

(e) Assume that we do not have the shared covariance matrix ($\Sigma_k \neq \Sigma_{k'}$). Calculate the decision boundary between classes 1 and 3 for $\Sigma_1 \neq \Sigma_3$.

Answer: When you don't assume a shared covariance matrix, you have to use a Quadratic Discriminant Analysis (QDA) approach. In QDA, you allow for different covariance matrices for each class. First, compute the discriminant functions:

$$g_i(x) = -\frac{1}{2}(x - \boldsymbol{\mu}_i)^T \Sigma_i^{-1}(x - \boldsymbol{\mu}_i) - \frac{1}{2}log|\Sigma_i| + log(\pi_i)$$

Here, you need to use the class-specific covariance matrices and means.

And then, set $g_1(x) = g_3(x)$:

$$\begin{aligned}&-\frac{1}{2}(x-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(x-\boldsymbol{\mu}_1) - \frac{1}{2}log|\Sigma_1| + log(\pi_1)\\ =&-\frac{1}{2}(x-\boldsymbol{\mu}_3)^T\Sigma_3^{-1}(x-\boldsymbol{\mu}_3) - \frac{1}{2}log|\Sigma_3| + log(\pi_3)\end{aligned}$$

$$\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_3^{-1})\boldsymbol{x} - (\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_3^T\boldsymbol{\Sigma}_3^{-1})\boldsymbol{x} + \cdots = 0$$

(f) Discuss the relationship between Fisher's LDA and GDA. Answer this question based on the assumptions that they make about the data and the advantages and limitations of them.

Answer:

Relationship between Fisher LDA and GDA:
Fisher LDA does **not have specific assumptions** about the distribution of the data while GDA assumes that the data is **normally distributed** within each class. Fisher LDA only provides **a linear decision boundary** while GDA can have **a non-linear decision boundary**. However, the assumption of normality can be a limitation of GDA. If the data is not normally distributed or the mean or covariance estimates are not accurate, GDA can produce poor classification performances. In addition, GDA finds one projection direction while LDA can find multiple projection directions and can be used as a dimensionality reduction technique.