STA 712 Challenge Assignment 1

Due: By Friday, November 10 at 12:00pm (noon) on Canvas.

Instructions:

- Submit your work as a single PDF. Use TEXto type and format any math and equations, either with a LATEXeditor (Overleaf, Texmaker, etc.), or directly in an R Markdown or Quarto document. Include all R code needed to reproduce your results in your submission.
- You are welcome to work with others on this assignment, but you must submit your own work.
- The goal of this assignment is for you to learn about a topic beyond the core material covered in class. If you get stuck, I am happy to chat over email or in office hours.
- You can probably find the answers to many of these questions online. It is ok to use online resources! But make sure to show all your work in your final submission.

Linear Discriminant Analysis (LDA) vs. Logistic Regression

The first topic in STA 712 is logistic regression. Logistic regression allows us to model the relationship between a binary response Y and set of covariates $\boldsymbol{X}$. In the logistic regression model,

$$P(Y_i = 1|\boldsymbol{X}_i) = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{X}_i}}{1 + e^{\boldsymbol{\beta}^T \boldsymbol{X}_i}}$$

.

However, logistic regression is not the only option for modeling $P(Y_i = 1|\boldsymbol{X}_i)$. Other classifiers, like neural networks, random forests, and support vector machines, also exist. In this assignment, we will study a classification method called *linear discriminant analysis* (LDA).

Overview of LDA

Let $Y \in \{0, 1\}$ be a binary response variable, and $\boldsymbol{X} \in \mathbb{R}^k$ be a vector of covariates. LDA assumes that, conditional on Y, $\boldsymbol{X}$ follows a multivariate normal distribution. That is,

$$\boldsymbol{X}|(Y = 0) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) \qquad \text{and} \qquad \boldsymbol{X}|(Y = 1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1 \in \mathbb{R}^k$ are the means, and $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ is the covariance matrix. *Note that LDA assumes the same covariance matrix for both distributions.*

1. Let $\pi_1 = P(Y_i = 1)$ be the marginal probability that $Y = 1$ in the population, and let $\pi_0 = 1 - \pi_1$. Use Bayes' theorem and (1) to show that, if the LDA model assumptions are correct, $P(Y_i = 1|\boldsymbol{X}_i)$ is given by

$$P(Y_i = 1|\boldsymbol{X}_i) = \frac{\pi_1 \exp\{-\frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_1)\}}{\pi_1 \exp\{-\frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_1)\} + \pi_0 \exp\{-\frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)\}}.$$

2. Suppose we observe data $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$. We fit the LDA model, estimating $\pi_1, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}$. The covariance $\boldsymbol{\Sigma}$ is estimated with a pooled sample estimate:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-2} \sum_{i=1}^{n} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i})^T,$$

where $\widehat{\boldsymbol{\mu}}_{Y_i} = \widehat{\boldsymbol{\mu}}_1$ if $Y_i = 1$, and $\widehat{\boldsymbol{\mu}}_{Y_i} = \widehat{\boldsymbol{\mu}}_0$ if $Y_i = 0$.

In this question, we will fit the model to the `dengue` data from class:

```
dengue <- read.csv("https://sta711-s23.github.io/homework/dengue.csv")
```

Let Y_i be a patient's dengue status, and let $\boldsymbol{X}_i = (WBC_i, PLT_i)$ be a patient's white blood cell count and platelet count. Fit the LDA model (1) to this dengue data, and report the estimates $\widehat{\pi}_1$, $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\mu}}_1$, and $\widehat{\boldsymbol{\Sigma}}$.

3. Now fit a logistic regression model with dengue status as the response, and WBC and PLT as predictors. Report your fitted coefficients $\widehat{\boldsymbol{\beta}}$ for the logistic regression model.

4. In R, make a plot showing the relationship between the predicted probabilities $\widehat{P}(Y_i = 1|\boldsymbol{X}_i)$ from logistic regression, and the predicted probabilities $\widehat{P}(Y_i = 1|\boldsymbol{X}_i)$ from LDA. Do the two methods give similar predictions?

5. It turns out that LDA is the "same" as logistic regression, when the LDA assumptions hold. Show that if (1) holds, then

$$\log\left(\frac{P(Y_i = 1|\boldsymbol{X})}{P(Y_i = 0|\boldsymbol{X}_i)}\right) = \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i.$$

Conclude that if the LDA assumptions hold, then the log-odds are a linear function of the covariates $\boldsymbol{X}$ (which is what we assume in logistic regression!).

LDA vs. logistic regression

If LDA is the "same" as logistic regression, why do both methods exist? Several reasons:

- LDA assumes the data come from multivariate normal distributions. If this parametric assumption doesn't hold (and it usually doesn't), then logistic regression and LDA are *not* the same, and logistic regression is more flexible.
- Fitting LDA is computationally much easier than fitting logistic regression. LDA just requires estimates $\widehat{\pi}_1$, $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\mu}}_1$, and $\widehat{\boldsymbol{\Sigma}}$, all of which have a closed form. This avoids iterative methods like Fisher scoring.
- If the LDA assumptions hold, then LDA and logistic regression are both trying to estimate the same parameters. But, since different estimation methods are used, the fitted probabilities are slightly different.

In the final part of this assignment, you will compare LDA and logistic regression in a small simulation. Suppose that $\boldsymbol{X}_i \in \mathbb{R}^2$, and (1) holds, with $\pi_1 = 0.5$, $\boldsymbol{\mu}_0 = (0, 0)^T$, $\boldsymbol{\mu}_1 = (0.5, 0.5)^T$, and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

6. In R, generate 10 training samples $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_{10}, Y_{10})$ from the LDA model. **Hint:** sample Y first, then sample $\boldsymbol{X}|Y$ from the appropriate multivariate normal distribution.

7. Using your training sample from question 6, fit LDA and logistic regression models.

8. Now generate 1000 test samples $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_{1000}, Y_{1000})$ from the LDA model. Using your fitted models from question 6 (do not re-fit the models on the test data!), calculate (a) the estimated probabilities $\widehat{P}(Y_i = 1|\boldsymbol{X}_i)$ using the fitted LDA model, (b) the estimated probabilities $\widehat{P}(Y_i = 1|\boldsymbol{X}_i)$ using the fitted logistic regression model, and (c) the true probabilities $P(Y_i = 1|\boldsymbol{X}_i)$ for each point (using the true parameters π_1, $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}$).

9. Which predictions (LDA or logistic regression) are closer, on average, to the true probabilities for your test data?

10. Repeat questions 6 – 9 200 times. When the LDA assumptions hold and the number of training samples is $n_{train} = 10$, which method – LDA or logistic regression – does a better job estimating the true probabilities?

11. Repeat question 10 for different training sizes n_{train}. How does the relative performance of LDA and logistic regression change as we increase n_{train}? Make a plot summarizing the performance of LDA and logistic regression at each value of n_{train}.