STA 712 Challenge Assignment 1

Due: Friday, September 23, 12:00pm (noon) on Canvas

Instructions:

- Submit your work as a single PDF. All work should be typed, with math and equations typeset using LaTeX or similar software. 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. This may require more work than a normal homework assignment, so I recommend starting early. 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 X. In the logistic regression model,

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

.

However, logistic regression is not the only option for modeling $P(Y_i = 1|X_i)$. 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 $X \in \mathbb{R}^k$ be a vector of covariates. LDA assumes that, conditional on Y, X follows a multivariate normal distribution. That is,

$$X|(Y = 0) \sim N(\mu_0, \Sigma) \qquad \text{and} \qquad X|(Y = 1) \sim N(\mu_1, \Sigma), \tag{1}$$

where $\mu_0, \mu_1 \in \mathbb{R}^k$ are the means, and $\Sigma \in \mathbb{R}^{k\times k}$ 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, $P(Y = 1|X)$ is given by

$$P(Y = 1|X) = \frac{\pi_1 \exp\{-\frac{1}{2}(X - \mu_1)^T\Sigma^{-1}(X - \mu_1)\}}{\pi_1 \exp\{-\frac{1}{2}(X - \mu_1)^T\Sigma^{-1}(X - \mu_1)\} + \pi_0 \exp\{-\frac{1}{2}(X - \mu_0)^T\Sigma^{-1}(X - \mu_0)\}}$$

2. We fit the LDA model estimating π_1, μ_0, μ_1, and Σ. In this question, we will fit the model to the `dengue` data from class. Let Y_i be a patient's dengue status, and let $X_i = (WBC_i, PLT_i)$ be a patient's white blood cell count and platelet count. Fit the LDA model (1) to this dengue data, and report the estimates $\widehat{\pi}_1$, $\widehat{\mu}_0$, $\widehat{\mu}_1$, and $\widehat{\Sigma}$.

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

4. In R, make a plot showing the relationship between the predicted probabilities $\widehat{P}(Y_i = 1|X_i)$ from logistic regression, and the predicted probabilities $\widehat{P}(Y_i = 1|X_i)$ 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

$$\log\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0) + (\mu_1-\mu_0)^T\Sigma^{-1}X.$$

   Conclude that if the LDA assumptions hold, then the log-odds are a linear function of the covariates 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 $\widehat{\pi}_1$, $\widehat{\mu}_0$, $\widehat{\mu}_1$, and $\widehat{\Sigma}$, 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.

6. Suppose that $X \in \mathbb{R}^2$, and (1) holds, with $\pi_1 = 0.5$, $\mu_0 = (0,0)^T$, $\mu_1 = (0.5, 0.5)^T$, and $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. In R, generate 100 samples $(X_1, Y_1), ..., (X_{100}, Y_{100})$ from the LDA model. **Hint:** sample Y first, then sample $X|Y$ from the appropriate multivariate normal distribution.

7. Using your sample from question 6, fit LDA and logistic regression models. Make a plot showing (a) the true probabilities $P(Y_i = 1|X_i)$ for each point (using the true parameters π_1, μ_0, μ_1, and Σ), (b) the estimated probabilities $\widehat{P}(Y_i = 1|X_i)$ using the fitted LDA model, and (c) the estimated probabilities $\widehat{P}(Y_i = 1|X_i)$ from the fitted logistic regression model. Which method – LDA or logistic regression – gives estimated probabilities $\widehat{P}(Y_i = 1|X_i)$ which are closer (on average) to the true probabilities $P(Y_i = 1|X_i)$?

8. Repeat questions 6 and 7 200 times. You do not need to make a plot for each repetition, just calculate summary statistics comparing the estimated probabilities to the true probabilities for LDA and logistic regression. If the LDA assumptions hold, which method does better – LDA or logistic regression?