CSE 5160 Machine Learning **Giovanni Almaraz, Doug Fry, Zachary Hampton**
Dr. Haiyan Qiao **02-03-2025**

Group Assignment 3

Part A (Textbook Chapter 4.8 Exercises: Q1, Q6, Q8)

Problem 1. ***Problem 1:*** *Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and the logit representation for the logistic regression model are equivalent.*

Answer. Solution:
Let

$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}.$$

We want to show this is equivalent to

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

Proof:

$$p(X) = \frac{1}{1 + e^{-z}}, \quad \text{where } z = \beta_0 + \beta_1 X.$$

Then

$$1 - p(X) = 1 - \frac{1}{1 + e^{-z}} = \frac{1 + e^{-z} - 1}{1 + e^{-z}} = \frac{e^{-z}}{1 + e^{-z}}.$$

Hence,

$$\frac{p(X)}{1 - p(X)} = \frac{\frac{1}{1+e^{-z}}}{\frac{e^{-z}}{1+e^{-z}}} = \frac{1}{e^{-z}} = e^{z}.$$

Taking the natural logarithm on both sides,

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \log(e^{z}) = z = \beta_0 + \beta_1 X.$$

Thus, (4.2) and (4.3) are indeed equivalent.

Problem 6. ***Problem 6:*** *Logistic Regression Probability Estimate*

Answer. Solution:

(a) Given that $\beta_0 = -6$, $\beta_1 = 0.05$, and $\beta_2 = 1$, we estimate the probability of getting an A for a student studying 40 hours with a GPA of 3.5 as follows:

$$P(Y = 1) = \frac{1}{1 + e^{-(-6+0.05 \cdot 40+1 \cdot 3.5)}} = \frac{1}{1 + e^{-(-6+2+3.5)}} = \frac{1}{1 + e^{-0.5}} \approx 0.378 \text{ (37.8\%)}.$$

(b) To find the number of hours a student with a GPA of 3.5 needs to study to have a 50% chance of getting an A, we set $P(Y = 1) = 0.5$ and solve for X_1:

$$0.5 = \frac{1}{1 + e^{-(-6+0.05 \cdot X_1+1 \cdot 3.5)}}$$

Simplifying:

$$\begin{aligned}
0.5(1 + e^{-(-6+0.05 \cdot X_1+3.5)}) &= 1 \\
1 + e^{-(-6+0.05 \cdot X_1+3.5)} &= 2 \\
e^{-(-6+0.05 \cdot X_1+3.5)} &= 1 \\
-(-6 + 0.05 \cdot X_1 + 3.5) &= 0 \\
6 - 0.05 \cdot X_1 - 3.5 &= 0 \\
2.5 - 0.05 \cdot X_1 &= 0 \\
-0.05 \cdot X_1 &= -2.5 \\
X_1 &= \frac{2.5}{0.05} = 50
\end{aligned}$$

Therefore, the student would need to study 50 hours to have a 50% chance of getting an A in the class.

Problem 8. ***Problem 8:*** *Comparison of Logistic Regression and K-Nearest Neighbors*

Answer. Solution:

We have two classification methods:

1. **Logistic Regression:**
   - Training error: 20%
   - Test error: 30%
2. **1-Nearest Neighbor (K=1):**
   - Average error: 18%

Even though KNN (K=1) has a lower average error, it is prone to overfitting and does not generalize well. Logistic regression, despite having a higher test error, is more stable and interpretable.

Thus, logistic regression is the better choice for classifying new observations in this case. However, using a better K value (e.g., K=5 or K=10) for KNN might improve its performance.

Part B (Stock Market Data: Logistic Regression & LDA)

Problem 1. ***Problem 1:*** *(a)–(d) Logistic Regression on the Stock Market Data*

Answer. Solution:

(a) Compute the testing error rate using all predictors Lag1, Lag2, Lag3, Lag4, Lag5.

(b) Identify which predictors can be removed to reduce the testing error (based on p-values or other criteria).

(c) Recompute the testing error after removing the less significant predictors.

(d) Given Lag1 $= 2.1$ and Lag2 $= -0.5$, calculate the predicted probability of the market going up.

Problem 2. ***Problem 2:*** *(a)–(c) LDA on the Stock Market Data*

Answer. Solution:

(a) Calculate $\Pr(Y = \text{UP})$ and $\Pr(Y = \text{DOWN})$ based on the training set.

(b) Compute the mean vector of $\mathbf{X}$ (the predictors) for each class (UP vs. DOWN).

(c) Discuss whether using a 70% posterior probability threshold ($\Pr(Y = \text{UP}|\mathbf{X} = x) \geq 0.70$) is feasible or advisable for predicting a market increase.