**BIOST 546**
**WINTER QUARTER 2021**

**Homework # 2**
**Due Via Online Submission to Canvas:**
**Monday, February 8 at 10 AM**

*Instructions:* You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

In this assignment you will perform binary classification on the Breast Cancer Wisconsin (Diagnostic) Data Set in the csv file `wdbc.data`. The dataset describes characteristics of the cell nuclei present in $n$ (sample size) images. Each image has multiple attributes, which are described in detail in `wdbc.names`. Here, we focus on the attributes in columns $\{2, 3, 4\}$, i.e.

- 2: Diagnosis (M = malignant, B = benign)
- 3: Average radius of the cell nuclei in each image
- 4: Average texture of the cell nuclei in each image

Specifically, our aim will be predicting a categorical variable $Y$ (Diagnosis – column 2), from the quantitative attributes $X_1$ (Average radius – column 3) and $X_2$ (Average texture – column 4).

1. Data exploration + Logistic Regression
   (a) Describe the data: sample size $n$, number of predictors $p$, and number of observations in each class.
   (b) Divide the data into a training set of 400 observations, and a test set; from now on, unless specified, work only on the training set.

(c) Make a scatterplot displaying $Y$ (color or shape encoded) and the predictors $X_1, X_2$ (on the $x$- and $y$-axis). Based on this scatterplot, do you think it will be possible to accurately predict the outcome from the predictors? Motivate your answer.

(d) Fit a logistic regression model to predict $Y$ and make a table, like Table 4.3 in the textbook, displaying the coefficient estimates, standard errors, and p-values (use command `summary`). Give an interpretation of the values of the coefficient estimates.

(e) Use the coefficient estimates to *manually* calculate the predicted probability $P(Y = \texttt{M} \mid (X_1, X_2) = (10, 12))$ writing explicitly every step. Compare your result with the prediction computed with `predict`.

(f) Suppose instead that for a given observation $x = (x_1, x_2)^T$, the estimated log-odds equals 0.7. What is $P(Y = \texttt{M} \mid X = x)$?

(g) Use the glm previously fitted and the Bayes rule to compute the predicted outcome $\hat{Y}$ from the associated probability estimates (computed with `predict`) both on the training and the test set. Then compute the confusion table and prediction accuracy (rate of correctly classified observations) both on the training and test set. Comment on the results.

(h) Plot an image of the decision boundary (like the one in Figure 2.13 in the textbook, but without the purple dashed line) as follows:

- Generate a dense set (e.g. 10000 observations) of possible values for the predictors $(X_1, X_2)$ within reasonable ranges; (the command `expand.grid` might come in handy)
- Use the glm model previously fitted to predict the outcome probabilities for every observation you have generated and use Bayes rule to compute the predicted outcomes;
- Plot predicted outcomes and associated predictors in a scatter plot together with the training set.

Generate the same plot for probability cutoff values of 0.25 and 0.75. Comment on the results.

(i) Plot the ROC curve, computed on the test set, for a dense grid of possible cutoffs (e.g. 20 intervals).

(j) Compute an estimate of the Area under the ROC Curve (AUC).

2. Linear discriminant analysis model

(a) Now fit a linear discriminant analysis model to the training set you created in Exercise 1. Make a table displaying the estimated 'Prior probabilities of groups' and 'Group means'. Describe in words the meaning of these estimates and how they are related to the posterior probabilities.

(b) Use the fitted model and Bayes rule to compute the predicted outcome $\hat{Y}$ from the predicted posterior probabilities, both on the training and test

set. Then, compute the confusion table and prediction accuracy both on the training and test set. Comment on the results.

(c) Plot an image of the decision boundary (follow the instructions in 1(h)). Generate the same plot for cutoff values of 0.25 and 0.75. Comment on the results.

(d) Plot the ROC curve, computed on the test set, for a dense grid of possible cutoffs (e.g. 20 intervals).

(e) Compute an estimate of the Area under the ROC Curve (AUC).

3. Repeat Exercise 2 with a quadratic discriminant analysis model.

4. Now we decide to use a kNN classifier.

(a) For all choices of $k = \{1, 2, 3, 4, 20\}$ (number of neighbors), compute the predicted outcome $\hat{Y}$ both on the training and test set. Then, compute the confusion table and prediction accuracy both on the training and test set. Comment on the results.

(b) Plot an image of the decision boundary (follow the instructions in 1(h)), for $k = \{1, 2, 3, 4, 20\}$ (number of neighbors). Comment on the results.

(c) Compute and plot the prediction accuracy (both on the training and test set), for $k = \{1, 2, \ldots, 19, 20\}$ (number of neighbors). Which value of $k$ would you choose? Comment on the results.

5. Discuss the performances of the 4 models you have applied in terms of prediction accuracy, on both the training and test set. If we were to pick the model with the highest prediction accuracy on the test set, discuss why kNN (with the optimally chosen $k$) has an unfair advantage in producing the highest estimated accuracy on the test set and how we could mitigate this issue.