Problem 1

Recall that researchers in Canada enrolled 75 patients in a study to see if vaccination for COVID-19 helped alleviate symptoms of long COVID, also known as post-COVID19 conditions or PCC[1]. Of the 75 subjects, 36 were vaccinated, and 39 remained unvaccinated. After a period of two months, the subjects returned and took a survey reporting whether their well-being had improved, worsened, or remained unchanged and whether they had more, less, or the same number of PCC symptoms.

(a) What kinds of variables are `WBscore` and `PCCsymp`?

(b) Treating `WBscore` and `PCCsymp` as categorical variables, train a logistic regression model using the "PCC_study_train.csv" file. Which variables are significant?

(c) Treating `WBscore` and `PCCsymp` as numeric variables, train a logistic regression model using the "PCC_study_train.csv" file. Which factors are significant?

(d) Apply both of your logistic regression models to the "PCC_study_test.csv" file and plot the ROC curves. What are the AUCs of the logistic regression models? Which would you recommend?

Problem 2

(a) With the same long covid symptom alleviation study data, train a decision tree for predicting vaccination status using the "PCC_study_train.csv" file (you may want to play around with various settings, such as the decision criterion). Create a visualization of the tree. Predict the probabilities of each subject being vaccinated using your tree then create the ROC curve.

(b) In Homework 2, you trained a Naive Bayes classifier using "PCC_study_train.csv" and calculated the AUC using "PCC_study_test.csv". This classifier assumed that the data were numeric values, (i.e., `WBscore_Worsened` to `0`, `WCscore_Unchanged` to `1`, `WBscore_Improved` to `2`, and similarly for `PCC_symp`). Predict the probabilities of each subject being vaccinated using `CategoricalNB` and create the ROC curve.

(c) Find the test set AUC for the Naive Bayes Classifier from Homework 2, the logistic regression model from Problem 1, and the decision tree from part (a). Use the AUCs to justify which model is best for classification.

Problem 3

Suppose you have a *single observation* y_i, and the predicted probability of the observation occurring is $\hat{pi}_i$.

(a) Write the predicted log odds of y_i occurring in terms of $\hat{p}_i$.

(b) Write $\hat{p}_i$ in terms of the predicted log odds.

(c) Write $1 - \hat{p}_i$ in terms of the predicted log odds.

(d) Confirm that the log likelihood function used in logistic regression, in terms of the log odds, is

[1] Nayyerabadi, Maryam, Lyvia Fourcade, Swarali A. Joshi, Prabha Chandrasekaran, Arpita Chakravarti, Chantal Massé, Marie-Lorna Paul et al. "Vaccination after developing long COVID: Impact on clinical presentation, viral persistence, and immune responses." International Journal of Infectious Diseases 136 (2023): 136-145.

$$\log \mathcal{L} = y_i \left(\beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}\right) - log(1 + e^{\beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}})$$

Use parts (b) and (c), and remember that since we only have one observation, there is no sum. You may also want to review rules for logs and exponents.

(e) Given that the cross entropy loss function is the negative of the log likelihood function used in logistic regression, show that the derivative of the cross entropy loss function with respect to the log odds is the negative residual. Use your equation from part (d), and carry out the calculation with the predicted log odds and replace them with predicted probabilities only after you have finished differentiating.

(f) Now, take the second derivative of the log likelihood function with respect to the predicted log odds. What is the result in terms of the predicted probability (i.e., replace the predicted log odds with predicted probability after you have finished differentiating). Recall that the Quotient Rule says that for a function $h(x) = \frac{f(x)}{g(x)}$, the derivative $h'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$, and simplify the expression as much as you can.

Problem 4

Consider the function

$$f(x, y) = 2x^4 + y^4 - x^2 - 3y^2 + 0.56$$

(a) Calculate the gradient of $f(x, y)$, $\nabla_{xy} f(x, y)$.

(b) Write a function in Python that, given initial values and a constant step size, will calculate the first K steps of the gradient descent algorithm for minimizing $f(x)$.

(c) Test your function from part (b) with initial values $(2, 4)$ and a step size of 0.01. Where is the minimum after 15 steps?

(d) Test your function from part (b) with initial values $(2, 4)$ and a step size of 0.02. Where is the minimum after 15 steps?

(e) Test your function from part (b) with initial values $(-1, -1)$ and a step size of 0.01. Where is the minimum after 15 steps?

(f) Using either Python or Wolfram Alpha, create surface and contour plots for $f(x, y)$. Interpret your answers to c, d, and e in light of your plots.