STATS 503, Homework 3

Please use a Jupyter notebook to write up your solutions. Submit your work through Canvas by uploading an html file that contains the html export of your notebook file. The html file should include all code (no cells hidden). Use markdown formatting to make clear which part of your notebook corresponds to each problem.

1. **[3 POINTS]** ISLP chapter 4, conceptual exercise #1
2. **[5 POINTS]** ISLP chapter 4, conceptual exercise #4
3. **[3 POINTS]** ISLP chapter 4, conceptual exercise #9
4. **[4 POINTS]** ISLP chapter 4, conceptual exercise #12, parts a, b, and c
5. **[3 POINTS]** Suppose we fit logistic regression to predict the probability that a STATS 503 student gets an A in the class, from two variables. The variables are average hours of study per week (X_1) and GPA in other statistics courses taken (X_2). The model estimates $\beta_0 = -4, \beta_1 = 0.05, \beta_2 = 1$. (Note: These are made-up numbers; do not try to predict your own grades with them.)
   (a) Predict the probability of getting an A for a student who studies 5 hours a week and has a GPA of 3.5 in other statistics courses.
   (b) What are the odds that this student will get an A?
   (c) How many hours a week does this student need to study for the model to predict a 50% chance of getting an A?
6. Download the `college_train` and `college_test` datasets from Canvas and load them as dataframes. Fit a `sklearn.linear_model.LogisticRegression` model to predict the `Private` variable from all other variables except for `Name`, using the `college_train` dataset. Let $\hat{p}(y \mid x)$ denote the fitted model.
   (a) **[3 POINTS]** For each sample in the `college_test` dataset, use the fit model to predict the probability that `Private` is `Yes` and use the fit model to predict the probability that `Private` is `No`. Use these probabilities to calculate the negative log likelihood of the testing datase for the model:
   $$-\sum_{i=1}^{n} \log \hat{p}(y_i \mid x_i),$$
   where n is the number of samples in `college_test` and each (x_i, y_i) pair is a sample from that dataset.
   (b) **[2 POINTS]** Now use $\hat{p}$ to create a hard classifier, also known as a decision rule. Specifically, consider the rule
   $$\hat{y}(x) = \begin{cases} \texttt{Yes} & \text{if } \hat{p}(\texttt{Yes} \mid x) > 0.5 \\ \texttt{No} & \text{otherwise.} \end{cases}$$

Using the test data, compute the false positive rate (FPR), true positive rate (TPR), false negative rate (FNR), and true negative rate (TNR) made by this decision rule.

(c) **[2 POINTS]** We will now use $\hat{p}$ to create a different hard classifier.

$$\hat{y}(x) = \begin{cases} \texttt{Yes} & \text{if } \hat{p}(\texttt{Yes} \mid x) > 0.9 \\ \texttt{No} & \text{otherwise.} \end{cases}$$

Using the test data, compute the FPR, TPR, FNR, and TNR for this new decision rule.

(d) **[3 POINTS]** We will now use $\hat{p}$ to create a family of different hard classifiers.

$$\hat{y}_t(x) = \begin{cases} \texttt{Yes} & \text{if } \hat{p}(\texttt{Yes} \mid x) > t \\ \texttt{No} & \text{otherwise.} \end{cases}$$

Using the test data, for each value of $t \in \{0.0, 1/100, 2/100, \ldots, 99/100, 1.0\}$, compute the FPR and TPR of the decision rule $\hat{y}_t$. Plot your results as an ROC curve.

Note: It is possible to plot the ROC curve with sci-kit's `roc_curve` *function, but for this homework we want you to do it "by hand," using the steps outlined above.*

(e) **[2 POINTS]** Use $\hat{p}$, the test data, and `sklearn.metrics.roc_auc_score` to compute the area under your ROC curve. The first argument should be the true labels from the testing dataset. The second argument should be the probability that `Private` is Yes, as predicted by your model.

Total points for this assignment: 30