Homework 3

E6690: Statistical Learning for Bio & Info Systems

P1. Suppose we collect data for a group of students in a statistics class with variables X_1 (hours studied), X_2 (undergrad GPA), and Y (receive an A). We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

(a) (5pt) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

(b) (5pt) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

P2. (10pt) Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Hint: Use Bayes' theorem.

P3. (20pt) Consider $X = [0.0\ 0.2\ 0.4\ 0.6\ 0.8\ 1.0]^\top$ as the independent variable and $y = [\text{false false false true false true}]^\top$ as the response. Write down the log-likelihood function, $l(\beta_0, \beta_1)$, for the logistic regression problem and first order optimality conditions. Note that this problem of finding optimal (β_0, β_1) can only be solved numerically. Use NewtonRaphson algorithm from Section 4.4.1, pp. 120-121, [ESL] book and perform 10 iterations. Hint: Use library(matlib) for calculating matrix inverses.

P4. (20pt) Let $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Find the distribution of $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$. For the bivariate case with

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

find a 2×2 $\boldsymbol{A}$ such that cov$(\boldsymbol{Y})$ is an identity matrix.

Hint: Consider eigen-decomposition.

P5. (20pt) Consider K different populations when the mean of each population may be different, but one may assume that the variance of each population is the same (σ^2). Let

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:\, y_i = k} x_i \quad \text{and} \quad \hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2,$$

n_k is the sample size of population k and the pair (x_i, y_i) corresponds to the ith observation; $y_i \in \{1, \ldots, K\}$. For $\alpha_1, \ldots, \alpha_K$, such that $\sum_{i=1}^K \alpha_i = 1$, define an unbiased pooled variance estimator:

$$\hat{\sigma}^2 = \sum_{k=1}^{K} \alpha_k \hat{\sigma}_k^2.$$

If $n = \sum_{k=1}^{K} n_k$, show that $\alpha_k = (n_k - 1)/(n - K)$ minimizes the variance of $\hat{\sigma}^2$ under the Gaussian assumption.

P6. (10pt) Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $\mathbb{P}[\text{Class is Red}|X]$:

$$0.1,\ 0.15,\ 0.2,\ 0.2,\ 0.55,\ 0.6,\ 0.6,\ 0.65,\ 0.7,\ \text{and } 0.75.$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach. The second approach is to classify based on the average probability. In this problem, what is the final classification under each of these two approaches?

P7. This problem involves `OJ` data set which is part of the `ISLR` package.

(a) (2pt) First run set.seed(1000), and then create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

(b) (2pt) Fit a tree to the training data, with `Purchase` as the response and the other variables as predictors. Use the `summary()` function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?

(c) (3pt) Type in the name of the tree object in order to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed.

(d) (2pt) Create a plot of the tree, and interpret the results.

(e) (3pt) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?

(f) (2pt) Apply the `cv.tree()` function to the training set in order to determine the optimal tree size.

(g) (3pt) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.

(h) (1pt) Which tree size corresponds to the lowest cross-validated classification error rate?

(i) (3pt) Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.

(j) (2pt) Compare the training error rate between the pruned and unpruned tree. Which is higher?

(k) (2pt) Compare the test error rates between the pruned and unpruned trees. Which is higher?