STAT 503 Homework 1

Due Friday, Jan 28, 2022

Part 1. Suppose there are two classes in a population and the prior for the first class is $\pi_1 = 1/3$.

1. Suppose there is just a single predictor available, which in class k ($k = 1$ or 2) follows the exponential distribution with parameter λ_k. Derive the Bayes rule and compute the Bayes risk. Compute the numeric values for the class boundaries and the risk when $\lambda_1 = 1$, $\lambda_2 = 2$.

2. Repeat part 1 if, instead of a single predictor, there are two independent predictors available, and both follow the exponential distribution with parameter λ_k in class k. Derive the Bayes rule and compute the Bayes risk. Compute the numeric values for the class boundaries and the risk when $\lambda_1 = 1$, $\lambda_2 = 2$.

Reminder: the probability density function of an exponential random variable with parameter $\lambda > 0$ is $p(x) = \lambda e^{-\lambda x}$, for $x > 0$.

Part 2. The dataset `College` available within the ISL book R package `ISLR2` contains data on 18 variables from 777 US colleges collected in 1995, such as number of applications, enrollment, estimated costs, etc. See the dataset description for full information. The task is to predict whether the school is public or private (variable `Private`) from the other 17 variables.

1. Perform exploratory data analysis of this dataset. Make sure you have at least one figure with side-by-side boxplots, at least one panel plot, and at least one scatterplot. Focus on illustrating the relationship between the class and the other variables, but also explore pairwise relationships between predictors and include interesting findings. Comment on each plot you include (no points given for plots not accompanied by explanation / interpretation). This part should not exceed 6 pages total.

2. Set aside 20% of the data, selected at random, to use as the test set. Apply LDA, QDA, and logistic regression. Report the training and the test errors for each classifier, both for each class separately and overall. Comment on the results.

3. Make a histogram of the test data projected onto the first LDA direction learned from the training data, using different colors or symbols to compare true class labels with predicted class labels (one for each classifier). Comment on any interesting features and on how the classifiers compare.

4. Plot the test data against two predictor variables of your choosing, using different colors or symbols to compare true class labels with predicted class labels (one for each classifier). Comment on any interesting features and on how the classifiers compare.

5. (Optional extra credit) Add class boundaries to each plot in part 4.