Stats 202 Practice Midterm Solutions

Jelena Markovic

October 24, 2018

1. (a) See page 395 in the book.

   (b) Yes, the clustering will be the same.

   At each step of an agglomerative clustering algorithm, we join the two clusters that are closest together. Suppose at some level in the dendrogram, the clusters are the same under d and d'. Let A and B be two clusters, and (a, b) be the pair of samples that are closest together under d, with $a \in A$ and $b \in B$. Since d' is a monotone transformation of d, the pair of points in A and B that are closest together under d' will also be (a, b). The single-linkage distance between clusters A and B is then $d(a, b)$ in the first case, and $d'(a, b)$ in the second case.

   Now, suppose that clusters A and B are the two clusters that are closes under d. By monotonicity again, A and B will be the most proximal clusters under d'. This implies that the next pair of clusters to be joined in the dendrogram is the same under both distances. By induction, the two dendrograms have the same structure, and the clustering with k clusters will be identical.

2. (a) Since the decision boundary seems very non-linear and there are only 2 predictors, I would use a k-nearest neighbors algorithm.

   (b) The decision boundary seems linear or close to linear, so I would use logistic regression (or LDA).

3. The one-standard error rule states we should choose the simplest model whose error lies within a standard error of the minimum error. The minimum error in the plot above is achieved at $k = 9$. The flexibility or variance of k-nearest neighbors decreases with k, so we would have to choose a model with $k = 9$. The model with $k = 10$ is the only model whose error lies within a standard error of the minimum error, so we would pick $k = 10$.

4. (a) See page 105 in the book describing KNN regression.

   (b) Cross-validation. For a range of K values you run estimate test MSE error via cross-validation, and pick K that minimizes this estimated test MSE. Instead of picking K that is the exact minimizer of Cross-Validation error, you can also use one standard error rule (described on page 214 in the book or Problem 3 of this practice exam).

5. (a) False. We might not have a better test error with QDA since QDA might be more flexible than necessary for this dataset. You can look up the first three scenarios in the book on page 153 for the examples where the decision boundary is linear and LDA outperforms QDA on all three of them.

   (b) False. When the boundaries are non-linear, we expect QDA to give better test error than LDA. However, this might not necessarily be true as you can have situations where methods fitting linear boundary outperform methods fitting quadratic boundary even when the true boundary is highly non-linear.

   For a related example see Figure 3.19 in the book, showing the linear regression can outperform KNN even when the true relationship is highly non-linear.