Problem 1

Two distances, d and d', are related by a monotone transformation:

$$d'(a,b) = f(d(a,b))$$

which satisfies $f(x) \geq f(y)$ if $x \geq y$.

Please refer to book ISLR page 390-396 for the description of hierarchical clustering with single linkage.
Yes we will get the same clustering result.

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.

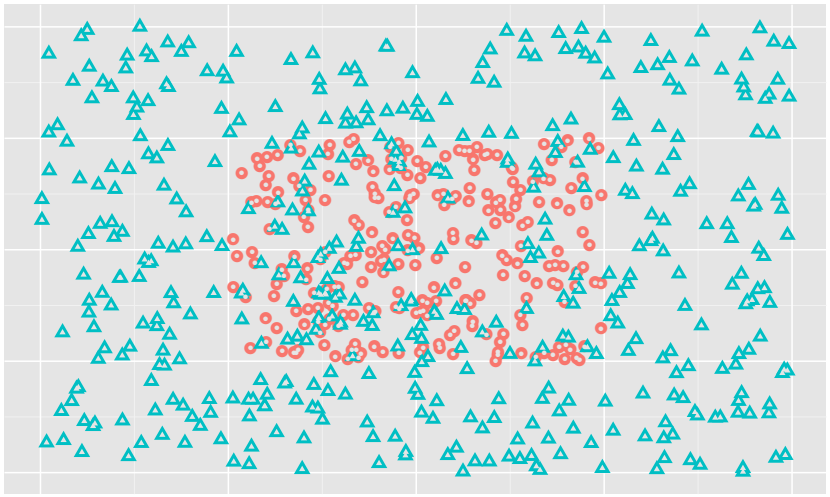
Now, suppose that A^* and B^* are the two clusters that are closest together under d. By monotonocity again, A^* and B^* will be the most proximal clusters under d'. This implies that the next pair of clusters to be joined in the dendrogram is the same under both distances. By induction, the two dendrograms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

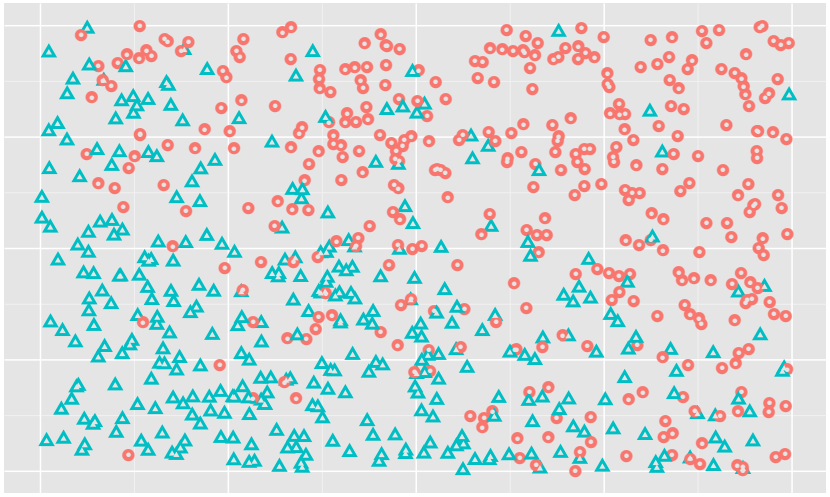
Problem 2

The figure below depicts two different two-class classification problems. Call the top figure A, the bottom figure B.

For A do you think logistic regression or K-nearest neighbors would be a better classifier? Explain.

For B do you think logistic regression or K-nearest neighbors would be a better classifier? Explain.





Top: Since the decision boundary seems very non-linear and there are only 2 predictors, I would use a k-nearest neighbors algorithm. The k-nearest neighbors algorithm classifies to the positive class if the estimated conditional probability

$$\hat{P}(Y = +|X = x) = \frac{1}{n} \sum_{i \in N_k(x)} \mathbf{1}(y_i = +)$$

is greater than 0.5.

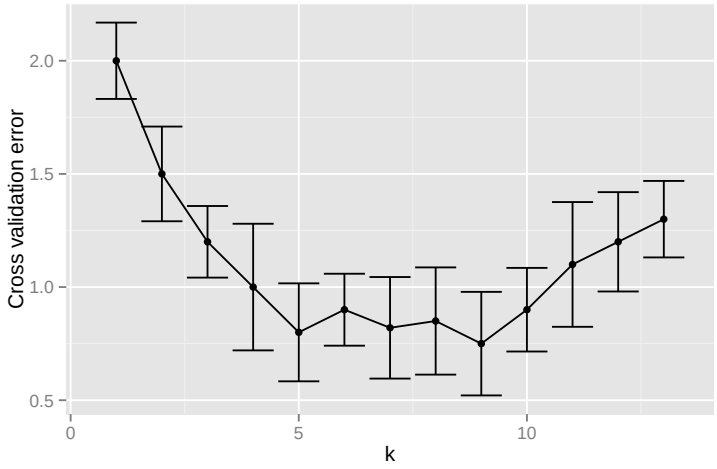
Bottom: The decision boundary seems linear or close to linear, so I would use LDA or logistic regression. Logistic regression assigns to positive if the estimated conditional probability

$$\hat{P}(Y = +|X = x) = \frac{e^{X \cdot \hat{\beta}}}{1 + e^{X \cdot \hat{\beta}}}$$

is greater than 0.5.

Problem 3

State and explain the one standard error rule for model selection using 10-fold cross validation. Apply it to select the optimal number of nearest neighbors in the plot below, which shows the cross-validation error and one standard error intervals as a function of k.



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 $k \geq 9$. The model with $k = 10$ is the only model whose error lies within a standard error of the minimum error, so we would pick $k = 10$.

Problem 4

The Advertising data set in the book consists of the sales of a product in 200 markets, along with the advertising budget in each market for three different media: TV, radio and print. You want to use k-nearest neighbor regression to predict the sales as a function of the spending on advertising on these three media.

(a) For a fixed value of k, explain how k-nearest neighbor regression predicts the sales number, given the advertising spending on TV, radio and print.

(b) How would you choose k to get a good prediction? Name a method for doing this and briefly explain the method.

a For the given numbers (TV, radio, print), find the K nearest ~~neighbors~~ observations (using Euclidean distance in 3-D space (TV, radio, print)). Then compute the average of the K corresponding Y-values (i.e. sales numbers) as the predictor.

b Using cross-validation. For example, leave one out cross-validation predicts each of the 200 sales values in turn from its K nearest neighbors, as desribed in a. Then we estimate the test error by averaging the 200 squared ~~distan~~ differences between predicted sale and actual sale. Then we choose the K that minimizes this estimated test error.

Problem 5

True or False, and explain briefly:
(a) Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary.
(b) If the Bayes decision boundary for a given problem is nonlinear, then we will achieve a superior test error rate using QDA rather than LDA.

a: False: QDA is more flexible than LDA and hence has more variance, which can result in a larger test error rate (e.g. in high-dimensional space)

b False: While QDA may have a smaller bias than LDA, its variance is larger and it's not clear that the bias-variance trade off goes in its favor. E.g. in high dimensional space the variance penalty may be larger than the gain in bias.