Stats 202: Data Mining and Analysis

Fall 2018

Students will be allowed one double-sided or two single-sided pages of notes.

Problem 1

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

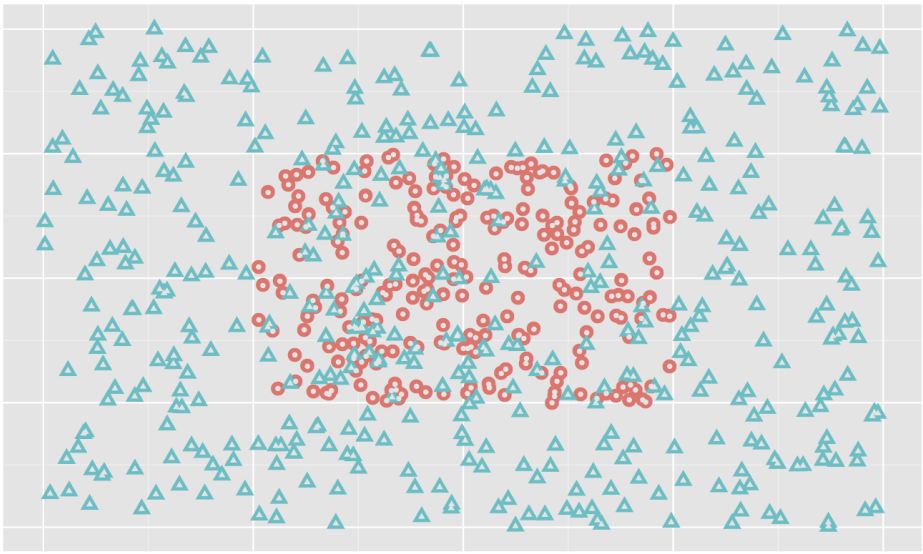
$$d'(a,b) = f(d(a,b))$$

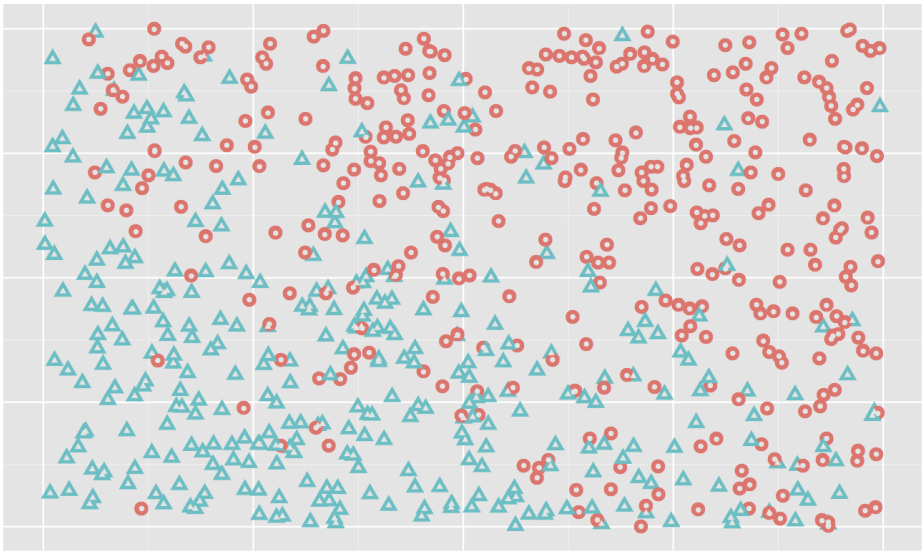
where f satisfies $f(x) \geq f(y)$ if $x \geq y$.

- Explain the method of single linkage hierarchical clustering.
- If you use distance d instead of d' will you get the same clustering? Explain.

Problem 2

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

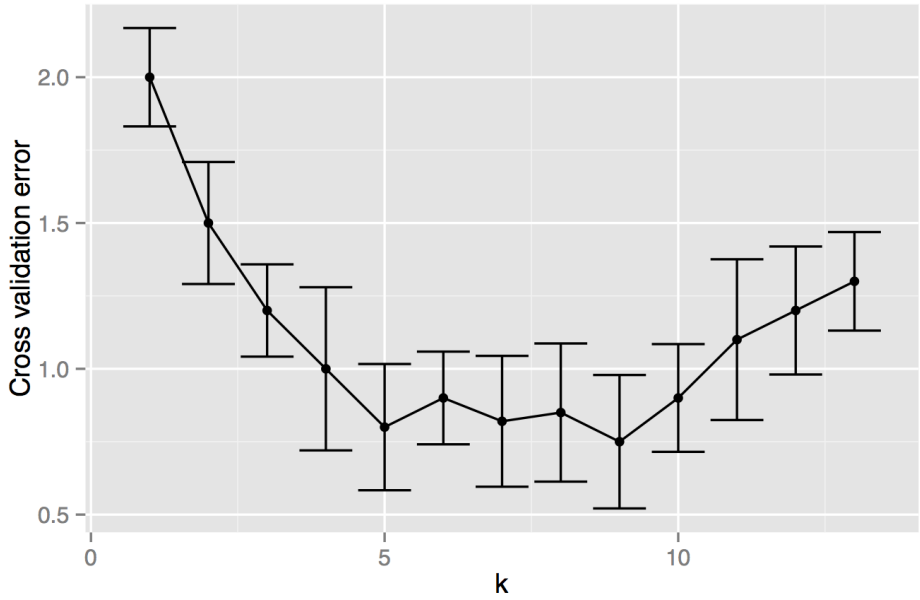




- For A do you think logistic regression or K-nearest neighbors would be a better classifier? Explain.
- For B do you think logistic regression or K-nearest neighbors would be a better classifier? Explain.

Problem 3

The figure below depicts cross-validation error in a regression setting with K-nearest neighbors.



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.

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.

- 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.
- How would you choose K to get a good prediction? Name a method for doing this and briefly explain the method.

Problem 5

True or False, and explain briefly:

- 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.
- 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.