STAT 730: Homework 6

1. Perform model-based clustering on the iris data, use all three species and all four measurements. Is the number of clusters picked via BIC the actual number of species in the data?

2. Perform hierarchical clustering using the sports ranking data from the multidimensional scaling chapter; obtain dedrograms for each of the four linkage methods. Favoring average linkage and Ward's, comment on the groupings; do the clusters "make sense?" How many groups would you pick based on the dendrograms?

3. MKB 11.2.1.

4. For the leisure-time data with all 20 activities, use linear discriminant analysis and logistic-regression with LASSO to discriminate between faculty and graduate students. What are the apparent (raw) error rates of each method? What are the cross-validated error rates? Extra-credit: which of the 20 activities are actually kept using LASSO with cross-validated $\hat{\lambda}$?

5. Fit a CART to the leisure-time data, again discriminating among graduate students and faculty, using AIC and display the tree. Which activities are used in the tree?