Stat 340 - Potential Quiz 3 Topics

Curse of Dimensionality

In the context of statistical/machine learning, the curse of dimensionality is the phenomenon that as the number p of features/explanatory variables gets very large, the number of training set observations falling in a neighborhood of a test point becomes very small. Or, turning that around, in order to ensure that a neighborhood around the test point includes a reasonable number of observations, the neighborhood must be very large if p is large. This is a problem for all methods, but is especially problematic for local methods like K nearest neighbors because there are not many nearby neighbors in the training data set to use for estimating the response at the test point.

You will not need to prove this or motivate it in any formal way, but you should understand and be able to repeat what's said in the paragraph above.

KNN Classification

Calculate the prediction from a KNN classifier by calculating the Euclidean distance between each training set observation and the test point and finding the proportion of the K nearest training set observations that are in each class; the predicted class is the most commonly-occuring class among those K nearest training set observations.

Logistic Regression

- Interpret coefficients from a logistic regression model in terms of odds. The problem set only featured this with a quantitative explanatory variable, but you should also know how to do this with a categorical variable - see Lab 07 problems 4 and 5 for examples with both quantitative and categorical explanatory variables. I will not ask you to interpret interaction terms in a logistic regression model (we did not discuss this in class).
- Conduct tests about one or more coefficients in a logistic regression model, using output from either `summary` or `anova`.

Train/Test

Explain how you would estimate test set error rate of a classification method using a train/test split:

- Randomly split the full data set into two parts: the training set and the test set
- Estimate any model parameters using the training set; the test set should not be involved in parameter estimation
- Make predictions for the test set
- Calculate the proportion of predictions for the test set that were incorrect

The result of this will be a number summarizing model performance for that particular test set; this is an estimate of performance on test sets more generally. We want a method with a low test set error rate.

Note that we hadn't yet talked about cross-validation as of the time problem set 07 was assigned, so cross-validation will not appear on this quiz.