Review for Final Exam

Rebecca C. Steorts, Duke University

STA 325

Agenda

- Final exam: November 30, 2017
- Final exam is cumulative
- Will be a major focus on topics after exam two

Quick review of topics

- Information retrieval
- Locality sensitive hashing
- Principle components analysis
- K-means clustering
- Hierarchical clustering
- How to specify the number of clusters
- Linear regression (will not be on the exam)
- Logistic regression
- Classification
- LDA and QDA
- Cross validation
- Bootstrapping
- Trees (will not be on the exam)

Question 1

What is the difference between supervised and unsupervised learning? What is one example of each topic from the course?

Question 1

What is the difference between supervised and unsupervised learning? What is one example of each topic from the course?

Unsupervised problems have no labeled data (the response variable y), where as supervised data problems have both the predictor and response variable (x, y).

K-means is unsupervised, while linear regression is supervised.

Question 2

What is re-sampling?

Question 2

A re-sampling method involves repeatedly drawing samples from a training data set and refitting a model to obtain addition information about that model.

Question 3

We want to estimate the test error associated with fitting a particular statistical learning method on a set of observations. Explain the simplest way of doing this (hint: validation approach).

Question 3

1. Randomly divide the available set of observations into two parts, a training set and a validation set or hold-out set.
2. Fit the model on the training set.
3. Use the resulting fitted model to predict the responses for the observations in the validation set.
4. The resulting validation set error rate is typically assessed using the MSE in the case of a quantitative response. This provides an estimate of the test error rate.

Question 4

What is one main drawback to the validation approach?

Question 4

There are two drawbacks to the validation approach:

1. The validation estimate of the test error rate can be highly variable, which depends on which observations are in the training and validations sets.
2. Only a subset of the observations (training set) are used to fit the model. The validation set error may tend to over-estimate the test error rate for the model fit on the entire data set.

Question 5

How can we easily fix the validation approach?

Question 5

We can either use LOOCV or k-fold CV. (You don't need to explain the methods unless asked).