Introduction to Machine Learning - Final Exam

Landon Buell

Spring 2020

Problem 1

Assume that you have a data set with a predictor (feature) X and the target Y. You run simple linear regression (Y X) and get the best fit with RSS=20 and TSS=120.

1. What is the covariance between X and Y if the variances of X and Y are Var(X)=15 and Var(Y)=20?

The *covariance* between variables X and Y is defined by:

$$Cov(X,Y) = E[XY] + E[X]E[Y] \tag{1}$$

2. Repeat if RSS=50 and TSS=40. Discuss the result.

Problem 2

Problem 3

1. Do you expect random forests to achieve a smaller or larger training error than bagged trees with the same number of trees? Justify your answer.

2. How would you expect the training error of bagged trees to compare with boosted trees? Justify your answer.

3. How would you decide how many trees to use for any particular data set in a random forest? It is best to choose the number that minimizes the training error? Why or why not?

Problem 4

This problem examines the differences between SVC (linear SVM), SVM with a polynomial kernel, and other linear classifiers.

1. For an arbitrary training set, would you expect for SVC (linear) or SVM (polynomial kernel) to work better on the training set? Why?

2. If the Bayes decision boundary between the two classes is linear, would you expect SVC or SVM to work better on the training set?

3. True or False: There is no need to use slack variables in SVMs with polynomial kernels because the decision boundary can be nonlinear. Justify your answer.

4. LDA, Logistic regression, and SVC (linear) all fit a linear decision boundary. Will their fits be the same? What would be some reasons for you to prefer LDA over SVC?

Problem 5