EEL709: Assignments Quiz

April 29, 2015

Solutions

1. Suppose you have trained a polynomial model $y(x) = w_0 + w_1 x + ... + w_M x^M$ via least-squares regression, and you find that it has a low training error, but a high testing error. Which of the following is likely to reduce the testing error?

   Increasing the number of training data points and re-training.
   Decreasing M and re-training.

2. When training a multi-variable regression model, which of these would be a reasonable way to gauge the importance of different features?

   Normalise each feature to have mean 0 and variance 1; then weights with higher absolute values will correspond to more important features.
   The greater the increase in cross-validation error on eliminating a feature, the more important it is.

3. Suppose you train two different neural networks (with differing parameters) on the same classification data set. The cross-entropy error is used for training. Which of the following are true? (Errors below refer to training errors.)

   The neural net with lower cross-entropy error could have higher classification error.

4. You are training an RBF SVM with the following parameters: C (slack penalty) and σ (spread of RBF kernel). How should you tweak the parameters to reduce overfitting?

   Reduce C, increase σ

5. In which of the below settings would it usually make sense to simply use a linear SVM, rather than choosing some non-linear kernel? (N refers to the number of training data points, D to the number of dimensions.)

   N << D
   N = D

6. Consider the following possible choices of error function in training a neural network for classification: cross-entropy error (I), classification error (II), and sum-of-squares error (III). Which of the following are true?

   (II) is problematic because it's non-differentiable; (I) is preferred to (III) because the former corresponds to maximising the likelihood of the data.

7. Consider the following statements with respect to reducing the number of features to be used in a supervised learning task. Which are true?

8. You run PCA on your 200-dimensional data set, and find that the top two principal components capture 90% of the variance. Based on this, you can conclude that:

   For clustering purposes, using the 2-D PCA space is likely to lead to little or no loss of relevant information, compared to the full feature space.
   There is a two-dimensional plane within your full feature space, such that pretty much all the data points are located close to that plane.

9. You run K-means on a labeled data set, then label each cluster with the most frequently occurring label within it and thus compute an accuracy figure for the clustering. You find that this accuracy is substantially lower than what a one-vs.-one logistic regression classifier could achieve on the same data. Which of the following are NOT valid reasons for this observation?

   The classification method is more powerful than the clustering method, in the sense that it can learn more complicated decision boundaries.