EEL709: Assignments Re-Quiz

May 12, 2015

Maximum Marks: 12

Instructions: Each question may have any number of correct answers, including zero. Clearly mark all answers which you think are correct. Each correct choice will earn one mark, whilst half a mark will be deducted for each incorrect choice. Unclear markings will be ignored.

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 high error on both training and testing data. Which of the following might reasonably reduce the testing error?

   (a) Decreasing the number of training data points and re-training.
   (b) Decreasing M and re-training.
   (c) Dividing each of the learnt weights by 2.
   (d) Increasing 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?

   (a) The greater the increase in testing error on eliminating a feature, the more important it is.
   (b) Normalise each feature to have mean 0 and variance 1; then weights with higher values will correspond to more important features.
   (c) Normalise each feature to have mean 0 and variance 1; then weights with lower values will correspond to more important features.
   (d) The greater the increase in training accuracy 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/accuracies below refer to training.)

   (a) The neural net with higher cross-entropy error is always expected to give lower classification accuracy.
   (b) The neural net with higher cross-entropy error might give higher classification accuracy.
   (c) Scenario (b) is possible only when the neural net with higher cross-entropy error has overfit the data.
   (d) Scenario (b) is possible only when the neural net with higher classification error has overfit the data.

4. You are training an RBF SVM with the following parameters: C (slack penalty) and σ (spread of RBF kernel). Suppose you have used a grid search to find optimal values for these parameters. Starting with this optimal setting, what will be the effect of increasing C?

   (a) Overfitting, if σ is kept unchanged.
   (b) Overfitting, as long as σ is not increased.
   (c) Underfitting, as long as σ is not decreased.
   (d) Overfitting, irrespective of any changes to σ.

5. In which of the below settings is an SVM with a non-linear kernel reasonably likely to work better than a linear SVM? (N refers to the number of training data points, D to the number of dimensions.)

   (a) N very large, irrespective of D.
   (b) N and D both similarly large.
   (c) D very small, compared to N.
   (d) N and D both similarly small.

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?

   (a) (I) and (II) are both likely to give similar results, because they penalise errors in similar ways.
   (b) (II) and (III) are both likely to give similar results, because they penalise errors in similar ways.
   (c) (III) is preferable to (II) because the former is differentiable.
   (d) (III) is preferable to (I) 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?

   (a) Reducing the number of features generally helps classifier performance, because more features means more noise.
   (b) For a fixed number of dimensions to be retained, getting those dimensions via PCA will in general work better than choosing them on the basis of a filter like correlation or entropy with respect to the class labels, because PCA retains the maximally informative dimensions.
   (c) Even if (b) is the case, the correlation and entropy filters would often be preferred because the original features are more interpretable than principal components.
   (d) Feature reduction in general can be helpful, despite loss of information, because it lessens the curse of dimensionality.
   (e) Feature reduction can only be helpful when there are redundant features, such that there is no loss of information on removing them.

8. You run PCA on your 200-dimensional data set, and find that the top two principal components capture 95% of the variance. Based on this, you can conclude that:

   (a) K-means run on the 2-D PCA space is likely to give similar results to K-means run on the full feature space.
   (b) Suppose the data consists of two classes. Training an SVM classifier in the 2-D PCA space is likely to lead to little or no loss of relevant information, compared to training in the full feature space.
   (c) The statement (a) would also be true for this data set if the two dimensions were obtained via a correlation or entropy filter with respect to the class labels.

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 valid reasons for this observation?

   (a) The clustering method was not able to construct an appropriate representation of the data, for the purposes of discriminating between the classes.
   (b) In real data, the number of natural clusters may not match the number of labeled classes.
   (c) In real data, classes may have substantial overlap, due to noise or other reasons.
   (d) The classification method can learn more complicated decision boundaries.