EEL709: Assignments Quiz

April 29, 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 low training error, but a high testing error. Which of the following is likely to reduce the testing error?

   (a) Increasing 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) Weights with higher values will correspond to more important features.
   (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 higher absolute values will correspond to more important features.
   (d) 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.)

   (a) The neural net with lower cross-entropy error will always have lower classification error as well.
   (b) The neural net with lower cross-entropy error could have higher classification error.
   (c) Scenario (b) is possible only when the neural net with lower cross-entropy error has overfit the data.
   (d) Scenario (b) is possible only when the neural net with lower 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). How should you tweak the parameters to reduce overfitting?

   (a) Increase C, reduce σ
   (b) Reduce C, increase σ
   (c) Increase both C and σ
   (d) Reduce C only (σ has no predictable effect on overfitting)

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.)

(a) $N << D$
(b) $N >> D$
(c) $N = D$
(d) None of the above

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) (II) is problematic because it's non-differentiable, but either of (I) or (III) should give the same result.
   (b) Any of the three could be easily used for backpropagation, but (I) is preferred because it corresponds to maximising the likelihood of the data.
   (c) (II) is problematic because it's non-differentiable; (I) is preferred to (III) because the former corresponds to maximising the likelihood of the data.
   (d) (II) is problematic because it's non-differentiable; (III) is preferred to (I) because the former corresponds to maximising the likelihood of the data.
   (e) Any of the three could be easily used for backpropagation, but (III) is preferred because it 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 leads to loss of information, hence can always be expected to lead to poorer performance.
   (b) For a fixed number of dimensions to be retained, getting those dimensions via PCA will in general lead to better performance than getting them from supervised feature selection, because PCA retains the maximally informative dimensions.
   (c) Whilst (b) is true, supervised feature selection is often preferred because the original features are more interpretable than principal components.
   (d) For a fixed number of dimensions to be retained, there is no predictable reason why either of PCA or supervised feature selection should lead to better performance; *a priori*, either approach is equally sensible.

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:

   (a) 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.
   (b) For supervised learning 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.
   (c) 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?

   (a) The clustering method wasn't trying to optimise the classification performance; it was optimising something else.
   (b) In real data, classes don't always correspond to tight clusters: sometimes, intra-class variation may be quite high.
   (c) In real data, classes aren't always well-separated: sometimes, inter-class separation may be quite low.
   (d) The classification method is more powerful than the clustering method, in the sense that it can learn more complicated decision boundaries.

EEL709: Assignments Quiz

April 29, 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. 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 cross-validation error on eliminating a feature, the more important it is.
   (b) Weights with higher values will correspond to more important features.
   (c) Normalise each feature to have mean 0 and variance 1; then weights with higher values will correspond to more important features.
   (d) Normalise each feature to have mean 0 and variance 1; then weights with higher absolute values will correspond to more important features.

2. 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?

   (a) The clustering method wasn't trying to optimise the classification performance; it was optimising something else.
   (b) The classification method is more powerful than the clustering method, in the sense that it can learn more complicated decision boundaries.
   (c) In real data, classes don't always correspond to tight clusters: sometimes, intra-class variation may be quite high.
   (d) In real data, classes aren't always well-separated: sometimes, inter-class separation may be quite low.

3. 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?

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

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?

   (a) Reduce C only (σ has no predictable effect on overfitting)
   (b) Increase C, reduce σ
   (c) Increase both C and σ
   (d) Reduce C, increase σ

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

   (a) For a fixed number of dimensions to be retained, getting those dimensions via PCA will in general lead to better performance than getting them from supervised feature selection, because PCA retains the maximally informative dimensions.
   (b) Reducing the number of features leads to loss of information, hence can always be expected to lead to poorer performance.
   (c) Whilst (a) is true, supervised feature selection is often preferred because the original features are more interpretable than principal components.
   (d) For a fixed number of dimensions to be retained, there is no predictable reason why either of PCA or supervised feature selection should lead to better performance; *a priori*, either approach is equally sensible.

6. 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.)

   (a) $N >> D$
   (b) $N = D$
   (c) $N << D$
   (d) None of the above

7. 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.)

   (a) The neural net with lower cross-entropy error could have higher classification error.
   (b) Scenario (a) is possible only when the neural net with lower cross-entropy error has overfit the data.
   (c) Scenario (a) is possible only when the neural net with lower classification error has overfit the data.
   (d) The neural net with lower cross-entropy error will always have lower classification error as well.

8. 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) Any of the three could be easily used for backpropagation, but (I) is preferred because it corresponds to maximising the likelihood of the data.
   (b) (II) is problematic because it's non-differentiable; (III) is preferred to (I) because the former corresponds to maximising the likelihood of the data.
   (c) (II) is problematic because it's non-differentiable; (I) is preferred to (III) because the former corresponds to maximising the likelihood of the data.
   (d) (II) is problematic because it's non-differentiable, but either of (I) or (III) should give the same result.
   (e) Any of the three could be easily used for backpropagation, but (III) is preferred because it corresponds to maximising the likelihood of the data.

9. 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:

   (a) 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.
   (b) 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.
   (c) For supervised learning 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.

EEL709: Assignments Quiz

April 29, 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. 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.)

   (a) $N >> D$
   (b) $N << D$
   (c) $N = D$
   (d) None of the above

2. 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?

   (a) The classification method is more powerful than the clustering method, in the sense that it can learn more complicated decision boundaries.
   (b) In real data, classes don't always correspond to tight clusters: sometimes, intra-class variation may be quite high.
   (c) The clustering method wasn't trying to optimise the classification performance; it was optimising something else.
   (d) In real data, classes aren't always well-separated: sometimes, inter-class separation may be quite low.

3. 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?

   (a) Reduce C only (σ has no predictable effect on overfitting)
   (b) Increase both C and σ
   (c) Reduce C, increase σ
   (d) Increase C, reduce σ

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

   (a) For a fixed number of dimensions to be retained, getting those dimensions via PCA will in general lead to better performance than getting them from supervised feature selection, because PCA retains the maximally informative dimensions.
   (b) Whilst (a) is true, supervised feature selection is often preferred because the original features are more interpretable than principal components.
   (c) Reducing the number of features leads to loss of information, hence can always be expected to lead to poorer performance.

(d) For a fixed number of dimensions to be retained, there is no predictable reason why either of PCA or supervised feature selection should lead to better performance; *a priori*, either approach is equally sensible.

5. 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.)

   (a) The neural net with lower cross-entropy error could have higher classification error.
   (b) The neural net with lower cross-entropy error will always have lower classification error as well.
   (c) Scenario (a) is possible only when the neural net with lower cross-entropy error has overfit the data.
   (d) Scenario (a) is possible only when the neural net with lower classification error has overfit the data.

6. 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:

   (a) 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.
   (b) For supervised learning 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.
   (c) 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.

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

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

8. 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) Any of the three could be easily used for backpropagation, but (I) is preferred because it corresponds to maximising the likelihood of the data.
   (b) (II) is problematic because it's non-differentiable; (I) is preferred to (III) because the former corresponds to maximising the likelihood of the data.
   (c) (II) is problematic because it's non-differentiable, but either of (I) or (III) should give the same result.
   (d) Any of the three could be easily used for backpropagation, but (III) is preferred because it corresponds to maximising the likelihood of the data.
   (e) (II) is problematic because it's non-differentiable; (III) is preferred to (I) because the former corresponds to maximising the likelihood of the data.

9. 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?

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

EEL709: Assignments Quiz

April 29, 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. 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:

   (a) For supervised learning 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.
   (b) 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.
   (c) 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.

2. 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?

   (a) In real data, classes don't always correspond to tight clusters: sometimes, intra-class variation may be quite high.
   (b) The clustering method wasn't trying to optimise the classification performance; it was optimising something else.
   (c) The classification method is more powerful than the clustering method, in the sense that it can learn more complicated decision boundaries.
   (d) In real data, classes aren't always well-separated: sometimes, inter-class separation may be quite low.

3. 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?

   (a) Increase C, reduce σ
   (b) Reduce C only (σ has no predictable effect on overfitting)
   (c) Increase both C and σ
   (d) Reduce C, increase σ

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

   (a) For a fixed number of dimensions to be retained, there is no predictable reason why either of PCA or supervised feature selection should lead to better performance; *a priori*, either approach is equally sensible.
   (b) For a fixed number of dimensions to be retained, getting those dimensions via PCA will in general lead to better performance than getting them from supervised feature selection, because PCA retains the maximally informative dimensions.

(c) Whilst (b) is true, supervised feature selection is often preferred because the original features are more interpretable than principal components.
(d) Reducing the number of features leads to loss of information, hence can always be expected to lead to poorer performance.

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.)

   (a) $N = D$
   (b) $N >> D$
   (c) $N << D$
   (d) None of the above

6. 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.)

   (a) The neural net with lower cross-entropy error will always have lower classification error as well.
   (b) The neural net with lower cross-entropy error could have higher classification error.
   (c) Scenario (b) is possible only when the neural net with lower classification error has overfit the data.
   (d) Scenario (b) is possible only when the neural net with lower cross-entropy error has overfit the data.

7. 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?

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

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

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

9. 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) (II) is problematic because it's non-differentiable, but either of (I) or (III) should give the same result.
   (b) Any of the three could be easily used for backpropagation, but (I) is preferred because it corresponds to maximising the likelihood of the data.
   (c) Any of the three could be easily used for backpropagation, but (III) is preferred because it corresponds to maximising the likelihood of the data.
   (d) (II) is problematic because it's non-differentiable; (I) is preferred to (III) because the former corresponds to maximising the likelihood of the data.
   (e) (II) is problematic because it's non-differentiable; (III) is preferred to (I) because the former corresponds to maximising the likelihood of the data.