ELL784: Introduction to Machine Learning

Assignments Quiz, Form: A

Maximum marks: 12

(Answer all questions on this question paper. Read all section-specific instructions carefully.)

Name: _____

Entry Number: _____

Section 1. Multiple choice questions

Instructions: Each question may have any number of correct choices. Clearly mark (tick) all choices you believe to be correct (1.5 marks for each correct choice, −0.5 for each incorrect choice).

1. Can cross-validation error be regarded as a reasonable estimate of testing error?
   (a) Yes, even if the error has been obtained after hyperparameter tuning
   (b) Yes, but only when no hyperparameter tuning has been done
   (c) Yes, but only when the number of folds is not too large
   (d) No, never

2. Which of the following are reasonable ways to select the number of clusters K for K-means?
   (a) Minimising the distortion function J as a function of K
   (b) Set a threshold τ, pick the smallest K such that $J < \tau$
   (c) Set a threshold τ, pick the smallest K such that the reduction in J from $K-1$ to K clusters is less than τ
   (d) Set a threshold τ, pick the smallest K such that the reduction in J from K to $K+1$ clusters is less than τ

3. In training an SVM with a polynomial kernel, one typically has two hyperparameters: the order of the polynomial d, and the slack penalty C. Suppose one does a grid search on these and obtains a contour plot showing pairs of values which correspond to the same cross-validation error. Moving along a given contour in the direction of increasing d, C will generally be
   (a) Increasing
   (b) Decreasing
   (c) Increasing for the part of the contour corresponding to overfitting, decreasing for the part corresponding to underfitting
   (d) Decreasing for the part of the contour corresponding to overfitting, increasing for the part corresponding to underfitting

4. Suppose you try quadratic (L2) regularisation on a regression model fit to data sets of different sizes sampled from the same population (*e.g.*, the data sets of size 20 and 100 you used in Assignment 1). The error function used is sum-of-squares error, $E(\mathbf{w}) = \sum_{n=1}^{N}(y(\mathbf{x}_n) - t_n)^2$. For each data set, you tune the regularisation parameter λ using cross-validation. What is the expected relation between the value of λ obtained and N, the size of the training data set used?
   (a) λ should increase as N increases
   (b) λ should decrease as N increases
   (c) The relation between λ and N depends on the dimension of the parameter vector $\mathbf{w}$
   (d) In general, λ should not depend on N

5. A neural network with 30 hidden units gave a cross-validation accuracy of 96% on a classification data set; when the number of units was increased to 40, the cross-validation accuracy was recorded as 91%. Which of the following are likely to increase the generalisation performance of the 40-hidden-unit network in this case?

   (a) Making the backpropagation convergence criterion more stringent

   (b) Decreasing the learning rate η

   (c) Adding a second hidden layer of similar dimension to the current one

   (d) Early stopping of backpropagation

6. Suppose you've been given a regression data set generated from a polynomial function plus some unknown kind of noise. You fit a regression function to it by minimising sum-of-squares error. In which of the following circumstances will the resulting model be expected to accurately recover the underlying polynomial, assuming you have provided for sufficient data and model complexity?

   (a) Only when the noise is Gaussian

   (b) Only when the noise is symmetric about zero

   (c) Only when the noise is zero-mean

   (d) Always

Section 2. Numerical/Short-answer questions

Instructions: Please write the answer immediately following each part of each question. No working needs to be shown, and marks may be deducted for unnecessary clutter.

7. Similar to Assignment 4, let's say you assess your clustering on a labeled data set containing 300 images each from 4 classes by assigning each cluster the label found most frequently within it.

   (a) Assuming you have taken the number of clusters K to be 20 or fewer, what is the lowest possible accuracy this could give you? **[1]**

   (b) In general, are there values of K for which this lowest possible accuracy will be greater than the above? If so, what is the smallest value of K for which this will happen, for the given data set? **[1]**

   (c) What is the smallest K at which 100% accuracy can theoretically be achieved for the given data set? **[0.5]**

   (d) Why is picking K so as to maximise this notion of accuracy not a good way of determining the number of clusters? **[0.5]**

ELL784: Introduction to Machine Learning

Assignments Quiz, Form: B

Maximum marks: 12

(Answer all questions on this question paper. Read all section-specific instructions carefully.)

Name: _____

Entry Number: _____

Section 1. Multiple choice questions

Instructions: Each question may have any number of correct choices. Clearly mark (tick) all choices you believe to be correct (1.5 marks for each correct choice, −0.5 for each incorrect choice).

1. Which of the following are reasonable ways to select the number of clusters K for K-means?
   (a) Minimising the distortion function J as a function of K
   (b) Set a threshold τ, pick the smallest K such that $J < \tau$
   (c) Set a threshold τ, pick the smallest K such that the reduction in J from $K-1$ to K clusters is less than τ
   (d) Set a threshold τ, pick the smallest K such that the reduction in J from K to $K+1$ clusters is less than τ
2. In training an SVM with a polynomial kernel, one typically has two hyperparameters: the order of the polynomial d, and the slack penalty C. Suppose one does a grid search on these and obtains a contour plot showing pairs of values which correspond to the same cross-validation error. Moving along a given contour in the direction of increasing d, C will generally be
   (a) Increasing
   (b) Decreasing
   (c) Increasing for the part of the contour corresponding to overfitting, decreasing for the part corresponding to underfitting
   (d) Decreasing for the part of the contour corresponding to overfitting, increasing for the part corresponding to underfitting
3. Can cross-validation error be regarded as a reasonable estimate of testing error?
   (a) Yes, even if the error has been obtained after hyperparameter tuning
   (b) Yes, but only when no hyperparameter tuning has been done
   (c) Yes, but only when the number of folds is not too large
   (d) No, never
4. A neural network with 30 hidden units gave a cross-validation accuracy of 96% on a classification data set; when the number of units was increased to 40, the cross-validation accuracy was recorded as 91%. Which of the following are likely to increase the generalisation performance of the 40-hidden-unit network in this case?
   (a) Making the backpropagation convergence criterion more stringent
   (b) Decreasing the learning rate η
   (c) Adding a second hidden layer of similar dimension to the current one
   (d) Early stopping of backpropagation

5. Suppose you've been given a regression data set generated from a polynomial function plus some unknown kind of noise. You fit a regression function to it by minimising sum-of-squares error. In which of the following circumstances will the resulting model be expected to accurately recover the underlying polynomial, assuming you have provided for sufficient data and model complexity?

   (a) Only when the noise is Gaussian

   (b) Only when the noise is symmetric about zero

   (c) Only when the noise is zero-mean

   (d) Always

6. Suppose you try quadratic (L2) regularisation on a regression model fit to data sets of different sizes sampled from the same population (*e.g.*, the data sets of size 20 and 100 you used in Assignment 1). The error function used is sum-of-squares error, $E(\mathbf{w}) = \sum_{n=1}^{N}(y(\mathbf{x}_n) - t_n)^2$. For each data set, you tune the regularisation parameter λ using cross-validation. What is the expected relation between the value of λ obtained and N, the size of the training data set used?

   (a) λ should increase as N increases

   (b) λ should decrease as N increases

   (c) The relation between λ and N depends on the dimension of the parameter vector $\mathbf{w}$

   (d) In general, λ should not depend on N

Section 2. Numerical/Short-answer questions

Instructions: Please write the answer immediately following each part of each question. No working needs to be shown, and marks may be deducted for unnecessary clutter.

7. Similar to Assignment 4, let's say you assess your clustering on a labeled data set containing 200 images each from 5 classes by assigning each cluster the label found most frequently within it.

   (a) Assuming you have taken the number of clusters K to be 20 or fewer, what is the lowest possible accuracy this could give you? **[1]**

   (b) In general, are there values of K for which this lowest possible accuracy will be greater than the above? If so, what is the smallest value of K for which this will happen, for the given data set? **[1]**

   (c) What is the smallest K at which 100% accuracy can theoretically be achieved for the given data set? **[0.5]**

   (d) Why is picking K so as to maximise this notion of accuracy not a good way of determining the number of clusters? **[0.5]**

ELL784: Introduction to Machine Learning

Assignments Quiz, Form: C

Maximum marks: 12

(Answer all questions on this question paper. Read all section-specific instructions carefully.)

Name: _____

Entry Number: _____

Section 1. Multiple choice questions

Instructions: Each question may have any number of correct choices. Clearly mark (tick) all choices you believe to be correct (1.5 marks for each correct choice, −0.5 for each incorrect choice).

1. A neural network with 30 hidden units gave a cross-validation accuracy of 96% on a classification data set; when the number of units was increased to 40, the cross-validation accuracy was recorded as 91%. Which of the following are likely to increase the generalisation performance of the 40-hidden-unit network in this case?
   (a) Making the backpropagation convergence criterion more stringent
   (b) Decreasing the learning rate η
   (c) Adding a second hidden layer of similar dimension to the current one
   (d) Early stopping of backpropagation
2. Suppose you've been given a regression data set generated from a polynomial function plus some unknown kind of noise. You fit a regression function to it by minimising sum-of-squares error. In which of the following circumstances will the resulting model be expected to accurately recover the underlying polynomial, assuming you have provided for sufficient data and model complexity?
   (a) Only when the noise is Gaussian
   (b) Only when the noise is symmetric about zero
   (c) Only when the noise is zero-mean
   (d) Always
3. Which of the following are reasonable ways to select the number of clusters K for K-means?
   (a) Minimising the distortion function J as a function of K
   (b) Set a threshold τ, pick the smallest K such that $J < \tau$
   (c) Set a threshold τ, pick the smallest K such that the reduction in J from $K-1$ to K clusters is less than τ
   (d) Set a threshold τ, pick the smallest K such that the reduction in J from K to $K+1$ clusters is less than τ
4. In training an SVM with a polynomial kernel, one typically has two hyperparameters: the order of the polynomial d, and the slack penalty C. Suppose one does a grid search on these and obtains a contour plot showing pairs of values which correspond to the same cross-validation error. Moving along a given contour in the direction of increasing d, C will generally be
   (a) Increasing
   (b) Decreasing
   (c) Increasing for the part of the contour corresponding to overfitting, decreasing for the part corresponding to underfitting
   (d) Decreasing for the part of the contour corresponding to overfitting, increasing for the part corresponding to underfitting

5. Can cross-validation error be regarded as a reasonable estimate of testing error?

   (a) Yes, even if the error has been obtained after hyperparameter tuning

   (b) Yes, but only when no hyperparameter tuning has been done

   (c) Yes, but only when the number of folds is not too large

   (d) No, never

6. Suppose you try quadratic (L2) regularisation on a regression model fit to data sets of different sizes sampled from the same population (*e.g.*, the data sets of size 20 and 100 you used in Assignment 1). The error function used is sum-of-squares error, $E(\mathbf{w}) = \sum_{n=1}^{N}(y(\mathbf{x}_n) - t_n)^2$. For each data set, you tune the regularisation parameter λ using cross-validation. What is the expected relation between the value of λ obtained and N, the size of the training data set used?

   (a) λ should increase as N increases

   (b) λ should decrease as N increases

   (c) The relation between λ and N depends on the dimension of the parameter vector $\mathbf{w}$

   (d) In general, λ should not depend on N

Section 2. Numerical/Short-answer questions

Instructions: Please write the answer immediately following each part of each question. No working needs to be shown, and marks may be deducted for unnecessary clutter.

7. Similar to Assignment 4, let's say you assess your clustering on a labeled data set containing 100 images each from 8 classes by assigning each cluster the label found most frequently within it.

   (a) Assuming you have taken the number of clusters K to be 20 or fewer, what is the lowest possible accuracy this could give you? **[1]**

   (b) In general, are there values of K for which this lowest possible accuracy will be greater than the above? If so, what is the smallest value of K for which this will happen, for the given data set? **[1]**

   (c) What is the smallest K at which 100% accuracy can theoretically be achieved for the given data set? **[0.5]**

   (d) Why is picking K so as to maximise this notion of accuracy not a good way of determining the number of clusters? **[0.5]**

ELL784: Introduction to Machine Learning

Assignments Quiz, Form: D

Maximum marks: 12

(Answer all questions on this question paper. Read all section-specific instructions carefully.)

Name: _____

Entry Number: _____

Section 1. Multiple choice questions

Instructions: Each question may have any number of correct choices. Clearly mark (tick) all choices you believe to be correct (1.5 marks for each correct choice, −0.5 for each incorrect choice).

1. A neural network with 30 hidden units gave a cross-validation accuracy of 96% on a classification data set; when the number of units was increased to 40, the cross-validation accuracy was recorded as 91%. Which of the following are likely to increase the generalisation performance of the 40-hidden-unit network in this case?
   (a) Making the backpropagation convergence criterion more stringent
   (b) Decreasing the learning rate η
   (c) Adding a second hidden layer of similar dimension to the current one
   (d) Early stopping of backpropagation
2. Which of the following are reasonable ways to select the number of clusters K for K-means?
   (a) Minimising the distortion function J as a function of K
   (b) Set a threshold τ, pick the smallest K such that $J < \tau$
   (c) Set a threshold τ, pick the smallest K such that the reduction in J from $K-1$ to K clusters is less than τ
   (d) Set a threshold τ, pick the smallest K such that the reduction in J from K to $K+1$ clusters is less than τ
3. In training an SVM with a polynomial kernel, one typically has two hyperparameters: the order of the polynomial d, and the slack penalty C. Suppose one does a grid search on these and obtains a contour plot showing pairs of values which correspond to the same cross-validation error. Moving along a given contour in the direction of increasing d, C will generally be
   (a) Increasing
   (b) Decreasing
   (c) Increasing for the part of the contour corresponding to overfitting, decreasing for the part corresponding to underfitting
   (d) Decreasing for the part of the contour corresponding to overfitting, increasing for the part corresponding to underfitting

4. Suppose you try quadratic (L2) regularisation on a regression model fit to data sets of different sizes sampled from the same population (*e.g.*, the data sets of size 20 and 100 you used in Assignment 1). The error function used is sum-of-squares error, $E(\mathbf{w}) = \sum_{n=1}^{N}(y(\mathbf{x}_n) - t_n)^2$. For each data set, you tune the regularisation parameter λ using cross-validation. What is the expected relation between the value of λ obtained and N, the size of the training data set used?

   (a) λ should increase as N increases

   (b) λ should decrease as N increases

   (c) The relation between λ and N depends on the dimension of the parameter vector $\mathbf{w}$

   (d) In general, λ should not depend on N

5. Can cross-validation error be regarded as a reasonable estimate of testing error?

   (a) Yes, even if the error has been obtained after hyperparameter tuning

   (b) Yes, but only when no hyperparameter tuning has been done

   (c) Yes, but only when the number of folds is not too large

   (d) No, never

6. Suppose you've been given a regression data set generated from a polynomial function plus some unknown kind of noise. You fit a regression function to it by minimising sum-of-squares error. In which of the following circumstances will the resulting model be expected to accurately recover the underlying polynomial, assuming you have provided for sufficient data and model complexity?

   (a) Only when the noise is Gaussian

   (b) Only when the noise is symmetric about zero

   (c) Only when the noise is zero-mean

   (d) Always

Section 2. Numerical/Short-answer questions

Instructions: Please write the answer immediately following each part of each question. No working needs to be shown, and marks may be deducted for unnecessary clutter.

7. Similar to Assignment 4, let's say you assess your clustering on a labeled data set containing 50 images each from 10 classes by assigning each cluster the label found most frequently within it.

   (a) Assuming you have taken the number of clusters K to be 20 or fewer, what is the lowest possible accuracy this could give you? **[1]**

   (b) In general, are there values of K for which this lowest possible accuracy will be greater than the above? If so, what is the smallest value of K for which this will happen, for the given data set? **[1]**

   (c) What is the smallest K at which 100% accuracy can theoretically be achieved for the given data set? **[0.5]**

   (d) Why is picking K so as to maximise this notion of accuracy not a good way of determining the number of clusters? **[0.5]**