Indian Institute of Technology Kanpur
CS771A Introduction to Machine Learning (Sem I, 2018-19)

QUESTION 1

Assignment Number: 1
Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: September 2, 2018

The decision tree will predict the label, that is majority in the leaf node, in case the leaf node does not contains homogeneous label.

$NumberofMisclassification$ at Tree A= $Numberofmisclassification$ on the left leaf node $+Numberofmisclassification$ on the right leaf node
$NumberofMisclassification$ at Tree A= $100 + 100$
$NumberofMisclassification$ at Tree A= 200
$Misclassification$ rate at Tree A= (200/800)
$Misclassification$ rate at Tree A= 0.25

$NumberofMisclassification$ at Tree B= $Numberofmisclassification$ on the left leaf node $+Numberofmisclassification$ on the right leaf node
$NumberofMisclassification$ at Tree B= $200 + 0$
$NumberofMisclassification$ at Tree B= 200
$Misclassification$ rate at Tree B= (200/800)
$Misclassification$ rate at Tree B= 0.25

Therefore, miscalssification rate for Decision Tree A and Decision Tree B is same.

$InformationGain = Entropy - WeightedEntropy$
$H_A(Y)$ is the entropy calculated at Tree A
$Entropy(H(Y)) = -\sum_{i=1}^{k} P(Y_i) * log_2(P(Y_i))$
$H_A(Y) = -\sum_{i=1}^{2} P(Y_1) * log_2(P(Y_1)) + P(Y_2) * log_2(P(Y_2))$
$H_A(Y) = -(400/800 * log_2(P(400/800)) + 400/800 * log_2(400/800))$
$H_A(Y) = -(1/2 * log_2(P((1/2)) + 1/2 * log_2(1/2))$
$H_A(Y) = -(2 * (1/2) * log_2(P((1/2))$
$H_A(Y) = -(-1)$
$H_A(Y) = 1$

Weighted Entropy at A(W_A)=$(|D_1|/|D|) * H_{D1A}(Y) + (|D_2|/|D|) * H_{D2A}(Y)$
$H_{D1A} = -(3/4 * log_2(3/4) + 1/4 * log_2(1/4))$
$H_{D1A} = -(3/4 * (-0.415) + 1/4(-2))$
$H_{D1A} = -(-0.31125 - 0.5)$
$H_{D1A} = 0.811$

$H_{D2A} = -(1/4 * log_2(1/4) + 3/4 * log_2(3/4))$
$H_{D2A} = -(1/4 * (-2) + 3/4(-0.415))$
$H_{D2A} = -(-0.5 - 0.31125)$
$H_{D2A} = 0.811$

$|D_1| = 400, |D_2| = 400, |D| = 800$
$W_A = 400/800 * 0.811 + 400/800 * 0.811$
$W_A = 1/2 * 0.811 + 1/2 * 0.811$
$W_A = 0.811$

Information Gain at Tree A(IG_A)=$H_A(Y) - W_A$
$IG_A = 1 - 0.811$
$IG_A = 0.189$

$H_B(Y)$ is the entropy calculated at Tree B
$Entropy(H(Y)) = -\sum_{i=1}^{k} P(Y_i) * log_2(P(Y_i))$
$H_B(Y) = -\sum_{i=1}^{2} P(Y_1) * log_2(P(Y_1)) + P(Y_2) * log_2(P(Y_2))$
$H_B(Y) = -(400/800 * log_2(P(400/800)) + 400/800 * log_2(400/800))$
$H_B(Y) = -(1/2 * log_2(P((1/2)) + 1/2 * log_2(1/2))$
$H_B(Y) = -(2 * (1/2) * log_2(P((1/2))$
$H_B(Y) = -(-1)$
$H_B(Y) = 1$

Weighted Entropy at B(W_B)=$(|D_1|/|D|) * H_{D1B}(Y) + (|D_2|/|D|) * H_{D2B}(Y)$
$H_{D1B} = -(1/3 * log_2(1/3) + 2/3 * log_2(2/3))$
$H_{D1B} = -(1/3 * (-1.599) + 2/3(-0.599))$
$H_{D1B} = -(-0.533 - 0.395)$
$H_{D1B} = 0.928$

$H_{D2B} = -(1/1 * log_2(1) + 0 * log_2(0))$
$H_{D2B} = -(1 * 0 + 0log_2(0))$
$H_{D2B} = 0$

$|D_1| = 600, |D_2| = 200, |D| = 800$
$W_B = 600/800 * 0.928 + 200/800 * 0$
$W_B = 0.696$

Information Gain at Tree B(IG_B)=$H_B(Y) - W_B$
$IG_B = 1 - 0.696$
$IG_B = 0.304$

Information Gain is more in Tree B than Tree A.

Though misclassification rate is same in both the decision tree A and B but decision tree B is better than decision tree A in terms of information gain.

Indian Institute of Technology Kanpur
CS771A Introduction to Machine Learning (Sem I, 2018-19)

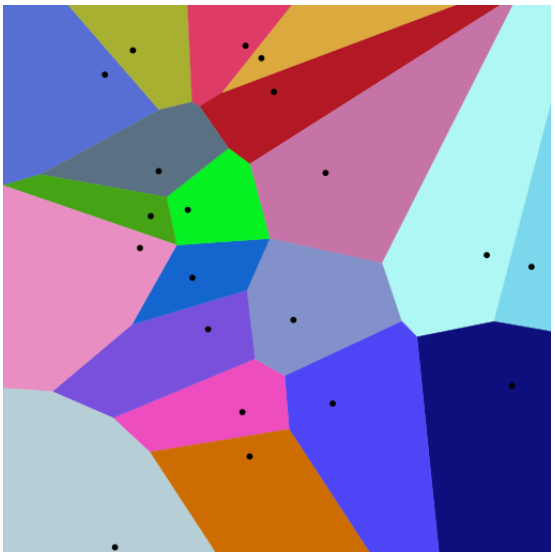
QUESTION

2

Assignment Number: 1
Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: September 2, 2018

A classifier is said to be consistent if the error rate of the classifier approaches Bayes optimal rate when given access to infinite amount of training data. Given, Bayes optimal error to be zero and the training data is noise free i.e there is no mislabelled data in the training dataset.

1NN classifier is consistent i.e the error approaches zero when given access to infinite amount of training data where no mislabelled data exists. When there is infinite amount of data, the entire input space is covered by the training data points. The Voronoi diagram of 1NN classifiers is given below:



For infinite set of training data the Voronoi diagram of 1NN classifier will be bounded to a region of the datapoint only.The test data points will overlap with any of the existing training data point and the classifier will predict the label which is the label of the existing training data point.Since,no training datapoint is mislabelled the label predicted for the test datapoints will also be correct. Therefore, the error rate approaches zero and hence 1NN classifier is consistent.

Indian Institute of Technology Kanpur
CS771A Introduction to Machine Learning (Sem I, 2018-19)

QUESTION

3

Assignment Number: 1
Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: September 2, 2018

We know for unregularized linear regression model,$\hat{w} = (X^T X)^{-1} X^T y$
For unregularized model $f(x^*) = \hat{w}^T x^*$
$f(x^*) = [(X^T X)^{-1} X^T y]^T x^*$
$f(x^*) = y^T X[(X^T X)^{-1}]^T x^*$
$f(x^*) = \sum_{n=1}^{N} L_n y_n$

Here L_n is the is the each entry of the matrix obtained by $X(X^T X)^{-1} x^*$.

In KNN $f(x^*) = \sum_{n=1}^{N} w_n y_n$ where $w_n = \frac{1}{\|x_* - x_n\|_p}$ which is the inverse of the distance of the test point to each training data point.

In unregularized linear regression model, the L_n is not dependent on the inverse of distance of each training data point to the test data point rather it only depends on the input feature matrix X and the new test data point x^*.

Indian Institute of Technology Kanpur
CS771A Introduction to Machine Learning (Sem I, 2018-19)

QUESTION

4

Assignment Number: 1
Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: September 2, 2018

The normal l_2 regularized least squares regression can be written as:
$L(w) = \sum_{n=1}^{N}(y_n - w^T x_n)^2 + (1/2)\lambda w^T w$ where λ is a constant value which have equal regularizing effect on each element of weight vector w.

The main objective of us in this question is to replace λ with an entity that have different effect for each element in weight vector w.
λ is replaced by a D*D diagonal matrix λ_D where the diagonal elements are $\lambda_1, \lambda_2, \ldots\ldots\ldots, \lambda_d$.

Objective function:$L(w) = \sum_{n=1}^{N}(y_n - w^T x_n)^2 + w^T \lambda_D w$
Differentiating the objective function with respect to w and equating that to 0

$\delta L(w)/\delta w = 0$

$\delta/\delta w[\sum_{n=1}^{N}(y_n - w^T x_n)^2] + \delta/\delta w[w^T \lambda_D w] = 0$

$2\sum_{n=1}^{N}(y_n - w^T x_n)\delta/\delta w[(y_n - w^T x_n)] + 2w\lambda_D = 0$

$\sum_{n=1}^{N}(y_n - w^T x_n)(-x_n^T) + w\lambda_D = 0$

$\sum_{n=1}^{N} x_n(y_n - x_n^T w) + w\lambda_D = 0$

$\sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} x_n x_n^T w + w\lambda_D = 0$

$\sum_{n=1}^{N} x_n x_n^T w + \lambda_D w = \sum_{n=1}^{N} x_n y_n$

$w = (\sum_{n=1}^{N} x_n x_n^T + \lambda_D)^{-1} \sum_{n=1}^{N} x_n y_n$

$w = (X^T X + \lambda_D)^{-1} X^T y$

Indian Institute of Technology Kanpur
CS771A Introduction to Machine Learning (Sem I, 2018-19)

QUESTION 5

Assignment Number: 1
Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: September 2, 2018

The objective function is: $L(S) = Trace[(Y - XBS)^T(Y - XBS)]$

The objective function is differentiated with respect to S and equated to 0 to find $\hat{S}$ where $\hat{S}$ is the $argmin_S Trace[(Y - XBS)^T(Y - XBS)]$

$\delta L(S) = 0$

$\frac{\delta}{\delta S} Trace[(Y - XBS)^T(Y - XBS)] = 0$
We can write,
$Trace[(Y - XBS)^T(Y - XBS)] = Trace[(Y^T - S^T B^T X^T)(Y - XBS)]$
[By using the identity $(AB)^T = B^T A^T$]

$Trace[(Y - XBS)^T(Y - XBS)] = Trace[Y^TY - Y^TXBS - S^TB^TX^TY + S^TB^TX^TXBS]$

$\frac{\delta}{\delta S} Trace[(Y-XBS)^T(Y-XBS)] = \frac{\delta}{\delta S} Trace[Y^TY-Y^TXBS-S^TB^TX^TY+S^TB^TX^TXBS]$

$= \frac{\delta}{\delta S} Trace[Y^TY] - \frac{\delta}{\delta S} Trace[Y^TXBS] - \frac{\delta}{\delta S} Trace[S^TB^TX^TY] + \frac{\delta}{\delta S} Trace[S^TB^TX^TXBS]$

$= 0 - (Y^TXB)^T - B^TX^TY + (B^TX^TXB + B^TX^TXB)S$

[By using the identity rules $\frac{\delta}{\delta X} Trace[AXB] = A^TB^T$, $\frac{\delta}{\delta X} Trace[(X^TA)] = A$ and $\frac{\delta}{\delta X} Trace[(X^TBX)] = BX + B^TX$]

$= 0 - B^TX^TY - B^TX^TY + (B^TX^TXB + B^TX^TXB)S$

$= -2B^TX^TY + 2B^TX^TXBS$

$$= -2B^TX^TY + 2B^TX^TXBS \tag{1}$$

The equation number (2) is equated to 0 to find the argmin of S:

$$\begin{aligned} -2B^TX^TY + 2B^TX^TXBS &= 0 \\ 2B^TX^TXBS &= 2B^TX^TY \\ S &= (B^TX^TXB)^{-1}B^TX^TY \end{aligned}$$

The objective function of standard multi-output regression is: $L(W) = Trace[(Y-XW)^T(Y-XW)]$

The objective function is differentiated with respect to W and equated to 0 to find $\hat{W}$ where $\hat{W}$ is the $argmin_W Trace[(Y - XW)^T(Y - XW)]$

$\delta L(W) = 0$

$\frac{\delta}{\delta W} Trace[(Y - XW)^T(Y - XW)] = 0$

We can write,

$Trace[(Y - XW)^T(Y - XW)] = Trace[(Y^T - W^T X^T)(Y - Xw)]$

[By using the identity $(AB)^T = B^T A^T$]

$Trace[(Y - XW)^T(Y - XW)] = Trace[Y^T Y - Y^T XW - W^T X^T Y + W^T X^T XW]$

$\frac{\delta}{\delta S} Trace[(Y - XW)^T(Y - XW)] = \frac{\delta}{\delta S} Trace[Y^T Y - Y^T XW - W^T X^T Y + W^T X^T XW]$

$= \frac{\delta}{\delta S} Trace[Y^T Y] - \frac{\delta}{\delta S} Trace[Y^T XW] - \frac{\delta}{\delta S} Trace[W^T X^T Y] + \frac{\delta}{\delta S} Trace[W^T X^T XW]$

$= 0 - (Y^T X)^T - X^T Y + (X^T X + (X^T X)^T)W$

[By using the identity rules $\frac{\delta}{\delta X} Trace[AXB] = A^T B^T$, $\frac{\delta T}{\delta X} race[(X^T A)] = A$ and $\frac{\delta}{\delta X} Trace[(X^T BX)] = BX + B^T X$]

$= 0 - X^T Y - X^T Y + (X^T X + X^T X)W$

$= -2X^T Y + 2X^T XW$

$$= -2X^T Y + 2X^T XW \tag{2}$$

The equation number (2) is equated to 0 to find the argmin of W:

$$\begin{aligned} -2X^T Y + 2X^T XW &= 0 \\ 2X^T XW &= 2X^T Y \\ W &= (X^T X)^{-1} X^T Y \end{aligned}$$

The two solutions are identical because:
$W = (X^T X)^{-1} X^T Y$ and $S = ((XB)^T XB)^{-1}(XB)^T Y$ because the form of equations remains same but only X is transformed byXB in the later solution.

Indian Institute of Technology Kanpur
CS771A Introduction to Machine Learning (Sem I, 2018-19)

QUESTION

6

Assignment Number: 1
Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: September 2, 2018

Method1: Convex
1) Mean of each seen classes is computed. (40,4096) matrix is returned.
2) Similarity matrix is computed using the seen class attribute matrix(ak) and unseen class attribute matrix(ac).
$similaritymatrix[i,j] = ac[i]^T * ak[j]$. 3) Similarity matrix is normalized such that sum of each row of similarity matrix equals 1.
4) Mean of unseen classes are computed.
$meanunseen[j] = meanunseen[j] + (similaritymatrix[j][i] * meanseen[i]$ where i ranges from 1 to 40. 5) Euclidean Distance of each test data point is calculated from the mean of unseen classes and the class closest to the test data point is predicted to be the class label of test data point.
6)Accuracy is calculated by number of correctly classified points/total number of points.

The accuracy achieved is 46.89 percent.

Method2: Regerssion
1) Mean of each seen classes is computed. (40,4096) matrix is returned.
2) W is calculated by performing matrix multiplications in the order $(A_s^T A_S + I)^{-1} A_s M_s$ where A_s is seen class attribute matrix, M_s is the feature matrix.
3) mean of unseen classes is computed using Wac where ac is unseen class attribute matrix.
4) Euclidean Distance of each test data point is calculated from the mean of unseen classes and the class closest to the test data point is predicted to be the class label of test data point.
5)Accuracy is calculated by number of correctly classified points/total number of points.

The accuracy achieved for different values of λare :

Value of λ	Accuracy (%)
0.01	58.09
0.1	59.54
1.0	67.39
10.0	73.28
20.0	71.68
50.0	65.08
100.0	56.47

Table 1: Accuracy for different values of λ

The value of $\lambda = 10$ gives the best accuracy.