Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 2, 2018

QUESTION 1

Part 1 : Computing training data misclassification rate

For tree A,

Assuming that left branch misclassifies examples of class 1 and right branch misclassifies examples of class 0, misclassification rate will be:

$$\frac{(100+100)}{800} = 0.25$$

For tree B,

Assuming that left branch misclassifies examples of class 0 and right branch misclassifies examples of class 1, misclassification rate will be:

$$\frac{(200+0)}{800} = 0.25$$

We can observe that training data misclassification rate turns out to be same for both the cases.

Part 2 : Evaluating the Information gain

$$\text{Information Gain } (IG) = H(S) - \frac{\|S1\|}{\|S\|}H(S1) - \frac{\|S2\|}{\|S\|}H(S2)$$

For tree A,

Probabilities at root node of the tree $= p_1 = p_2 = 0.5$
Probability of class 0 at left node $= 0.75$ and probability of class 1 at left node $= 0.25$
Probability of class 0 at right node $= 0.25$ and probability of class 1 at right node $= 0.75$

$$H(S) = -0.5\log 0.5 - 0.5\log 0.5 = 1$$

$$H(S1) = -0.75\log 0.75 - 0.25\log 0.25 = 0.811$$

$$H(S2) = -0.25\log 0.25 - 0.75\log 0.75 = 0.811$$

Therefore,

$$IG = 1 - \frac{400}{800} * 0.811 - \frac{400}{800} * 0.811 = 0.189$$

For tree B,

Probabilities at root node of the tree $= p_1 = p_2 = 0.5$
Probability of class 0 at left node $= 0.33$ and probability of class 1 at left node $= 0.67$
Probability of class 0 at right node $= 1$ and probability of class 1 at right node $= 0$

$$H(S) = -0.5\log 0.5 - 0.5\log 0.5 = 1$$

$$H(S1) = -0.33 \log 0.33 - 0.67 \log 0.67 = 0.918$$

$$H(S2) = 0$$

Therefore,

$$IG = 1 - \frac{600}{800} * 0.918 - \frac{200}{800} * 0 = 0.312$$

Part 3 : Result Analysis

We can observe that while training data misclassifcation rate turns out to be same for both the cases, information gain is higher for the tree B which indicates that tree B is a better decision making model in comparison to tree A.

This observation makes sense in the way that whereas tree B perfectly identifies the labels for right node and misidentifes the labels in left node with probability 0.33, in tree A both the nodes misidentify labels with probability 0.25.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 2, 2018

QUESTION 2

As stated in class slides that if given lots of training data, 1-nearest neighbour has an error-rate that is no worse than twice of the Bayes optimal classifier, therefore, this holds true if the 1-nearest neighbour classification algorithm has access to infinite amounts of training data.

Therefore, we have

$$\text{Error rate of 1-NN} \leq 2 * (\text{Bayes Optimal Error rate})$$

We are supposed to consider the noise-free setting (i.e., every training input is labeled correctly). In this case, the Bayes Optimal error rate is zero which implies

$$\text{Error rate of 1-NN} \leq 0$$

Therefore, 1-nearest neighbour algorithm is consistent in this setting.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 2, 2018

QUESTION 3

For the unregularized linear regression model, prediction at test input x_* is given by :

$$y_* = f(x_*) = w^T x_* = x_*^T w$$

And the solution for the weight function w is as follows :

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

So, the above equation can be written as :

$$y_* = f(x_*) = x_*^T (X^T X)^{-1} X^T y$$

which further is equivalent to

$$y_* = f(x_*) = \sum_{n=1}^{N} x_*^T (X^T X)^{-1} x_n y_n$$

The equation given in the problem is :

$$f(x_*) = \sum_{n=1}^{N} w_n y_n$$

Comparing the two equations, w_n is as follows :

$$w_n = x_*^T (X^T X)^{-1} x_n$$

Difference in these weights and weights in weighted version of K-nearest neighbours :

Let $(X^T X)^{-1}$ be denoted by a square matrix M, then w_n can be expressed as $w_n = x_*^T M x_n$. From this equation, we can see that w_n is a measure of similarity between two vectors x_* and x_n except that the similarity is modulated by a symmetric matrix M (something similar to modulation in distance in case of Mahalanobis distance).

On the other hand, the weights in weighted version of K-nearest neighbours are proportional to inverse of distance between the vectors x_* and x_n.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 2, 2018

QUESTION 4

The l_2 regularized least squares regression objective is written as :

$$\mathcal{L}(w) = \sum_{n=1}^{N}(y_n - w^T x)^2 + \frac{\lambda}{2} w^T w$$

In this case, the extent of regularization is same on all the features (and is controlled by λ).

An alternative of this which still uses l_2 regularization but the extent of regularization is different for each entry w_d is as follows :

$$\mathcal{L}(w) = \sum_{n=1}^{N}(y_n - w^T x)^2 + w^T M w$$

where the matrix M is a d-dimensional diagonal matrix whose each diagonal entry (M_{ii}) corresponds to the extent of regularization for the i^{th} entry w_i.

The above expression for $\mathcal{L}(w)$ can also be expressed in matrix form as follows :

$$\mathcal{L}(w) = (y - Xw)^T(y - Xw) + w^T M w$$

Derivation of the closed form expression for the weight vector w :

To get the closed form solution, differentiate the expression for $\mathcal{L}(w)$ w.r.t w and set it to 0 i.e.

$$\frac{\partial \mathcal{L}(w)}{\partial w} = 0$$

Applying the chain rule,

$$(y - Xw)^T \frac{\partial}{\partial w}(y - Xw) + (\frac{\partial}{\partial w}(y - Xw)^T)(y - Xw) + (M + M^T)w = 0$$

$$-(y - Xw)^T X - X^T(y - Xw) + (M + M^T)w = 0$$

$$-2X^T(y - Xw) + (M + M^T)w = 0$$

Now, since M is a diagonal matrix, therefore, $M = M^T$ which further implies

$$-2X^T(y - Xw) + 2Mw = 0$$

$$-2X^T y + 2X^T X w + 2Mw = 0$$

$$(X^T X + M)w = X^T y$$

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

So, the final closed form expression for w is :

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

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 2, 2018

QUESTION 5

First we need to verify that

$$\sum_{n=1}^{N}\sum_{m=1}^{M}(y_{nm}-w_m^Tx_n)^2 = \text{TRACE}[(Y-XW)^T(Y-XW)]$$

Let S = Y - XW. Then, the nm^{th} element of S is given by $y_{nm}-w_m^Tx_n$. Now, for S^T its mn^{th} element is same as nm^{th} element of S. Therefore, the mm^{th} diagonal element of S^TS is given by $\sum_{n=1}^{N}(y_{nm}-w_m^Tx_n)^2$.

Now, since TRACE is the sum of diagonal elements of a matrix. Therefore,

$$\text{TRACE}[S^TS] = \sum_{m=1}^{M}\sum_{n=1}^{N}(y_{nm}-w_m^Tx_n)^2$$

Finding closed form solution for $\hat{S}$:

Given,

$$\hat{S} = \arg\min_S \text{TRACE}[(Y-XBS)^T(Y-XBS)]$$

Differentiate the expression w.r.t S and set it to 0.

$$\frac{\partial}{\partial S}\text{TRACE}[(Y-XBS)^T(Y-XBS)] = 0$$

$$\text{TRACE}[\frac{\partial}{\partial S}(Y-XBS)^T(Y-XBS)] = 0$$

$$\text{TRACE}[2(Y-XBS)\frac{\partial}{\partial S}(Y-XBS)] = 0$$

$$\text{TRACE}[2(XB)^T(Y-XBS)] = 0$$

$$-2(XB)^T(Y-XBS) = 0$$

$$-2(XB)^TY + 2(XB)^TXBS = 0$$

$$(XB)^TXBS = (XB)^TY$$

$$S = ((XB)^TXB)^{-1}(XB)^TY$$

Clearly, we can see that the form of solution obtained is identical to solution in case of standard multi-output regression except the fact that input matrix X in latter is replaced by XB in this case.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Sahil Bansal
Roll Number: 150614
Date: September 2, 2018

QUESTION 6

Method 1:

Test-set classification accuracy : 46.89320388349515

Method 2:

List of accuracies obtained for different values of λ is as follows :

λ	**Test-set classification accuracy (in %)**
0.01	58.090614886731395
0.1	59.54692556634305
1	67.39482200647248
10	73.28478964401295
20	71.6828478964401
50	65.08090614886731
100	56.47249190938511

We can observe that $\lambda = 10$ gives the best test-set classification accuracy.