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

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 2, 2018

QUESTION 1

1. **For Tree A:**
From leaf node 1 min. number of element in a class is 100 and from leaf node 2 also min. number of element in a class is 100. Therefore,

$$MissclassificationRate = \frac{100+100}{800} = \frac{1}{4} \tag{1}$$

For Tree B:
From leaf node 1 min. number of element in a class is 200 and from leaf node 2 min. number of element in a class is 0. Therefore,

$$MissclassificationRate = \frac{200+0}{800} = \frac{1}{4} \tag{2}$$

Yes, they are equal!

2. **For Tree A:**

$$H(S) = -\frac{1}{2}(log(\frac{1}{2}) + log(\frac{1}{2})) = 1 \tag{3}$$

$$H(S_1) = -\frac{1}{4}(log(\frac{1}{4}) + 3log(\frac{3}{4})) = 0.811 \tag{4}$$

$$H(S_2) = -\frac{1}{4}(log(\frac{1}{4}) + 3log(\frac{3}{4})) = 0.811 \tag{5}$$

$$\Rightarrow IG_A = 1 - \frac{1}{2}0.811 - \frac{1}{2}0.811 = 0.189 \tag{6}$$

For Tree B:

$$H(S) = -\frac{1}{2}(log(\frac{1}{2}) + log(\frac{1}{2})) = 1 \tag{7}$$

$$H(S_1) = -\frac{1}{3}(log(\frac{1}{3}) + 2log(\frac{2}{3})) = 0.918 \tag{8}$$

$$H(S_2) = -\frac{2}{2}log(\frac{2}{2}) = 0. \tag{9}$$

$$\Rightarrow IG_B = 1 - \frac{3}{4}0.918 = 0.311 \tag{10}$$

Since $IG_B > IG_A$. Therefore, Tree B is better decision tree!

3. From (1) its not possible to predict which tree to select but from (2) we can decide Tree B to better. This makes sense because entropy is better deciding factor for decision tree.

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

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 2, 2018

QUESTION 2

Yes, **1NN** will be consistent in this case!
Since, there are infinite numbers of training data and each of them are correctly labeled without any noise. That means whenever you get a test data you can always find a training data completing close to it probability of finding such point will tend to 1 when the numbers of train data goes to infinity! Therefore, you can always classify them with no error in classifcation. In simple words you already have that test data in your training set!

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

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 2, 2018

QUESTION 3

Given: $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Therefore, a prediction at test input $\mathbf{x}_*$ can be written as $y_* = \hat{\mathbf{w}}^T\mathbf{x}_* = \mathbf{x}_*^T\hat{\mathbf{w}}$
Therefore,

$$y_* = \mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{11}$$

$$\Rightarrow y_* = \mathbf{W}\mathbf{y} \tag{12}$$

where, $\mathbf{W} = \mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Therefore, $\mathbf{W} = (w_1 w_2 ... w_N)$ comes out to be a 1xN matrix. We can also write $\mathbf{y} = (y_1 y_2 ... y_N)^T$.
Therefore,

$$y_* = \mathbf{W}\mathbf{y} = \sum_{n=1}^{N} w_n y_n \tag{13}$$

Therefore, w_n are the n^{th} index of the 1xN matrix $\mathbf{W}$. Knowing that $\mathbf{X}$ is matrix whose rows are the N training vectors $\mathbf{x}_n$, w_n can be written as:

$$w_n = \mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_n \tag{14}$$

So, w_n depends on the input $\mathbf{x}_*$ and all the training data's from $\mathbf{x}_1$ to $\mathbf{x}_n$. Since, there exist X^TX term in the expressin of w_n. Which is not in case with weighted kNN, where individual weights depends only on $\mathbf{x}_*$ and $\mathbf{x}_n$. Also, $\mathbf{x}_*$ comes in numerator in case of this while for kNN it comes in denominator. Another difference is that w_n are expressed as products of $\mathbf{x}_*$ while in kNN they are expressed as sum in denominator.

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

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 2, 2018

QUESTION 4

We can use a diagonal matrix $\mathbf{M} = diag(m_{11}m_{22}...m_{DD})$ to define a new regularization term as $\mathbf{w}_T\mathbf{M}\mathbf{w}$, inspired from mahalanobis distance!
Therefore, our loss becomes:

$$\mathbf{L}(\mathbf{w}) = \sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 + \mathbf{w}_T\mathbf{M}\mathbf{w} \tag{15}$$

We can also write it as,

$$\mathbf{L}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \mathbf{w}_T\mathbf{M}\mathbf{w} \tag{16}$$

$$\Rightarrow \frac{\partial \mathbf{L}(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\mathbf{M}\mathbf{w} \tag{17}$$

For closed form solution put $\frac{\partial \mathbf{L}(\mathbf{w})}{\partial \mathbf{w}} = 0$. Therefore,

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{w} + \mathbf{M}\mathbf{w} \tag{18}$$

$$\Rightarrow \mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X} + \mathbf{M})\mathbf{w} \tag{19}$$

$$Therefore, \mathbf{w} = (\mathbf{X}^T\mathbf{X} + \mathbf{M})^{-1}\mathbf{X}^T\mathbf{y} \tag{20}$$

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 2, 2018

QUESTION

5

$$\mathbf{L}(\mathbf{W}) = TRACE[(\mathbf{Y} - \mathbf{XBS})^T(\mathbf{Y} - \mathbf{XBS})] \tag{21}$$

$$= TRACE[(\mathbf{Y}^T - \mathbf{S}^T\mathbf{B}^T\mathbf{X}^T)(\mathbf{Y} - \mathbf{XBS})] \tag{22}$$

$$\Rightarrow \mathbf{L}(\mathbf{W}) = TRACE[\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{XBS} - \mathbf{S}^T\mathbf{B}^T\mathbf{X}^T\mathbf{Y} + \mathbf{S}^T\mathbf{B}^T\mathbf{X}^T\mathbf{XBS}] \tag{23}$$

Using the identities we get,

$$\Rightarrow \frac{\partial \mathbf{L}(\mathbf{W})}{\partial \mathbf{S}} = -(\mathbf{Y}^T\mathbf{XB})^T - \mathbf{B}^T\mathbf{X}^T\mathbf{Y} + ((\mathbf{B}^T\mathbf{X}^T\mathbf{XB}) + (\mathbf{B}^T\mathbf{X}^T\mathbf{XB})^T)\mathbf{S} \tag{24}$$

For closed form solution keeping $\mathbf{B}$ constant put $\frac{\partial \mathbf{L}(\mathbf{W})}{\partial \mathbf{S}} = 0$. Therefore,

$$\mathbf{B}^T\mathbf{X}^T\mathbf{Y} = (\mathbf{B}^T\mathbf{X}^T\mathbf{XB})\mathbf{S} \tag{25}$$

$$\Rightarrow \mathbf{S} = (\mathbf{B}^T\mathbf{X}^T\mathbf{XB})^{-1}\mathbf{B}^T\mathbf{X}^T\mathbf{Y} \tag{26}$$

writing $\mathbf{B}^T\mathbf{X}^T = (\mathbf{XB})^T$,

$$\Rightarrow \mathbf{S} = ((\mathbf{XB})^T\mathbf{XB})^{-1}(\mathbf{XB})^T\mathbf{Y} \tag{27}$$

replacing $\mathbf{XB} = \mathbf{V}$, we get,

$$\Rightarrow \mathbf{S} = (\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T\mathbf{Y} \tag{28}$$

Therefore, from the above equation when putting $\mathbf{XB} = \mathbf{V}$ we can see that the solution is identifical to the solution of standard multi-output regression where $\mathbf{X}$ have been transformed to $\mathbf{V}$.

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

Student Name: Shashi Kant Gupta
Roll Number: 160645
Date: September 2, 2018

QUESTION
6

Method 1:
Test accuracy with Euclidean distance is 46.8932038835
I had also tried implementing the mahalanobis distance but not much improvement observed. Following are the results based on number of iterations I runned to optimise theta:
Test accuracy for iter = 5 is: 47.1844660194
Test accuracy for iter = 10 is: 48.2362459547
Test accuracy for iter = 15 is: 48.8996763754
Test accuracy for iter = 20 is: 49.7249190939
Test accuracy for iter = 25 is: 50.1294498382
Test accuracy for iter = 30 is: 50.2103559871

Method 2:
Test accuracy for $\lambda = 0.01$ is: 58.0906148867
Test accuracy for $\lambda = 0.1$ is: 59.5469255663
Test accuracy for $\lambda = 1$ is: 67.3948220065
Test accuracy for $\lambda = 10$ is: 73.284789644
Test accuracy for $\lambda = 20$ is: 71.6828478964
Test accuracy for $\lambda = 50$ is: 65.0809061489
Test accuracy for $\lambda = 100$ is: 56.4724919094
$\lambda = 10$ gives the best result!