Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Manish Agrawal
Roll Number: 231110028
Date: September 15, 2023

QUESTION 1

SOLUTION-1

The given loss function is:

$$(\hat{w}_c, \hat{M}_c) = \text{argmin}_{w_c,\mathbf{M_c}} \sum_{x_n:y_n=c} \frac{1}{N_c}(x_n - w_c)^T \mathbf{M_c}(x_n - w_c) - \log |\mathbf{M_c}| \tag{1}$$

To get the optimum values of w_c we take the first order partial derivative with respect to w_c then make that equal to zero to get the stationary point.
$\frac{\partial(\hat{w}_c,\hat{M}_c)}{\partial \hat{w}_c} = 0$
then we get following equation::

$$-2\frac{1}{N_c}\sum_{x_n,y_n=c} \mathbf{M_c}(x_n - w_c) = 0$$

$$\sum_{x_n,y_n=c} \mathbf{M_c}(x_n - w_c) = 0$$

$$\mathbf{M_c}\sum_{x_n,y_n=c} x_n = \mathbf{M_c} w_c \sum_{x_n,y_n=c} 1$$

$$\boxed{\sum_{x_n,y_n=c} 1 = N_c}$$

$$w_c N_c = \sum_{x_n,y_n=c} x_n$$

$$w_c = \frac{1}{N_c}\sum_{x_n,y_n=c} x_n$$

$$\boxed{w_c = \mu_n}$$

To get the optimum values of $\mathbf{M_c}$ we take the first order partial derivative with respect to $\mathbf{M_c}$ then make that equal to zero to get the stationary point.
$\frac{\partial(\hat{w}_c,\hat{M}_c)}{\partial \hat{\mathbf{M}}_\mathbf{c}} = 0$
then we get following equation::

$$\sum_{x_n,y_n=c} \frac{1}{N_c}(x_n - w_c)^T(x_n - w_c) - \frac{1}{\mathbf{M}_c} = 0$$

$$\frac{1}{\mathbf{M}_c} = \sum_{x_n,y_n=c} \frac{1}{N_c}(x_n - w_c)^T(x_n - w_c)$$

$$\mathbf{M}_c^{-1} = \sum_{x_n, y_n = c} \frac{1}{N_c}(x_n - w_c)^T(x_n - w_c)$$

$$\boxed{\mathbf{M}_c = \left(\sum_{x_n, y_n = c} \frac{1}{N_c}(x_n - w_c)^T(x_n - w_c)\right)^{-1}}$$

Let's take $\mathbf{M}_c = \mathbf{I}$ where $\mathbf{I}$ =identity matrix

$$(\hat{w}_c, \hat{M}_c) = \text{argmin}_{w_c, \mathbf{M_c} = \mathbf{I}} \sum_{x_n : y_n = c} \frac{1}{N_c}(x_n - w_c)^T \mathbf{I}(x_n - w_c) - \log|\mathbf{I}|$$

$$\boxed{\hat{w}_c = argmin_{w_c} \sum_{x_n, y_n = c} \frac{1}{N_c} ||(\mathbf{x}_n - \mathbf{w}_c)||_2^2}$$

The equation provided represents the standard squared Euclidean distance.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Manish Agrawal
Roll Number: 231110028
Date: September 15, 2023

QUESTION 2

SOLUTION-2

Yes, the one-nearest-neighbor (1-NN) algorithm is consistent in the noise-free setting. In this scenario, where all training inputs are correctly labeled, the 1-NN algorithm consistently assigns a label to a test input based on the nearest neighbor in the training data. As the number of training examples approaches infinity, the likelihood of finding the exact match for any test input increases. Consequently, the algorithm's error rate approaches the Bayes optimal error rate of zero, making it consistent in this setting.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Manish Agrawal
Roll Number: 231110028
Date: September 15, 2023

QUESTION 3

SOLUTION-3

In Decision Tree a good criteria to choose a feature to spliton if we were doing regression, is variance in the outputs. For Decision tree regression, variance in the outputs can be used to assess purity.the objective is to find feature splits that minimize the variance of the labels within each child node. A suitable criterion for this purpose is the reduction in variance, often referred to as "variance reduction". For each candidate split, calculate the variance of the labels in the child nodes weighted by the number of examples in each child node:

$$\text{Weighted Variance Reduction} = \text{Variance}(P) - \left(\frac{n_{\text{left}}}{n} \cdot \text{Variance}(Leftchild) + \frac{n_{\text{right}}}{n} \cdot \text{Variance}(Rightchild)\right)$$

The criteria aim to maximize the reduction in variance when splitting, effectively minimizing the variance within each child node. Thus at each node when we choose a feature to split then at that time feature must be chosen such that the variance among labels decreases with each split so when variance is low then homogeneity of the set of real-valued labels of the examples at each node is high

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Manish Agrawal
Roll Number: 231110028
Date: September 15, 2023

QUESTION 4

SOLUTION-4

for linear regression model prediction given by

$$f(x_*) = W^T x^*$$

where W is given by:

$$W = (X^T X)^{-1} X^T Y$$

so therefore

$$=> f(x_*) = ((X^T X)^{-1} X^T Y)^T x_*$$

$$\boxed{\therefore W^T x_* = x_*^T W}$$

$$=> f(x_*) = x_*^T ((X^T X)^{-1} X^T Y)$$
$$=> f(x_*) = (x_*^T (X^T X)^{-1} X^T) Y$$

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

$$\boxed{\therefore \sum_{n=1}^{N} w_n y_n = WY}$$

$$\boxed{W = (x_*^T (X^T X)^{-1} X^T)}$$

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

so here weight vector w depends on X means depends on all training input x_1 to x_n and depends on input x_* but directly there is no distance term between input x_* and x_n but in K-nearest neighbours,weights are calculated by using the inverse distances between input x_* and each training set input(from x_1 to x_n)so it usually consider the proximity of test input to the all data points.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Manish Agrawal
Roll Number: 231110028
Date: September 15, 2023

QUESTION 5

SOLUTION-5

Loss function for linear regression model ::

$$L(W) = \sum_{n=1}^{N} (y_n - w^T x_n)^2$$

$$\therefore \sum_{n=1}^{N} (y_n - w^T x_n)^2 = \|\mathbf{Y} - \mathbf{W^T}.\mathbf{X}\|^2$$

Now we are replacing the input x_n by $\tilde{x}_n$

$\tilde{x}_n = x_n \odot m_n,$ where $\odot$ denotes elementwise product and m_n denotes the $D \times 1$ binary mask vector

then new loss function is:

$$L(W) = \|\mathbf{Y} - (\mathbf{X}.\mathbf{M})\mathbf{W}\|^2$$

the expected value of this new loss function E[L(W)]
This needs to be minimized to get the optimum value of w

$$L(w) = \arg\min_{w} E[\| \; Y - (M.X)W\|^2]$$

$$\boxed{\therefore \| \; W\|^2 = W^T.W}$$

$$L(w) = \arg\min_{w} E[((Y - (M.X)W)^T (Y - (M.X)W)]$$

$$=> L(w) = \arg\min_{w} (E[Y^T Y] - 2E[W^T (M.X)^T Y] + E[W^T (M.X)^T (M.X)W]) - - - eq1$$

M is a random variable as we know and it is given that it follows Bernouli Distribution

$$m_{nd} \sim \text{Bernoulli}(p)$$

Here, $m_{nd} = 1$ means that the feature x_{nd} was retained, and $m_{nd} = 0$ means that the feature x_{nd} was masked or zeroed.

$$\therefore E[M] = p$$

$$E[(1 - M)] = 1 - p$$

As we already know that::

$$E[x] = \sum_{n=1}^{N}(x_i.p(x_i))$$

in eq1 we have 3 Expactation term::
1.

$$E[Y^T.T] = Y^T.Y$$

because it is constant term which does not depends on M
2.

$$E[W^T(M.X)^TY] = W^TE[(M.X)^T].Y$$
$$=> E[W^T(M.X)^TY] = p.(W^TX^TY) + (1-p).0$$
$$=> E[W^T(M.X)^TY] = p.W^TX^TY$$

3.

$$E[W^T(M.X)^T(M.X)W] = W^TE[(M.X)^T(M.X)].W$$

$E[w^T(M*X)^T(M*X)w]$: This term involves $(M*X)^T(M*X)w.$ here we have two thing diagonal values and non diagonal values.In case of non diagonal values both random variable m_n, i and m_n, j are different so we get expactation independently as $p^2(X^T.X)$ and in case of diagonal values when i=j then both random variable are same that time expaction is p(X$^T.X$)

$$=> E[W^T(M.X)^T(M.X)W] = p^2W^T(X^T.X).W + p.W^T(X^T.X).W$$

Now, using eq1

$$L(w) = \arg\min_w(Y^T.Y - p.W^TX^TY + p^2W^T(X^T.X).W + p.W^T(X^T.X).W)$$

this can be written as::

$$L(w) = \arg\min_w \left\|Y - p \cdot W^T \cdot X\right\|^2 - p \cdot W^T \cdot (X^T \cdot X) \cdot W$$

So the final equation equivalent to equivalent to minimizing a regularized loss function where $\left\|Y - p \cdot W^T \cdot X\right\|^2$ is loss in prediction (like in linear regression) and $p \cdot W^T \cdot (X^T \cdot X) \cdot W$ might be a regularize term which regularize the weight vector

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Manish Agrawal
Roll Number: 231110028
Date: September 15, 2023

QUESTION 6

SOLUTION-6

Method-1. Accuracy in percentage is 46.89320388349515

Method-2. Accuracy in percentage is for λ 0.01 is 58.090614886731395.

Accuracy in percentage is for λ 0.1 is 59.54692556634305.

Accuracy in percentage is for λ 1.0 is 67.39482200647248.

Accuracy in percentage is for λ 10.0 is 73.28478964401295.

Accuracy in percentage is for λ 20.0 is 71.68284789644012.

Accuracy in percentage is for λ 50.0 is 65.08090614886731.

Accuracy in percentage is for λ 100.0 is 56.47249190938511.

The best accuracy is at $\lambda = 10.0$.