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

QUESTION 1

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: September 15, 2023

We can find optimal values of w_c and M_c, by finding the first-order derivative of our objective function with respect to w_c and M_c and equating it with zero.

Our objective function, as given in the question is

$$L_{(\hat{w}_c, \hat{M}_c)} = arg \min_{w_c, M_c} \sum_{(x_n:y_n=c)} \frac{1}{N_c} \left(x_n - w_c\right)^T M_c \left(x_n - w_c\right) - \log |M_c| \tag{1}$$

To obtain w_c, we need to partially differentiate the above objective function with respect to w_c and then equate to zero to get the optimal value of w_c,

$$\frac{\partial L}{\partial w_c} = -\frac{2}{N_c} \sum_{(x_n:y_n=c)} M_c(x_n - w_c) \tag{2}$$

Moving independent terms on the other side.

$$\sum_{(x_n:y_n=c)} (x_n - w_c) = 0$$

$$\sum_{(x_n:y_n=c)} x_n - \sum_{(x_n:y_n=c)} w_c = 0$$

$$\sum_{(x_n:y_n=c)} x_n = \sum_{(x_n:y_n=c)} w_c$$

$$\sum_{(x_n:y_n=c)} x_n = w_c N_c$$

$$w_c = \frac{1}{N_c} \sum_{(x_n:y_n=c)} x_n \tag{3}$$

Partially differentiating equation (1) with respect to M_c and then equating to zero to get optimal value of M_c,

$$\frac{\partial L}{\partial M_c} = \frac{1}{N_c} \sum_{(x_n : y_n = c)} (x_n - w_c)^T (x_n - w_c) - M_c^{-1}$$

$$\sum_{x_n, y_n = c} \left(\frac{1}{N_c} (x_n - w_c)(x_n - w_c)^T - (M_c^{-1}) \right) = 0$$

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

$$M_c = \frac{1}{N_c} \left(\sum_{x_n, y_n = c} (x_n - w_c)(x_n - w_c) \right)^{-1} \tag{4}$$

and if M_c is the Identity matrix,

$$argmin_{w_c, M_c} = \frac{1}{N_c} \sum_{x_n : y_n = c} (x_n - w_c)^T \mathbf{I} (x_n - w_c) - \log |\mathbf{I}|$$

$$w_c = argmin_{w_c} \frac{1}{N_c} \sum_{x_n, y_n = c} \|x_n - w_c\|_2^2$$

The above equation simply represents standard squared Euclidean distance.

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

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: September 15, 2023

QUESTION 2

Yes, **1NN**, will be consistent.
We know that there is an infinite number of training data with correct labels without any noise. Using 1NN, test data will always find the closest training data if the training data is infinite. Therefore we can always classify it without error in classification.

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

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: September 15, 2023

QUESTION 3

We can utilize variance as a node-selection feature in a decision tree for regression with real-valued labels. This measures how homogeneous each node's set of real-valued labels is.
You can accomplish this by following these steps.
The variance of the labels in a node should first be calculated. The variance of a node represents how much the labels inside it deviate from the mean. Low variance suggests homogeneity because it means the labels are near the mean.

$$\textbf{Weighted Variance Reduction} = \textbf{Variance}(P) - \frac{n_{\textbf{left}}}{n} \cdot \textbf{Variance}(\textbf{LeftChild}) + \frac{n_{\textbf{right}}}{n} \cdot \textbf{Variance}(\textbf{RightChild})$$

Now that the node variance has been determined, we independently determine the variance of the labels of the node's left and right children.
The characteristic that produces the lowest weighted average variance in child nodes is then selected.
The division that causes the smallest rise in variance is selected

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

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: September 15, 2023

QUESTION 4

In unregularized linear regression, the prediction for a test input x_* is given by,

$$f(x_*) = \hat{w}^\top x_* \tag{5}$$

where, $\hat{w} = (X^\top X)^{-1} X^\top y$.

Substituting $\hat{w}$ in the equation (5)

$$y_* = \left(\left(X^\top X\right)^{-1} X^\top y\right)^\top x_*$$

$$y_* = (X^\top y)^\top ((X^\top X)^{-1})^\top x_*$$

$$y_* = (X^\top y)^\top ((X^\top X)^\top)^{-1} x_*$$

$$y_* = y^\top X (X^\top X)^{-1} x_*$$

$$y_* = y^\top \hat{w}$$

$$y_* = \sum_{n=1}^{N} w_n y_n \tag{6}$$

In the given equation, each weight denoted as $\hat{w}_n$ signifies the significance of individual features in capturing the linear connection between features and the input data during the training process.

In the context of K-Nearest Neighbors (KNN), each weight denoted as w_n highlights the significance of nearby training examples when making predictions. This implies that closer training examples are accorded greater importance in the prediction process compared to those that are farther away.

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

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: September 15, 2023

QUESTION 5

The new loss function is defined as:

$$\sum_{n=1}^{N}(y_n - w^T\tilde{x})^2 \tag{7}$$

where $x_n = x_n \cdot m_n$ and m_n is a binary mask vector with $m_{nd} \sim$ Bernoulli(p).
finding the expected value of this new loss function:

$$E[P] = E\left[\sum_{n=1}^{N}(y_n - w^T\tilde{x})^2\right] \tag{8}$$

We can apply the linearity of expectation:

$$E[P] = \sum_{n=1}^{N} E\left[(y_n - w^T\tilde{x})^2\right]$$

focusing on the expectation of a single term in the summation:

$$E\left[(y_n - w^T\tilde{x})^2\right] = E\left[(y_n - w^T \cdot x_n m_n)^2\right]$$

Expanding the square and using the linearity of expectation:

$$E\left[(y_n - w^T \cdot x_n m_n)^2\right] = E\left[(y_n - w^T \cdot x_n m_n) \cdot (y_n - w^T \cdot x_n m_n)\right]$$

Now, consider that m_{nd} is a binary random variable with $E[m_{nd}] = p$:

$$E\left[(y_n - w^T \cdot x_n m_n) \cdot (y_n - w^T \cdot x_n m_n)\right] = E\left[y_n^2 - 2y_n w^T x_n m_n + (w^T \cdot x_n m_n)^2\right]$$

Now, focus on the middle term, and since m_{nd} is binary:

$$E\left[-2y_n w^T x_n m_n\right] = -2y_n w^T x_n E[m_n] = -2py_n w^T x_n$$

Now, we can rewrite the expected value of the single term in the summation as:

$$E\left[(y_n - w^T \cdot x_n m_n) \cdot (y_n - w^T \cdot x_n m_n)\right] = E\left[y_n^2 - 2py_n w^T x_n + (w^T \cdot x_n m_n)^2\right]$$

Now, substitute this back into the expression for the expected value of the new loss function:

$$E[P] = \sum_{n=1}^{N} E\left[y_n^2 - 2py_n w^T x_n + (w^T \cdot x_n m_n)^2\right] \tag{9}$$

Now, let's sum over all data points:

$$E[P] = \sum_{n=1}^{N} E\left[y_n^2 - 2py_n w^T x_n + (w^T \cdot x_n m_n)^2\right] \tag{10}$$

This expression is equivalent to minimizing the following regularized loss function:

$$L(w) = \sum_{n=1}^{N} \left(E[y_n^2] - 2py_n w^T x_n + p(w^T \cdot x_n)^2\right) \tag{11}$$

This regularised loss function has a regularisation term that penalizes the magnitude of w based on the L2-norm (squared magnitude). As a result, minimizing a regularised loss function using L2 regularisation is the same as minimizing the expected value of the new loss function.

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

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: September 15, 2023

QUESTION 6

Method 1:
The prediction accuracy of the model is: **46.89320388349515**
Method 2:
The prediction accuracy of the model at $\lambda = 0.01$ is: **58.090614886731395**
The prediction accuracy of the model at $\lambda = 0.1$ is: **59.54692556634305**
The prediction accuracy of the model at $\lambda = 1$ is: **67.39482200647248**
The prediction accuracy of the model at $\lambda = 10$ is: **73.28478964401295**
The prediction accuracy of the model at $\lambda = 20$ is: **71.68284789644012**
The prediction accuracy of the model at $\lambda = 50$ is: **65.08090614886731**
The prediction accuracy of the model at $\lambda = 100$ is: **56.47249190938511**

The best accuracy of the model is at lambda = 10