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

Student Name: Hemant Kumar
Roll Number: 210434
Date: September 15, 2023

QUESTION 1

My solution to problem 1

for calculating the optimal values, we need to calculate the derivative of the function w.r.t. w_c and M_c and equate them to zero

solving for w_c

$\frac{\partial}{\partial w_c}\left[\frac{1}{N_c}\sum_{x_n:y_n=c}(x_n-w_c)^\top M_c(x_n-w_c)-\log(|M_c|)\right]=0$
$\frac{1}{N_c}\sum_{x_n:y_n=c}(-2)(M_c(x_n-w_c))=0$

after solving the value of w_c is $\frac{1}{Nc}\sum_{x_n:y_n=c}x_n$

now similarly solving for M_c

$\frac{\partial}{\partial M_c}\left[\frac{1}{N_c}\sum_{x_n:y_n=c}(x_n-w_c)^\top M_c(x_n-w_c)-\log(|M_c|)\right]=0$

after solving the value of M_c is $\left[\frac{1}{Nc}\sum_{x_n:y_n=c}(x_n-w_c)(x_n-w_c)^\top)\right]^{-1}$

If M_c is an identity matrix, the loss function reduces to a simple quadratic term that measures the squared distance between data points x_n and the cluster center w_c for the class c. It's a loss in clustering algorithms, where the goal is to minimize the sum of squared distances between data points and their cluster centers.

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

Student Name: Hemant Kumar
Roll Number: 210434
Date: September 15, 2023

QUESTION

2

My solution to problem 2

Yes, one nearest neighbor will be consistent in this case. because, there is an infinite amount of training data and each of them is correctly labeled without any noise, which means whenever we get a test data point we can always find a training data point completely close to its probability of finding such a point will tend to be 1, when the amount of train data goes to infinity. Therefore, we can always classify them with no errors in classification. In short, we always have that test data in our training set which ensures Bayes optimal error rate.

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

Student Name: Hemant Kumar
Roll Number: 210434
Date: September 15, 2023

QUESTION 3

My solution to problem 3

When constructing Decision Trees for regression, we want to choose a feature to split on that will quantify the homogeneity of the set of real-valued labels at each node. Here is the criteria that I feel will qualify the constraint:

Variance Reduction: In this criteria, we will calculate the variance of the labels at each node and then select the feature that results in the greatest reduction in variance when split. split which will result in a more homogeneous subset of labels will result in a large reduction in variance.

$$\mathbf{Var}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - X_{mean})^2$$

X is a set of real-valued labels, where X_i represents the i-th label, and X_{mean} is the mean of the labels.

To calculate the variance reduction, we first compute the variance of the labels at the parent node, and then for each possible split, we compute the weighted average of the variances of the labels at the child nodes. The variance reduction is the difference between the parent node variance and the weighted average of the child node variances. The feature that results in the greatest variance reduction is chosen as the splitting feature.

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

Student Name: Hemant Kumar
Roll Number: 210434
Date: September 15, 2023

QUESTION

4

My solution to problem 4

Given, $\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$. So, the prediction at any test input $\mathbf{x}_*$ is given by $y_* = \hat{\mathbf{w}}^T\mathbf{x}_* = \mathbf{x}_*^T\hat{\mathbf{w}}$

Therefore,

$$y_* = \mathbf{x}_*^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$
$$\Rightarrow y_* = \mathbf{W}\mathbf{y}$$

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

Now,

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

So, w_n is the n^{th} element of the 1xN matrix $\mathbf{W}$. we know that $\mathbf{X}$ is a matrix whose rows are the N training vectors $\mathbf{x}_n$, w_n can be written in the following form:

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

Here, w_n depends on the input $\mathbf{x}_*$ and all the training data from $\mathbf{x}_1$ to $\mathbf{x}_n$. Since there exist X^TX term in the expression 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 the numerator in this given case, while for kNN it comes in the denominator. Another difference is that w_n are expressed as products of $\mathbf{x}_*$ while in kNN they are expressed as the sum in the denominator.

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

Student Name: Hemant Kumar
Roll Number: 210434
Date: September 15, 2023

QUESTION 5

My solution to problem 5

Considering linear regression model with minimizing the squared loss function

$$\sum_{n=1}^{N} \left(y_n - \mathbf{w^T x_n}\right)^2$$

masking off the features of $\mathbf{x_n}$ by replacing $\mathbf{x_n}$ with $\widetilde{\mathbf{x}}_\mathbf{n} = \mathbf{x_n} \circ \mathbf{m_n}$, where $\mathbf{m_n}$ denotes Dx1 binary mask vector, with $m_{nd} \sim \text{Bernoulli}(p)$

New loss function after masking inputs is $\sum_{n=1}^{N} \left(y_n - \mathbf{w^T}\widetilde{\mathbf{x}}_\mathbf{n}\right)^2$

Now, Calculating the Expected value of the new loss function

$$\begin{aligned}
E\left(\sum_{n=1}^{N} \left(y_n - \mathbf{w^T}\widetilde{\mathbf{x}}_\mathbf{n}\right)^2\right) &= \sum_{n=1}^{N} E\left[\left(y_n - \mathbf{w^T}\widetilde{\mathbf{x}}_\mathbf{n}\right)^2\right] \\
&= \sum_{n=1}^{N} E\left[y_n^2 - 2y_n\mathbf{w^T}\widetilde{\mathbf{x}}_\mathbf{n} + \left(\mathbf{w^T}\widetilde{\mathbf{x}}_\mathbf{n}\right)^2\right] \\
&= \sum_{n=1} y_n^2 - 2y_n E\left[\mathbf{w^T}\widetilde{\mathbf{x}}_\mathbf{n}\right] + E\left[\left(\mathbf{w^T}\widetilde{\mathbf{x}}_\mathbf{n}\right)^2\right] \\
&= \sum_{n=1} y_n^2 - 2y_n p\mathbf{w^T x_n} + E\left[\left(\mathbf{w^T}\widetilde{\mathbf{x}}_\mathbf{n}\right)^2\right] \\
&= \sum_{n=1} \left[\left(y_n - p\mathbf{w^T x_n}\right)^2\right] - \sum_{n=1}^{D} \left\{\left(p\mathbf{w^T x_n}\right)^2 + \sum_{i=1}^{D} E\left[\left(w_i\widetilde{x}_{ni}\right)^2\right] + \sum_{i\neq j} E\left[w_i x_{ni} w_j x_{nj}\right]\right\} \\
&= \sum_{n=1} \left[\left(y_n - p\mathbf{w^T x_n}\right)^2\right] - \sum_{i=1}^{D} \left\{p^2\left(\sum_{i=1}^{D} w_i^2 x_{ni}^2 + \sum_{i\neq j} w_i x_{ni} w_j x x_{nj}\right)\right. \\
&\left. + \sum_{i=1}^{D} w_i^2 x_{ni}^2 + p^2 \sum_{i\neq j} w_{i,j} x_{ni} w_j x x_{nj}\right\} \\
&= \sum_{n=1} \left(y_n - p\mathbf{w^T x_n}\right)^2 + pq \sum_{n=1}^{N}\sum_{i=1}^{D} w_i^2 x_{ni}^2 \\
&= \sum_{n=1} \left(y_n - p\mathbf{w^T x_n}\right)^2 + \sum_{i=1}^{D} w_i^2 C_i \quad \text{where } C_i \text{ is const. wrt } w
\end{aligned}$$

The equation comes out in the form of ridge regression. so minimizing the expected value of our new loss function is equivalent to minimizing the ridge regression.

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

Student Name: Hemant Kumar
Roll Number: 210434
Date: September 15, 2023

QUESTION 6

My solution to problem 6

Method 1

I attempted to implement the **Mahalanobis distance**

Here are the outcomes based on the varying number of iterations to optimize the theta:

Test accuracy for 10 iter. is: 47.86407766990292
Test accuracy for 15 iter. is: 48.64077669902913
Test accuracy for 20 iter. is: 49.385113268608414
Test accuracy for 25 iter. is: 49.95145631067961
Test accuracy for 30 iter. is: 50.06472491909385
Test accuracy for 35 iter. is: 50.2588996763754
Test accuracy for 40 iter. is: 50.30744336569579
Test accuracy for 45 iter. is: 49.75728155339806

Results indicate that increasing the number of iterations improved the model's test accuracy only up to a certain point.

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 test set accuracy.