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

Student Name: Ogirala Deeven Kumar
Roll Number: 210681
Date: January 16, 2024

QUESTION 1

To find the optimal values of $\mathbf{w}_c$ and $\mathbf{M}_c$ of the given objective function, we need to minimize the function with respect to $\mathbf{w}_c$ and $\mathbf{M}_c$.

The objective function,

$$f(\hat{\mathbf{w}}_c, \hat{\mathbf{M}}_c) = arg \min_{\mathbf{w}_c,\mathbf{M}_c} \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \frac{1}{N_c}(\mathbf{x}_n - \mathbf{w}_c)^T \mathbf{M}_c(\mathbf{x}_n - \mathbf{w}_c) - \log|\mathbf{M}_c|$$

Partial derivative with respect to $\mathbf{w}_c$,

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{w}_c} = 0 \Rightarrow & \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \frac{1}{N_c}(-2\mathbf{M}_c(\mathbf{x}_n - \mathbf{w}_c)) = 0 \\
\Rightarrow & (\sum_{\mathbf{x}_c:\mathbf{y}_c=c} \mathbf{x}_n - \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \mathbf{w}_c) = 0 \\
\Rightarrow & \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \mathbf{x}_n = N_c \mathbf{w}_c \\
\Rightarrow & \mathbf{w}_c = \frac{1}{N_c} \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \mathbf{x}_n \\
\Rightarrow & \mathbf{w}_c = \bar{\mathbf{x}}_c, \text{ where } \bar{\mathbf{x}}_c \text{ is average of } \mathbf{x}_n \text{ in class c}
\end{aligned}$$

Partial derivative with respect to $\mathbf{M}_c$,

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{M}_c} = 0 \Rightarrow & \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \frac{1}{N_c}(\mathbf{x}_n - \mathbf{w}_c)(\mathbf{x}_n - \mathbf{w}_c)^T - (\mathbf{M}_c^{-1})^T = 0 \\
\Rightarrow & (\mathbf{M}_c^{-1})^{-1} = \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \frac{1}{N_c}[(\mathbf{x}_n - \mathbf{w}_c)(\mathbf{x}_n - \mathbf{w}_c)^T]^{-1} \\
\Rightarrow & \mathbf{M}_c = \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \frac{1}{N_c}[(\mathbf{x}_n - \bar{\mathbf{x}}_c)(\mathbf{x}_n - \bar{\mathbf{x}}_c)^T]^{-1}
\end{aligned}$$

If $\mathbf{M}_c$ is an identity matrix, then the objective function is,

$$f(\hat{\mathbf{w}}_c) = arg \min_{\mathbf{w}_c} \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \frac{1}{N_c}(\mathbf{x}_n - \mathbf{w}_c)^T(\mathbf{x}_n - \mathbf{w}_c) = arg \min_{\mathbf{w}_c} \sum_{\mathbf{x}_c:\mathbf{y}_c=c} \frac{1}{N_c}||\mathbf{x}_n - \bar{\mathbf{x}}_c||^2$$

The given model is reduced to **k-means clustering** model when the matrix $\mathbf{M}_c$ is an identity matrix.

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

Student Name: Ogirala Deeven Kumar
Roll Number: 210681
Date: January 16, 2024

QUESTION 2

Yes, the one-nearest-neighbor algorithm is consistent in this case. Because, if there are infinite training data and has noise-free setting i.e., every training input is labeled correctly, then there exists at least one data which is nearest to the give testing data and the probability of finding this data is approximately equal to 1. Hence the classifier's Bayes optimal error rate will be almost zero.

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

Student Name: Ogirala Deeven Kumar
Roll Number: 210681
Date: January 16, 2024

QUESTION 3

In regression, where the labels are real-valued, a valid criterion to choose a feature to split on is the **variance reduction** or **mean squared error reduction**. Here we quantify the reduction in the variance of the real-valued label that results from the split.

The idea is to select a feature and a splitting threshold that minimizes the variance of the labels in the child nodes compared to the variance in the parent node.

Let the training data be $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$,

1. Calculate the variance of the labels in the parent node:
   $Variance_{parent} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})^2$, where $\bar{y}$ is the mean of the labels in the parent node.
2. For each feature and threshold, divide the data into two child nodes i.e., left and right.
3. Calculate the variance of the labels in the left and right nodes. (similar to that of the parent node)
4. Calculate the variance reduction for this:
   $$VarianceReduction = Variance_{parent} - (\frac{N_{left}}{N}\ Variance_{left} + \frac{N_{right}}{N}\ Variance_{right})$$
5. Choose the feature and threshold that maximize the variance reduction.
6. Repeat the process for left node and right node.

The intuition behind this criterion is that a good split is one that reduces the variability of the labels within each child node. By maximizing the variance reduction, we aim to find the split that best separates the data into groups with more similar real-valued labels.

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

QUESTION 4

Student Name: Ogirala Deeven Kumar
Roll Number: 210681
Date: January 16, 2024

Given, $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The prediction of at the test case $\mathbf{x}_*$ is

$$\begin{aligned} y_* &= \hat{\mathbf{w}}^T\mathbf{x}_* \\ &= \mathbf{x}_*^T\hat{\mathbf{w}} \\ &= \mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \Rightarrow y_* &= \mathbf{W}\mathbf{y} \end{aligned}$$

where, $\mathbf{W} = \mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. i.e. $\mathbf{W}$ is a $1 \times N$ matrix - has weights of N features, $\mathbf{W} = (w_1 w_2 w_3 ... w_N)$ and $\mathbf{y}$ is a $N \times 1$ column matrix - has the prediction values, $\mathbf{y} = (y_1 y_2 y_3 ... y_N)^T$.

Since,

$$\begin{aligned} y_* &= \mathbf{W}\mathbf{y} \\ \Rightarrow y_* &= \sum_{n=1}^{N} w_n y_n \end{aligned}$$

where, w_n is the weight of n^{th} feature and it can be written as $w_n = \mathbf{x}_*^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_n^T$, cause we just need the n^{th} row vector from matrix $\mathbf{X}$. Here w_n depends on all the training data x_1 to x_n because the expression has the term $(\mathbf{X}^T\mathbf{X})^{-1}$. This is different in KNN, where weights depend only on x_* and x_n. Here w_n is the product of x_* where as in KNN it is the summation of x_*.

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

Student Name: Ogirala Deeven Kumar
Roll Number: 210681
Date: January 16, 2024

QUESTION 5

Given, Loss function using masked input is

$$\begin{aligned}L(M) &= \sum_{n=1}^{N}(y_n - \mathbf{w}^T\tilde{x})^2\\ \Rightarrow L(M) &= ||\mathbf{y} - \tilde{\mathbf{X}}\mathbf{w}||^2\\ \Rightarrow L(M) &= ||\mathbf{y} - (\mathbf{R}\times\mathbf{X})\mathbf{w}||^2\end{aligned}$$

The input is dropped out such that any input dimension is retained with probability $\mathbf{p}$, then expected value of $L(M$ is $E_{R\ Bernoulli(n,p)}[L(M)]$. let it be $E[L(M)]$
Now the expectancy is minimised with respect to $\mathbf{w}$. The objective function of this is

$$\begin{aligned}L(\mathbf{w}) &= arg\min_{\mathbf{w}}\{E[||\mathbf{y} - (\mathbf{R}\times\mathbf{X})\mathbf{w}||^2]\}\\ \Rightarrow L(\mathbf{w}) &= arg\min_{\mathbf{w}}\{E[(\mathbf{y} - (\mathbf{R}\times\mathbf{X})\mathbf{w})^T(\mathbf{y} - (\mathbf{R}\times\mathbf{X})\mathbf{w})]\}\end{aligned}$$

Let $\mathbf{e} = \mathbf{y} - (\mathbf{R}\times\mathbf{X})\mathbf{w}$ then $\mu = E[\mathbf{e}] = \mathbf{y} - p\mathbf{X}\mathbf{w}$,

$$\begin{aligned}&\Rightarrow L(\mathbf{w} = arg\min_{\mathbf{w}}\{E[\mathbf{e}^T\mathbf{e}]\}\\ &\Rightarrow L(\mathbf{w} = arg\min_{\mathbf{w}}\{\mu^T\mu + TRACE[(\mathbf{k}-\mu)^T(\mathbf{k}-\mu)]\}\\ &\Rightarrow L(\mathbf{w} = arg\min_{\mathbf{w}}\{(\mathbf{y} - p\mathbf{X}\mathbf{w})^T(\mathbf{y} - p\mathbf{X}\mathbf{w}) + p(1-p)TRACE[\mathbf{X}\mathbf{w}\mathbf{w}^T\mathbf{X}^T]\}\\ &\Rightarrow L(\mathbf{w} = arg\min_{\mathbf{w}}\{||\mathbf{y} - p\mathbf{X}\mathbf{w}||^2 + p(1-p)||(\sqrt{diag\mathbf{X}^T\mathbf{X}})\mathbf{w}||^2\}\end{aligned}$$

Comparing $L(\mathbf{w})$ with ridge regression objective function,

$$\begin{aligned}L_{ridge}(\mathbf{w}) &= arg\min_{\mathbf{w}}\{(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w}\}\\ &= arg\min_{\mathbf{w}}\{||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 + \lambda||\mathbf{w}||^2\}\end{aligned}$$

Here we can observe that the objective function $L(w)$ is similar to that of ridge regression. So we can minimize a regularized loss function where the term $p(1-p)||\sqrt{diag\mathbf{X}^T\mathbf{X}})\mathbf{w}||^2$ is the regularizer and $||\mathbf{y} - p\mathbf{X}\mathbf{w}||^2$ is squared loss.

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

Student Name: Ogirala Deeven Kumar
Roll Number: 210681
Date: January 16, 2024

QUESTION 6

Method 1: The accuracy is 46.8932038835

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!