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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: September 15, 2023

QUESTION 1

The given optimization function for class c is

$$L_c(\boldsymbol{w}_c, \mathbf{M}_c) = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} (\boldsymbol{x}_n - \boldsymbol{w}_c)^\top \mathbf{M}_c (\boldsymbol{x}_n - \boldsymbol{w}_c) - \log |\mathbf{M}_c|$$

Since $\mathbf{M}_c$ is positive definite matrix, Hessian of $(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top \mathbf{M}_c (\boldsymbol{x}_n - \boldsymbol{w}_c)$ will be positive definite, which is a sufficient condition for strict convexity. Hence, we will use first order optimality on the objective function to obtain closed form optimal values of $\boldsymbol{w}_c$ and $\mathbf{M}_c$.

1. Finding $\hat{\boldsymbol{w}}_c$

Taking the gradient of $L_c(\boldsymbol{w}_c, \mathbf{M}_c)$ with respect to $\boldsymbol{w}_c$ and equating it to zero will give us $\hat{\boldsymbol{w}}_c$.

$$\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \boldsymbol{w}_c} = \frac{\partial}{\partial \boldsymbol{w}_c} \left(\sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} (\boldsymbol{x}_n - \boldsymbol{w}_c)^\top \mathbf{M}_c (\boldsymbol{x}_n - \boldsymbol{w}_c) - \log |\mathbf{M}_c| \right)$$

Using $\frac{\partial}{\partial \boldsymbol{s}} (\boldsymbol{x} - \boldsymbol{s})^\top \mathbf{A} (\boldsymbol{x} - \boldsymbol{s}) = -(\mathbf{A} + \mathbf{A}^\top)(\boldsymbol{x} - \boldsymbol{s})$,

$$\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \boldsymbol{w}_c} = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \frac{\partial}{\partial \boldsymbol{w}_c} \Big((\boldsymbol{x}_n - \boldsymbol{w}_c)^\top \mathbf{M}_c (\boldsymbol{x}_n - \boldsymbol{w}_c) \Big)$$

$$\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \boldsymbol{w}_c} = - \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \Big((\mathbf{M}_c + \mathbf{M}_c^\top)(\boldsymbol{x}_n - \boldsymbol{w}_c) \Big)$$

Now let us equate $\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \boldsymbol{w}_c}$ to 0

$$- \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \Big((\mathbf{M}_c + \mathbf{M}_c^\top)(\boldsymbol{x}_n - \boldsymbol{w}_c) \Big) = 0$$

($\mathbf{M}_c$ is positive definite $\iff$ $\mathbf{M}_c^\top$ is positive definite) $\implies$ $\mathbf{M}_c + \mathbf{M}_c^\top$ is positive definite. Since $\mathbf{M}_c + \mathbf{M}_c^\top$ is positive definite, the solution of the above equation will be:

$$\sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} (\boldsymbol{x}_n - \boldsymbol{w}_c) = 0$$

$$\sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \boldsymbol{w}_c = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \boldsymbol{x}_n$$

$$\frac{N_c}{N_c} \boldsymbol{w}_c = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \boldsymbol{x}_n$$

$$\hat{\boldsymbol{w}}_c = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \boldsymbol{x}_n$$

2. Finding $\hat{\mathbf{M}}_c$

Taking the gradient of $L_c(\boldsymbol{w}_c, \mathbf{M}_c)$ with respect to $\mathbf{M}_c$ and equating it to zero will give us $\hat{\mathbf{M}}_c$.

$$\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \mathbf{M}_c} = \frac{\partial}{\partial \mathbf{M}_c}\left(\sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top \mathbf{M}_c(\boldsymbol{x}_n - \boldsymbol{w}_c) - \log|\mathbf{M}_c|\right)$$

$$\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \mathbf{M}_c} = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}\frac{\partial}{\partial \mathbf{M}_c}\Big((\boldsymbol{x}_n - \boldsymbol{w}_c)^\top \mathbf{M}_c(\boldsymbol{x}_n - \boldsymbol{w}_c)\Big) - \frac{\partial}{\partial \mathbf{M}_c}\log|\mathbf{M}_c|$$

Using $\frac{\partial}{\partial \mathbf{A}}(\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{x}^\top$

$$\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \mathbf{M}_c} = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \boldsymbol{w}_c)(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top - \frac{\partial}{\partial \mathbf{M}_c}\log|\mathbf{M}_c|$$

$$\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \mathbf{M}_c} = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \boldsymbol{w}_c)(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top - \frac{1}{|\mathbf{M}_c|}\frac{\partial}{\partial \mathbf{M}_c}|\mathbf{M}_c|$$

Using $\frac{\partial}{\partial \mathbf{A}}|\mathbf{A}| = |\mathbf{A}|(\mathbf{A}^{-1})^\top$

$$\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \mathbf{M}_c} = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \boldsymbol{w}_c)(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top - \frac{1}{|\mathbf{M}_c|}|\mathbf{M}_c|(\mathbf{M}_c^{-1})^\top$$

$$\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \mathbf{M}_c} = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \boldsymbol{w}_c)(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top - (\mathbf{M}_c^{-1})^\top$$

Now let us equate $\frac{\partial L_c(\boldsymbol{w}_c, \mathbf{M}_c)}{\partial \mathbf{M}_c}$ to 0

$$\sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \boldsymbol{w}_c)(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top - (\mathbf{M}_c^{-1})^\top = 0$$

$$(\mathbf{M}_c^{-1})^\top = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \boldsymbol{w}_c)(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top$$

$$\mathbf{M}_c^{-1} = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \boldsymbol{w}_c)(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top$$

$$\mathbf{M}_c^{-1} = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \boldsymbol{w}_c)(\boldsymbol{x}_n - \boldsymbol{w}_c)^\top$$

$$\hat{\mathbf{M}}_c = \left(\sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}(\boldsymbol{x}_n - \hat{\boldsymbol{w}}_c)(\boldsymbol{x}_n - \hat{\boldsymbol{w}}_c)^\top\right)^{-1}$$

Substituting $\hat{\boldsymbol{w}}_c = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}\boldsymbol{x}_n$, we get

$$\hat{\mathbf{M}}_c = \left(\sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}\left(\boldsymbol{x}_n - \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}\boldsymbol{x}_n\right)\left(\boldsymbol{x}_n - \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c}\boldsymbol{x}_n\right)^\top\right)^{-1}$$

3. If $\mathbf{M}_c$ is an identity matrix

$$\mathbf{M}_c = \mathbf{I}$$

$$L_c(\boldsymbol{w}_c, \mathbf{M}_c) = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} (\boldsymbol{x}_n - \boldsymbol{w}_c)^\top \mathbf{M}_c (\boldsymbol{x}_n - \boldsymbol{w}_c) - \log |\mathbf{M}_c|$$

$$L_c(\boldsymbol{w}_c, \mathbf{M}_c = \mathbf{I}) = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} (\boldsymbol{x}_n - \boldsymbol{w}_c)^\top \mathbf{I} (\boldsymbol{x}_n - \boldsymbol{w}_c) - \log |\mathbf{I}|$$

$$L_c(\boldsymbol{w}_c, \mathbf{M}_c = \mathbf{I}) = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} (\boldsymbol{x}_n - \boldsymbol{w}_c)^\top (\boldsymbol{x}_n - \boldsymbol{w}_c)$$

$$L_c(\boldsymbol{w}_c, \mathbf{M}_c = \mathbf{I}) = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} (\boldsymbol{x}_n - \boldsymbol{w}_c)^\top (\boldsymbol{x}_n - \boldsymbol{w}_c)$$

$$L_c(\boldsymbol{w}_c, \mathbf{M}_c = \mathbf{I}) = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} (||\boldsymbol{x}_n - \boldsymbol{w}_c||)^2$$

Substituting optimal value of $\boldsymbol{w}_c$,

$$L_c(\boldsymbol{w}_c = \hat{\boldsymbol{w}}_c, \mathbf{M}_c = \mathbf{I}) = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \left(||\boldsymbol{x}_n - \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \boldsymbol{x}_n || \right)^2$$

$$L_c(\boldsymbol{w}_c = \hat{\boldsymbol{w}}_c, \mathbf{M}_c = \mathbf{I}) = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} ||\boldsymbol{x}_n - \boldsymbol{\mu}_c||^2$$

where $\boldsymbol{\mu}_c = \boldsymbol{w}_c = \sum_{\boldsymbol{x}_n : y_n = c} \frac{1}{N_c} \boldsymbol{x}_n$. Here $\boldsymbol{\mu}_c$ is centroid of class c inputs

In this special case, optimizing this loss function is equivalent to solving **LwP** as both measure the Euclidean distances from a centroid of the class. $\hat{\boldsymbol{w}}_c$ will not change in this case.

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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: September 15, 2023

QUESTION 2

In the given setting, one-nearest-neighbour algorithm will be **consistent**.

One-nearest-neighbour classifier has an error rate of at most twice of Bayes optimal error rate.

$$E_{Bayes} \leq E_{1NN} \leq 2E_{Bayes}(1 - E_{Bayes}) \leq 2E_{Bayes}$$

- Cover and Hart (1967)

As number of training examples approach infinity and since they are labelled correctly in the given setting, both E_{Bayes} and $2E_{Bayes}$ approach 0. This implies that E_{1NN} approaches 0, hence it is consistent in this setting where noise is absent. Note that it is not necessarily consistent when the noise is present.

Since there are infinite training examples, for every test input, we can always find a training example labelled correctly in close proximity of test input. Therefore, we can classify the test inputs with a minimal error rate. This error rate tends to 0 as the number of training examples approach infinity because test input will tend to coincide with its nearest training input.

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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: September 15, 2023

QUESTION 3

In Decision Tree Classifier, we decided the features for nodes based on the Information Gain. Information Gain was defined as difference in entropy before split and after the split of that node. The entropy captured the homogenity/diversity of the labels. If the split was such that more homogenous labels were grouped together, the purity was high and the entropy was low. If the split was such that more diverse labels were grouped together, the purity was low and the entropy was high.

Extending the analogy of entropy to regression, we want a metric that can measure the homogenity/diversity of a set of real valued labels. One such metric could be **variance**. If the variance of the labels grouped together is low, then their homogenity/purity is high. If the variance of the labels grouped together is high, then their diversity is high and purity is low.

Let the variance of labels of set S be $Var(S)$. Suppose we assign a feature F to a node and it splits S into smaller disjoint sets $S_1, S_2, ...S_k$. Here k is the total number of distinct values taken by feature F in training data inputs and S_i is the set of training data inputs whose feature F takes i^{th} distinct value.

Similar to Information Gain in classification, we can define Information Gain in regression as

$$IG_{Regression}(F) = Var(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} Var(S_i)$$

The feature F which gives highest $IG_{Regression}(F)$ will be chosen as a feature to split the data for the node at that level.

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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: September 15, 2023

QUESTION 4

Given a closed form solution $\hat{\boldsymbol{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$ for an unregularized linear regression, the predicted output $y_* = f(\boldsymbol{x}_*)$ for test input $\boldsymbol{x}_*$ can be obtained by:

$$f(\boldsymbol{x}_*) = \hat{\boldsymbol{w}}^\top\boldsymbol{x}_*$$

$$f(\boldsymbol{x}_*) = \langle\hat{\boldsymbol{w}}, \boldsymbol{x}_*\rangle = \langle\boldsymbol{x}_*, \hat{\boldsymbol{w}}\rangle$$

$$f(\boldsymbol{x}_*) = \boldsymbol{x}_*^\top\hat{\boldsymbol{w}}$$

$$f(\boldsymbol{x}_*) = \boldsymbol{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$$

Let $\boldsymbol{w}_* = \boldsymbol{x}_*^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$

$$f(\boldsymbol{x}_*) = \boldsymbol{w}_*\boldsymbol{y}$$

Note that $\boldsymbol{w}_*$ is an $1 \times N$ matrix as the dimensions of $\boldsymbol{x}_*^\top$ is $1 \times D$, $(\mathbf{X}^\top\mathbf{X})^{-1}$ is $D \times D$ and $\mathbf{X}^\top$ is $D \times N$. Dimension of $\boldsymbol{y}$ is $N \times 1$. Let $\boldsymbol{y} = [y_1, y_2...y_N]^\top$ and $\boldsymbol{w}_* = [w_1, w_2...w_N]$. Then we can write

$$f(\boldsymbol{x}_*) = \boldsymbol{w}_*\boldsymbol{y} = \sum_{n=1}^{N} w_n y_n$$

w_n in **Linear Regression**

$$w_n = \langle\boldsymbol{x}_*^\top, ((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)_n\rangle$$

where $((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)_n$ is n^{th} column of $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$

w_n measures the cosine similarity between the test input and n^{th} example in training data input with the weighted matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$

w_n in **KNN**

$$w_n = \frac{1}{d(\boldsymbol{x}_*, \boldsymbol{x}_n)} = \frac{1}{\sqrt{(\boldsymbol{x}_* - \boldsymbol{x}_n)^\top(\boldsymbol{x}_* - \boldsymbol{x}_n)}}$$

w_n in KNN too measures the similarity between the test input and n^{th} example in training data input, but this similarity is different than that of Linear Regression. Here it measures similarity based on Euclidean Distance.

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: September 15, 2023

QUESTION 5

The new loss function is

$$L(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2$$

where $\tilde{\boldsymbol{x}}_n = \boldsymbol{x}_n \circ \boldsymbol{m}_n$

Calculating Expectation of $\boldsymbol{m}_{nd}$ using the property of Bernoulli variable,

$$E(m_{nd}) = p$$

Calculating Expectation of $\tilde{x}_{nd}$,

$$\tilde{x}_{nd} = x_{nd} \cdot m_{nd}$$

$$E(\tilde{x}_{nd}) = E(x_{nd} \cdot m_{nd}) = x_{nd} \cdot E(m_{nd}) = p x_{nd}$$

Calculating Expectation of loss function,

$$E(L(\boldsymbol{w})) = E\left(\sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2\right)$$

By linearity of Expectation,

$$E(L(\boldsymbol{w})) = \sum_{n=1}^{N} E\left((y_n - \boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2\right)$$

$$E(L(\boldsymbol{w})) = \sum_{n=1}^{N} E(y_n^2 - 2y_n \boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n + (\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2)$$

$$E(L(\boldsymbol{w})) = \sum_{n=1}^{N} (E(y_n^2) - 2E(y_n \boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n) + E((\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2))$$

$$E(L(\boldsymbol{w})) = \sum_{n=1}^{N} (y_n^2 - 2y_n E(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n) + E((\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2))$$

Calculating Expectation of $\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n$,

$$E(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n) = E\left(\sum_{d=1}^{D} w_d \tilde{x}_{nd}\right) = \sum_{d=1}^{D} E(w_d \tilde{x}_{nd}) = \sum_{d=1}^{D} w_d E(\tilde{x}_{nd}) = \sum_{d=1}^{D} p w_d x_{nd} = p \sum_{d=1}^{D} w_d x_{nd} = p \boldsymbol{w}^\top \boldsymbol{x}_n$$

Calculating Expectation of $(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2$,

$$Var(z) = E(z^2) - (E(z))^2$$

$$E(z^2) = Var(z) + (E(z))^2$$
$$E((\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2) = Var(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n) + (E(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n))^2$$

Finding $Var(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)$,

$$Var(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n) = Var\left(\sum_{d=1}^{D} w_d \tilde{x}_{nd}\right)$$

Dimensions are independent (covariance is equal to 0). So linearity of variance holds.

$$Var(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n) = \sum_{d=1}^{D} Var(w_d \tilde{x}_{nd}) = \sum_{d=1}^{D} Var(w_d x_{nd} m_{nd}) = \sum_{d=1}^{D} w_d^2 x_{nd}^2 Var(m_{nd})$$

Variance of Bernoulli variable is $p(1-p)$

$$Var(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n) = p(1-p)\sum_{d=1}^{D} w_d^2 x_{nd}^2$$

Substituting,

$$E((\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2) = Var(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n) + (E(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n))^2$$

$$E((\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2) = p(1-p)\sum_{d=1}^{D} w_d^2 x_{nd}^2 + (p\boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

Substituting in $E(L(\boldsymbol{w}))$,

$$E(L(\boldsymbol{w})) = \sum_{n=1}^{N}(y_n^2 - 2y_n E(\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n) + E((\boldsymbol{w}^\top \tilde{\boldsymbol{x}}_n)^2))$$

$$E(L(\boldsymbol{w})) = \sum_{n=1}^{N}(y_n^2 - 2y_n p\boldsymbol{w}^\top \boldsymbol{x}_n + p(1-p)\sum_{d=1}^{D} w_d^2 x_{nd}^2 + (p\boldsymbol{w}^\top \boldsymbol{x}_n)^2)$$

$$E(L(\boldsymbol{w})) = \sum_{n=1}^{N}\left((y_n - p\boldsymbol{w}^\top \boldsymbol{x}_n)^2 + p(1-p)\sum_{d=1}^{D} w_d^2 x_{nd}^2\right)$$

$$E(L(\boldsymbol{w})) = \sum_{n=1}^{N}(y_n - p\boldsymbol{w}^\top \boldsymbol{x}_n)^2 + p(1-p)\sum_{n=1}^{N}\sum_{d=1}^{D} w_d^2 x_{nd}^2$$

$$E(L(\boldsymbol{w})) = L'(\boldsymbol{w}) = L'_{unreg}(\boldsymbol{w}) + L'_{reg}(\boldsymbol{w})$$

Here, we can see that expectation value of the given loss function is in the form of a regularized loss function. So minimizing this given loss function would be equivalent to minimizing the regularized loss function.

The new unregularized loss function is

$$L'_{unreg}(\boldsymbol{w}) = \sum_{n=1}^{N}(y_n - p\boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

And the regularizer function is

$$L'_{reg}(\boldsymbol{w}) = p(1-p)\sum_{n=1}^{N}\sum_{d=1}^{D} w_d^2 x_{nd}^2$$

$$L'_{reg}(\boldsymbol{w}) = \boldsymbol{w}^\top (p(1-p)\sum_{n=1}^{N} diag(\boldsymbol{x}_n \boldsymbol{x}_n^\top))\boldsymbol{w}$$

Let $\lambda = p(1-p)\sum_{n=1}^{N} diag(\boldsymbol{x}_n \boldsymbol{x}_n^\top)$

$$L'_{reg}(\boldsymbol{w}) = \lambda ||\boldsymbol{w}||^2$$

$$L'(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - p\boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \lambda ||\boldsymbol{w}||^2$$

where $\lambda = p(1-p)\sum_{n=1}^{N} diag(\boldsymbol{x}_n \boldsymbol{x}_n^\top)$

Let $b \sim Bernoulli(p)$

$$E(b) = 1 \cdot p + 0 \cdot (1-p) = p$$
$$E(b^2) = 1^2 \cdot p + 0^2 \cdot (1-p) = p$$
$$Var(b) = E(b^2) - (E(b))^2 = p - p^2 = p(1-p)$$

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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: September 15, 2023

QUESTION
6

.

Method 1

Classification Accuracy on Test Set = **46.89320388349515 %**

Method 2

Classification Accuracy on Test Set for Different Lambdas

λ	Accuracy
0.01	58.090614886731395 %
0.1	59.54692556634305 %
1	67.39482200647248 %
10	**73.28478964401295 %**
20	71.68284789644012 %
50	65.08090614886731 %
100	56.47249190938511 %

Maximum Accuracy (73.28478964401295 %) is given by $\lambda = 10$