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

Student Name: Havi Bohra
Roll Number: 210429
Date: September 15, 2023

QUESTION 1

Useful Properties from *matrix cookbook*:

1. $\frac{\delta}{\delta \mathbf{s}}(\mathbf{x}-\mathbf{s})^T\mathbf{W}(\mathbf{x}-\mathbf{s}) = -2\mathbf{W}(\mathbf{x}-\mathbf{s})$ (86)

2. $\frac{\delta \mathbf{a}^T\mathbf{X}\mathbf{a}}{\delta \mathbf{X}} = \mathbf{a}\mathbf{a}^T$ (72)

3. $\frac{\delta log(det(\mathbf{X}))}{\delta \mathbf{X}} = (\mathbf{X}^T)^{-1}$ (57)

Let S_c be the set $\{\mathbf{x}_n : y_n = c\}$

Take $\mathcal{L}(\mathbf{w}_c, \mathbf{M}_c) = \frac{1}{N_c}\sum_{\mathbf{x}_n \in S_c}((\mathbf{x}_n - \mathbf{w}_c)^T\mathbf{M}_c(\mathbf{x}_n - \mathbf{w}_c)) - log|\mathbf{M}_c|$

Using above properties we get,

$\frac{\delta \mathcal{L}(\mathbf{w}_c,\mathbf{M}_c)}{\delta \mathbf{w}_c} = -\frac{2}{N_c}\sum_{\mathbf{x}_n \in S_c}(\mathbf{M}_c(\mathbf{x}_n - \mathbf{w}_c))$ (i)

$\frac{\delta \mathcal{L}(\mathbf{w}_c,\mathbf{M}_c)}{\delta \mathbf{M}_c} = \frac{1}{N_c}\sum_{\mathbf{x}_n \in S_c}((\mathbf{x}_n - \mathbf{w}_c)(\mathbf{x}_n - \mathbf{w}_c)^T - (\mathbf{M}_c^T)^{-1})$ (ii)

Now equating (i) and (ii) with 0 to get optimal values,

$-\frac{2}{N_c}\sum_{\mathbf{x}_n \in S_c}(\mathbf{M}_c(\mathbf{x}_n - \mathbf{w}_c)) = 0 \implies \tilde{\mathbf{w}_c} = \frac{1}{N_c}\sum_{\mathbf{x}_n \in S_c}\mathbf{x}_n$ $\{\ \mathbf{M}_c \text{ is invertible as } \det(\mathbf{M}_c) > 0\ \}$

$\frac{1}{N_c}\sum_{\mathbf{x}_n \in S_c}((\mathbf{x}_n - \mathbf{w}_c)(\mathbf{x}_n - \mathbf{w}_c)^T - (\mathbf{M}_c^T)^{-1}) = 0 \implies \tilde{\mathbf{M}}_c = (\sum_{\mathbf{x}_n \in S_c}\frac{1}{N_c}(\mathbf{x}_n - \tilde{\mathbf{w}_c})(\mathbf{x}_n - \tilde{\mathbf{w}_c})^T)^{-1}$

Special Case:
when $\mathbf{M}_c$ is an identity matrix then $\mathcal{L}$ reduces to average of square of distances $\mathbf{w}_c$ from all points in class c which is minimum at mean of all points i.e. $\tilde{\mathbf{w}_c} = \frac{1}{N_c}\sum_{\mathbf{x}_n \in S_c}\mathbf{x}_n$, which is same as above, hence no speciality but we can see notice that $\mathcal{L}$ has become similar to that of LwP.

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

Student Name: Havi Bohra
Roll Number: 210429
Date: September 15, 2023

QUESTION 2

Yes! 1-NN is consistent.
In the view that we have infinite training data, a new test input will fall exactly on one of the training inputs and we should get the same label (i.e. it's correct label) as noise-free setting, if we use 1-NN approach. In other words, given a test data point, we can always find a training data point close to it, and probabilty of getting such point tends to 1 as training data tends to infinite. Therefore we can classify it with zero error.

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

Student Name: Havi Bohra
Roll Number: 210429
Date: September 15, 2023

QUESTION 3

We can use **Variance** as a criteria for splitting when performing regression with real-valued labels. Variance measures the spread or dispersion of the labels within a node. Lower variance implies higher homogeneity of the labels within the node.

For a node N that contains a set of examples with real-valued labels $y_1, y_2, ..., y_n$,
Variance(N) $= \frac{1}{n}\sum(y_i - \mu)^2$

n is the number of examples in the node N.
y_i is the actual label of the i-th example in the node N.
μ is the mean (average) of the labels in the node N, $\mu = \frac{1}{n}\sum y_i$

To choose the feature to split on in a decision tree for regression, we would select the feature that minimizes the variance the most when used as a splitting criterion. This means that when we split the data based on this feature, the resulting child nodes should have the lowest possible variance values. In other words, you want to reduce the spread of the labels within each child node, making the predictions more homogeneous.

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

Student Name: Havi Bohra
Roll Number: 210429
Date: September 15, 2023

QUESTION 4

Solution to regression equation gives, $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
Let $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, a D $\times$ N matrix
$\implies \hat{\mathbf{w}} = \mathbf{A}\mathbf{y} \implies \hat{w_i} = \sum_{j=1}^{N} A_{ij}y_j \ \ \forall i = 1, 2, ..., D$
Now prediction for test input $\mathbf{x}_*$ is
$\mathrm{f}(\mathbf{x}_*) = \hat{\mathbf{w}}^T\mathbf{x}_* = \sum_{i=1}^{D} \hat{w_i}x_{*i} = \sum_{i=1}^{D}\sum_{j=1}^{N} A_{ij}y_jx_{*i}$
$= \sum_{j=1}^{N}\sum_{i=1}^{D} A_{ij}x_{*i}y_j = \sum_{j=1}^{N}(\sum_{i=1}^{D} A_{ij}x_{*i})y_j$

$\implies \mathrm{f}(\mathbf{x}_*) = \sum_{j=1}^{N} w_j \cdot y_j$

where $w_j = \sum_{i=1}^{D} A_{ij}x_{*i} \ \ \forall j = 1, 2, ..., N; \ \ \mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

Analysis:
Now these new weights in the above expression are different from *KNN* in the following sense:
1. These weights depends on the matrix $\mathbf{A}$ which in turn depends on entire training input, which is in contrast to weighted - *KNN* where weights only depends on nearest neighbors
2. These A_{ij} becomes a kind of constant for a given data input, hence weights are having linear relationship with test input, whereas in weighted - *KNN* weights are inversely proportional to nearest neighbors which are not same for all points.

$\mathbf{E}(\sum_{n=1}^{N}(y_n - \mathbf{w}^T\tilde{\mathbf{x}_n})^2)$

Using Linearity of Expectation

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

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

$= \sum_{n=1}^{N}(\mathbf{E}(y_n^2) - 2y_n\mathbf{w}^T\mathbf{E}(\tilde{\mathbf{x}_n}) + (\sum_{i\neq j} w_iw_j\mathbf{E}(\tilde{x_{n,i}x_{n,j}}) + \sum_i w_i^2\mathbf{E}(\tilde{x_{n,i}^2})))$

$= \sum_{n=1}^{N}(y_n^2 - 2py_n\mathbf{w}^T\mathbf{x}_n + (\sum_{i\neq j} p^2w_iw_jx_{n,i}x_{n,j} + \sum_i pw_i^2x_{n,i}^2))$

$\{\mathbf{E}(Bernoulli(p)) = p;\ \mathbf{E}(\tilde{\mathbf{x}_n}) = p\mathbf{x}_n;\ \mathbf{E}(\tilde{x_{n,i}x_{n,j}}) = p^2x_{n,i}x_{n,j},\ if(i \neq j);\ \mathbf{E}(\tilde{x_{n,i}^2}) = px_{n,i}^2\}$

$= \sum_{n=1}^{N}(y_n^2 - 2py_n\mathbf{w}^T\mathbf{x}_n + \sum_{i,j} p^2w_iw_jx_{n,i}x_{n,j} + \sum_i(p-p^2)w_i^2x_{n,i}^2)$

$= \sum_{n=1}^{N}(y_n^2 - 2py_n\mathbf{w}^T\mathbf{x}_n + p^2(\mathbf{w}^T\mathbf{x}_n)^2) + \sum_{n=1}^{N}\sum_i(p-p^2)w_i^2x_{n,i}^2$

$= \sum_{n=1}^{N}(y_n - p\mathbf{w}^T\mathbf{x}_n)^2 + \sum_{n=1}^{N}\sum_i(p-p^2)w_i^2x_{n,i}^2$

Hence we got minimizing the expected value of this new loss function equivalent to minimizing a regularized loss function where
$\mathcal{L} = \sum_{n=1}^{N}(y_n - p\mathbf{w}^T\mathbf{x}_n)^2$ with the Regularization Term $\mathbf{R}(\mathbf{w}) = (p-p^2)\sum_{n=1}^{N}\sum_i w_i^2x_{n,i}^2$
(Note p(1-p) $\geq$ 0 as 0 $\leq$ p $\leq$ 1)

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

Student Name: Havi Bohra
Roll Number: 210429
Date: September 15, 2023

QUESTION 6

Method 1: Accuracy = 46.89%

Method 2: Table below contains accuracy acheived for various values of λ, clearly $\lambda = 10$ gives the best accuracy

λ	Accuracy (in %)
0.01	58.09
0.1	59.55
1	67.40
10	73.28
20	71.68
50	65.08
100	56.47