Rohan Mahendra Chaudhari

USCID : 6675-5653-85

1 Problem 1 - Nearest Neighbor Classification

1.1 What is the prediction for the test point star when K = 4? Explain why.

Solution

K = 4 indicates that we look for the 4 nearest neighbors to the test point star. Here, the test point star (x) would be predicted as a "Triangle" as per the given data. The main reason for this prediction is that as per the K nearest neighbor algorithm, every neighbor (x_n) votes for a particular label (y_n). Since K = 4 here, we only consider the 4 closest neighbors to the test point star (x) and are only concerned with the votes of those 4 points. Thus, looking at the data visually, it is evident that the 4 closest points to the test point star are - Square, Triangle, Triangle and Open Circle with respect to L2/Euclidean distance. Considering only these 4 points, naturally a particular point (x_n) is bound to vote for its own label (y_n) and thus on aggregating we get the majority as Triangle with a count of 2 and thus we can conclude that the test point star is predicted as a "Triangle" when K = 4.

1.2 What is the diamond classified as for K = N? Explain why.

Solution

K = N indicates that we consider the whole data set of N samples as the nearest neighbors of the test point diamond (x). Here, the test point diamond (x) would be predicted as "Triangle" as per the given data. The reason for this prediction is that here K=N where N is the sample size and thus while determining the label for the new test point diamond we must consider all the sample points. As per the K nearest neighbor algorithm it is clear that every sample point (x_n) will vote for a particular label (y_n) which would naturally be the label it belongs to and thus on aggregating these votes the label for the test point diamond (x) would be determined by which label gets the majority votes. Therefore, considering the entire data set, it is clear that the label "Triangle" would get the most votes as it occurs the most number of times and since K=N here we can conclude that the test point diamond would in-fact be classified as a "Triangle".

1.3 Suppose one performs leave-one-out validation (that is, N-fold cross validation) to choose the best hyper-parameter K. List the triangles that are correctly classified (as a validation point) in this process for the run with K = 1.

Solution

In leave-one-out validation we consider each data point out in turn, fit our model on the remaining, and then see how we did on the held out data point. After we've gone through this process for each possible data point, the final score is the proportion of data points that were classified correctly when held out.

Here k = 1, and thus we are required to look only for the nearest neighbor to the held out point i.e. only 1 point. Thus, according to the question we must find out how many Triangles were correctly classified. To do this, we consider each triangle individually and check if its nearest neighbor is also a triangle or not. If yes, then it has been correctly classified else not.

1. Triangle at (1,1) (approximately)
   Since, the Open Circle is the closest point and not another Triangle, we can conclude that the Triangle at position (1,1) is wrongly classified.

2. Triangle at (2,4) (approximately)
   Since, the Open Circle is the closest point and not another Triangle, we can conclude that the Triangle at position (2,4) is wrongly classified.

3. Triangle at (3,2) (approximately)
   Since, the triangle is the closest point to the triangle, we can conclude that the triangle at the position (3,2) is correctly classified.

4. Triangle at (3,2.5) (approximately)
   Since, the triangle is the closest point to the triangle, we can conclude that the triangle at the position (3,2.5) is correctly classified.

5. Triangle at (5,1) (approximately)
   Since, the Square is the closest point and not another Triangle, we can conclude that the Triangle at position (5,1) is wrongly classified.

6. Triangle at (5,3) (approximately)
   Since, the Open Circle is the closest point and not another Triangle, we can conclude that the Triangle at position (5,3) is wrongly classified.

Therefore, 2 out of the 5 triangles are classified correctly which are the ones present at (3,2) and (3,2.5).

2 Problem 2 Linear Regression

2.1 Find the closed form of $\boldsymbol{w}_*^{'}$

Solution

Given equation $\boldsymbol{w}_*^{'} = argmin_{w \in R^D}\ \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$
On differentiating the above equation and equating the differential to 0 we get,

$$2(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}) + (M + M^T)\boldsymbol{w} = 0$$
$$2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + (M + M^T)\boldsymbol{w} = 2\boldsymbol{X}^T\boldsymbol{y}$$
$$(2\boldsymbol{X}^T\boldsymbol{X} + (M + M^T))\boldsymbol{w} = 2\boldsymbol{X}^T\boldsymbol{y}$$

Thus the closed form of $\boldsymbol{w}_*^{'}$ will be,

$$\boldsymbol{w}^* = (2\boldsymbol{X}^T\boldsymbol{X} + (M + M^T))^{-1}2\boldsymbol{X}^T\boldsymbol{y}$$

Considering the given statement in the question (L2 regularization is clearly a special case with M being the identity matrix scaled by λ), on differentiating we get,

$$2(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{y}) + 2\lambda\boldsymbol{w} = 0$$
$$2(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} + \lambda\boldsymbol{w}) = 2\boldsymbol{X}^T\boldsymbol{y}$$
$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{w} = \boldsymbol{X}^T\boldsymbol{y}$$

Thus another representation of the closed form of $\boldsymbol{w}_*^{'}$ will be,

$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{y})$$

2.2

2.2.1 Assume σ is fixed and given, find the maximum likelihood estimation for w_*. In other words, first write down the probability of seeing the outcomes $y_1, ..., y_N$ given $x_1, ..., x_N$ as a function of the value of w_*; then find the value of w_* that maximizes this probability. You can assume X^TX is invertible where X is the data matrix as used in Problem 2.1.

Log Likelihood,
LL = $\sum_{n=1}^{N}$ log($\frac{1}{\sigma\sqrt{2\pi}}$ exp($\frac{-(y-\boldsymbol{w}_*^T\boldsymbol{x}_n)^2}{2\sigma^2}$))
LL = $\sum_{n=1}^{N}$ ((log(1) - log($\sqrt{2\pi\sigma^2}$) + log(exp($\frac{-(y-\boldsymbol{w}_*^T\boldsymbol{x}_n)^2}{2\sigma^2}$))
LL = $\sum_{n=1}^{N}$ (-log ($\sqrt{2\pi\sigma^2}$)) + ($\frac{-(y-\boldsymbol{w}_*^T\boldsymbol{x}_n)^2}{2\sigma^2}$)
LL = -N log($\sqrt{2\pi}\sigma$) - $\frac{1}{2\sigma^2}$ $\sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2$

MLE for $\boldsymbol{w}_*$
Evaluate $\frac{dLL}{dw_*} = 0$
$\frac{dLL}{dw_*} = 0 + \frac{d}{dw_*}($ -$\frac{1}{2\sigma^2}$ $\sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2)$

$\frac{dLL}{dw_*} = 0 + \sum_{n=1}^{N} \frac{d}{dw_*}(-\frac{1}{2\sigma^2}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2)$
$\frac{dLL}{dw_*} = -\frac{1}{2\sigma^2}\sum_{n=1}^{N} \frac{d}{dw*}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2$
$0 = -\frac{1}{2\sigma^2}\sum_{n=1}^{N} 2(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)(-\boldsymbol{x}_n)$
$0 = \sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)(-\boldsymbol{x}_n)$
$0 = \sum_{n=1}^{N}(-y\boldsymbol{x}_n + \boldsymbol{w}_*^T\boldsymbol{x}_n^2)$
$\sum_{n=1}^{N} y\boldsymbol{x}_n = \sum_{n=1}^{N} \boldsymbol{w}_*^T\boldsymbol{x}_n^2$
$\sum_{n=1}^{N} y\boldsymbol{x}_n = \boldsymbol{w}_*^T \sum_{n=1}^{N} \boldsymbol{x}_n^2$

$\boldsymbol{w}_*^T = \frac{\sum_{n=1}^{N} y\boldsymbol{x}_n}{\sum_{n=1}^{N} \boldsymbol{x}_n^2}$

2.2.2 Now consider σ as a parameter of the probabilistic model too, that is, the model is specified by both w_* and σ. Find the maximum likelihood estimation for w_* and σ.

Log Likelihood,
LL = $\sum_{n=1}^{N}$ log($\frac{1}{\sigma\sqrt{2\pi}}$ exp($\frac{-(y-\boldsymbol{w}_*^T\boldsymbol{x}_n)^2}{2\sigma^2}$))
LL = $\sum_{n=1}^{N}$ ((log(1) - log($\sqrt{2\pi\sigma^2}$) + log(exp($\frac{-(y-\boldsymbol{w}_*^T\boldsymbol{x}_n)^2}{2\sigma^2}$))
LL = $\sum_{n=1}^{N}$ (-log ($\sqrt{2\pi\sigma^2}$)) + ($\frac{-(y-\boldsymbol{w}_*^T\boldsymbol{x}_n)^2}{2\sigma^2}$)
LL = -N log($\sqrt{2\pi}\sigma$) - $\frac{1}{2\sigma^2}$ $\sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2$

MLE for $\boldsymbol{w}_*$
Evaluate $\frac{dLL}{dw_*} = 0$
$\frac{dLL}{dw_*} = 0 + \frac{d}{dw_*}($ -$\frac{1}{2\sigma^2}$ $\sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2)$
$\frac{dLL}{dw_*} = 0 + \sum_{n=1}^{N} \frac{d}{dw_*}(-\frac{1}{2\sigma^2}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2)$
$\frac{dLL}{dw_*} = -\frac{1}{2\sigma^2}\sum_{n=1}^{N} \frac{d}{dw*}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2$
$0 = -\frac{1}{2\sigma^2}\sum_{n=1}^{N} 2(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)(-\boldsymbol{x}_n)$
$0 = \sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)(-\boldsymbol{x}_n)$
$0 = \sum_{n=1}^{N}(-y\boldsymbol{x}_n + \boldsymbol{w}_*^T\boldsymbol{x}_n^2)$
$\sum_{n=1}^{N} y\boldsymbol{x}_n = \sum_{n=1}^{N} \boldsymbol{w}_*^T\boldsymbol{x}_n^2$
$\sum_{n=1}^{N} y\boldsymbol{x}_n = \boldsymbol{w}_*^T \sum_{n=1}^{N} \boldsymbol{x}_n^2$

$\boldsymbol{w}_*^T = \frac{\sum_{n=1}^{N} y\boldsymbol{x}_n}{\sum_{n=1}^{N} \boldsymbol{x}_n^2}$

MLE for σ
Evaluate $\frac{dLL}{d\sigma} = 0$
$\frac{dLL}{d\sigma} = \frac{-N\sqrt{2\pi}}{\sqrt{2\pi}\sigma} - (\frac{-2}{2\sigma^3})\sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2$
$\frac{dLL}{d\sigma} = \frac{-N}{\sigma} + (\frac{1}{\sigma^3})\sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2$
$0 = \frac{-N}{\sigma} + (\frac{1}{\sigma^3})\sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2$
$\frac{N}{\sigma} = (\frac{1}{\sigma^3})\sum_{n=1}^{N}(y - \boldsymbol{w}_*^T\boldsymbol{x}_n)^2$

$\frac{N\sigma^3}{\sigma} = \sum_{n=1}^{N}(y - \boldsymbol{w}_*^T \boldsymbol{x}_n)^2$
$N\sigma^2 = \sum_{n=1}^{N}(y - \boldsymbol{w}_*^T \boldsymbol{x}_n)^2$
$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(y - \boldsymbol{w}_*^T \boldsymbol{x}_n)^2$
Substitute $\boldsymbol{w}_*^T$ in the above equation,
$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(y - \frac{\sum_{n=1}^{N} y\boldsymbol{x}_n}{\sum_{n=1}^{N} \boldsymbol{x}_n^2}\boldsymbol{x}_n)^2$

$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y - \frac{\sum_{n=1}^{N} y\boldsymbol{x}_n}{\sum_{n=1}^{N} \boldsymbol{x}_n^2}\boldsymbol{x}_n)^2}$

Thus,
$\boldsymbol{w}_{*\,MLE}^T = \frac{\sum_{n=1}^{N} y\boldsymbol{x}_n}{\sum_{n=1}^{N} \boldsymbol{x}_n^2}$

$\sigma_{MLE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y - \frac{\sum_{n=1}^{N} y\boldsymbol{x}_n}{\sum_{n=1}^{N} \boldsymbol{x}_n^2}\boldsymbol{x}_n)^2}$

3 Problem 3 Convergence of Perceptron Algorithm

3.1 Show that if the algorithm makes a mistake, the update rule moves the weight $\boldsymbol{w}_k$ towards the direction of the optimal weights $\boldsymbol{w}_{opt}$. Specifically, suppose in iteration k we have $y_i \neq$ sign$(\boldsymbol{w}_k^T \boldsymbol{x}_i)$. Prove

$$\boldsymbol{w}_{k+1}^T \boldsymbol{w}_{opt} \geq \boldsymbol{w}_k^T \boldsymbol{w}_{opt} + \gamma$$

Solution
As per the question, we must consider the expression:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + y_i \, \boldsymbol{x}_i$$

Consider $(\boldsymbol{w}_{k+1}$ - $\boldsymbol{w}_k)^T \boldsymbol{w}_{opt}$ and consider the property of $\boldsymbol{w}_{opt}$

$$\boldsymbol{w}_{k+1}^T \boldsymbol{w}_{opt} = \boldsymbol{w}_k^T \boldsymbol{w}_{opt} + y_i \, \boldsymbol{x}_i^T \boldsymbol{w}_{opt}$$

$y_i \, \boldsymbol{x}_i^T \boldsymbol{w}_{opt} = | \, \boldsymbol{x}_i^T \boldsymbol{w}_{opt} \, |$, $\boldsymbol{w}_{opt}$ perfectly classifies all the N data points and thus perfectly separates the data. Thus by definition of γ,

$$\boldsymbol{w}_{k+1}^T \boldsymbol{w}_{opt} \geq \boldsymbol{w}_k^T \boldsymbol{w}_{opt} + \gamma \|\boldsymbol{w}_{opt}\|$$

And as we know from the question that $\|\boldsymbol{w}_{opt}\| = 1$
Thus the final equation would be, $\boldsymbol{w}_{k+1}^T \boldsymbol{w}_{opt} \geq \boldsymbol{w}_k^T \boldsymbol{w}_{opt} + \gamma$

3.2 Show that the length of the weight vector does not increase by a large amount when the algorithm makes a mistake. More specifically, if in iteration k we have $y_i \neq$ sign$(\boldsymbol{w}_k^T \boldsymbol{x}_i)$, then

$$\|\boldsymbol{w}_{k+1}\|^2 \leq \|\boldsymbol{w}_k\|^2 + 1$$

Solution

Consider the LHS of the given equation $\|\boldsymbol{w}_{k+1}\|^2$ and substitute $\boldsymbol{w}_{k+1}$

$\|\boldsymbol{w}_{k+1}\|^2 = \boldsymbol{w}_{k+1}^T\ \boldsymbol{w}_{k+1} = (\boldsymbol{w}_k + y_i\ \boldsymbol{x}_i)^T\ (\boldsymbol{w}_k + y_i\ \boldsymbol{x}_i) = \|\boldsymbol{w}_k\|^2 + 2y_i\ \boldsymbol{w}_k^T\ \boldsymbol{x}_i + {y_i}^2\ \boldsymbol{x}_i^T\ \boldsymbol{x}_i$

Input $\boldsymbol{x}_i$ has norm 1 and the algorithm has made a mistake so $y_i \neq \text{sign}(\boldsymbol{w}_k^T \boldsymbol{x}_i)$

$\|\boldsymbol{w}_{k+1}\|^2 = \|\boldsymbol{w}_k\|^2 + 2y_i \boldsymbol{w}_k^T\ \boldsymbol{x}_i + {y_i}^2\ \boldsymbol{x}_i^T\ \boldsymbol{x}_i \leq \|\boldsymbol{w}_k\|^2 + 1$

3.3 Using results from Problem 3.1 and 3.2, show that for any iteration k+1,with M being the total number of mistakes the algorithm has made so far for the first k iterations, we have

$$\text{M}\gamma \leq \|\boldsymbol{w}_{k+1}\| \leq \sqrt{M}$$

Solution

By repeatedly applying results from Problem 3.1 all the mistakes and summing them up

$\boldsymbol{w}_{k+1}^T\ \boldsymbol{w}_{opt} \geq \boldsymbol{w}_1^T \boldsymbol{w}_{opt} + M\gamma\|\boldsymbol{w}_{opt}\|$

From the Algorithm given, it is safe to assume that the $\boldsymbol{w}_1$ would be 0.

$\boldsymbol{w}_{k+1}^T\ \boldsymbol{w}_{opt} \geq M\gamma\|\boldsymbol{w}_{opt}\|$

On using the given hint $\boldsymbol{a}^T \boldsymbol{b} \leq \|\boldsymbol{a}\|\ \|\boldsymbol{b}\|$ for any 2 vectors $\boldsymbol{a}$ and $\boldsymbol{b}$

$M\gamma\|\boldsymbol{w}_{opt}\| \leq \boldsymbol{w}_{k+1}^T\ \boldsymbol{w}_{opt} \leq \|\boldsymbol{w}_{k+1}\|\ \|\boldsymbol{w}_{opt}\|$

$M\gamma \leq \|\boldsymbol{w}_{k+1}\|$

Similarly, we use the results of problem 3.2 repeatedly and sum them up as well to get,

$\|\boldsymbol{w}_{k+1}\|^2 \leq \|\boldsymbol{w}_1\|^2 + M$

Since it is given that $\boldsymbol{w}_1$ is 0, we get

$\|\boldsymbol{w}_{k+1}\|^2 \leq \text{M}$

$\|\boldsymbol{w}_{k+1}\| \leq \sqrt{M}$

Therefore from $M\gamma \leq \|\boldsymbol{w}_{k+1}\|$ and $\|\boldsymbol{w}_{k+1}\| \leq \sqrt{M}$ we can conclude that,

$M\gamma \leq \|\boldsymbol{w}_{k+1}\| \leq \sqrt{M}$

3.4 Using result of Problem 3.3,conclude M $\leq \gamma^{-2}$

Solution

From Problem 3.3 we get $M\gamma \leq \|\boldsymbol{w}_{k+1}\| \leq \sqrt{M}$

Thus considering only the terms with M and γ, we get

$\text{M}\gamma \leq \sqrt{M}$

Divide $\sqrt{M}$ on both sides and flip the inequality, we get

$\text{M}\gamma/\sqrt{M} \geq 1$

Now we again divide the above equation by γ and flip the inequality yet again, we get

$\text{M}/\sqrt{M} \leq 1/\gamma$

On simplifying, we get

$\sqrt{M} \leq \gamma^{-1}$

We then square on both sides to get,
M $\leq \gamma^{-2}$ which is the required conclusion.