Assignment 4

Foundations of Machine Learning

IIT-Hyderabad

Sept-Dec 2021

Submitted by:

Ankita Jain

BM21MTECH14001

1. Non-Uniform Weights in Linear Regression

1. Given: A dataset : (x_n, t_n) ; $n = 0, 1, \ldots N$

Non negative weighting factor $g_n > 0$ for each data point

Resultant error function:

$$E_D(w) = \frac{1}{2} \sum_{n=1}^{N} g_n (t_n - w^T \phi(x_n))^2 \quad -(1)$$

$\phi(\cdot) \rightarrow$ any representation of data

(a) To find w^* that minimizes $E_D(w)$

Soln Let's first evaluate the derivative of the above error function $-(1)$ w.r.t. w

$$\frac{\partial E_D(w)}{\partial w} = \frac{1}{2} \sum_{n=1}^{N} g_n (t_n - w^T \phi(x_n)) \cdot 2 \cdot (-\phi^T(x_n))$$

$\boxed{\phi w^T \phi(x) = \phi^T(x) w}$ Hence

Equating the derivative to 0,

$$\sum_{n=1}^{N} g_n (t_n - w^T \phi(x_n)) \, \phi^T(x_n) = 0$$

Substituting

$$\sum_{n=1}^{N} g_n t_n \phi^T(x_n) - \sum_{n=1}^{N} g_n w^T \phi(x_n) \phi^T(x_n) = 0$$

$$\phi = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots \\ \phi_{21} & & \\ \vdots & & \end{bmatrix}$$

$t = [t_1 \cdots t_n]$

Substituting $\sqrt{g_n}\, \phi(x_n) = \phi'(x_n)$ & $\sqrt{g_n}\, t_n = t_n'$ (feature mapping)

$$\sum_{n=1}^{N} t_n' \phi'(x_n)^T - w^T \sum_{n=1}^{N} \phi'(x_n) \phi'(x_n)^T = 0$$

(Putting the above in vectorized form)

$$\phi'(x_n)^T t - w^T \phi'(x_n)^T \phi'(x_n) = 0$$

$$\boxed{\therefore w = (\phi^T \phi)^{-1} \phi^T t}$$

(b) To find two alternatives of the above weighted sum of squares error function in terms of

(i) Data dependent noise variance

Data dependent noise variance is indicated by β in the following equation:

$$\boxed{\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - w_{ML}^T\phi(x_n)\}^2} \quad (2)$$

also known as maximizing the log likelihood function w.r.t. noise parameter β

Now, if we substitute $\beta_{ML} = g_n\beta^{-1}$;

our eqn (2) becomes:

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum g_n\{t_n - w\phi(x_n)\}$$

Hence we see this is one of the interpretation in terms of data independent noise variance

(b) Replicated Data Points

the variable g_n which has been termed as a non negative weight for an individual data point (x_n, t_n) can also be the number of times a single data point has been replicated i.e the effective number of times the data point has been repeated ~~its sam~~ can be treated the same as a weight attached to a specific data point

2. Bayes Optimal Classifier:

2. Given 5 hypothesis that could guide a robot to move either Forward (F), Left (L) or Right (R)

$P(h_i \mid D)$	$P(F \mid h_i)$	$P(L \mid h_i)$	$P(R \mid h_i)$
0.4	1	0	0
0.2	0	1	0
0.1	0	0	1
0.1	0	1	0
0.2	0	1	0

MAP Estimate:

MAP hypothesis is defined as follows:

$$h_{MAP} \equiv \boxed{\underset{h \in H}{\operatorname{argmax}} P(h \mid D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D \mid h) P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D \mid h) P(h)$$

where
$D \rightarrow$ data
$h \rightarrow$ hypothesis

From the table above

$$= \underset{h \in H}{\operatorname{argmax}} (0.4, 0.2, 0.1, 0.1, 0.2)$$

$$= h \rightarrow h_1 \rightarrow \boxed{\text{Robot should move Forward}}$$

Bayes Optimal Estimate

Bayes Optimal Classification is defined as:

$$h_{BO} \equiv \underset{h \in H}{\operatorname{argmax}} \sum_{h \in H} P(v_j|h_i)\, P(h_i|D)$$

where V → set of possible classification of the examples

From the table ~~provided~~ provided:

$$\sum_{h \in H} P(F|h_i)\, P(h_i|D) = \overset{h_1}{(1 * 0.4)} = 0.4$$

$$\sum_{h \in H} P(L|h_i)\, P(h_i|D) = \overset{h_2}{(1 * 0.2)} + \overset{h_4}{(1 * 0.1)} + \overset{h_5}{(1 * 0.2)} = 0.5$$

$$\sum_{h \in H} P(R|h_i)\, P(h_i|D) = \overset{h_3}{(1 * 0.1)} = 0.1$$

$\therefore\ h_{BO} = \operatorname{argmax}\,(0.4, 0.5, 0.1)$

Bayes Optimal recommends the robot to move Left

As we see both recommendations are not same

No other classification method using the same hypothesis space & prior knowledge can outperform Bayes Optimal Classifier on the average

3. VC-Dimension:

3. Given: 1D data x where Hypothesis H is paremetrized by $\{p, q\}$ &

$$h_{p,q} = \begin{cases} 1 & ; \text{iff } p < x \leq q \\ 0 & ; \text{otherwise} \end{cases}$$

To find VC dimension of $\mathcal{H}$

Soln Let's assume ~~two~~ VC = 2
then for any 2 values, H must be able to shatter this subset of x

$x = \{x_1, x_2\}$ where $p < x_1, x_2 \leq q$

possible dichotomies

	Label			
x_1	0	0	1	1
x_2	0	1	0	1

$\because$ In this all $x_1, x_2 < q$ &

$x_1, x_2 > p$

the H can shatter one subset of x Hence VC dimension is atleast 2

Now, let's assume VC = 3

then H must be able to shatter any one subset of x (x_1, x_2, x_3) to prove the above.

Labels

x_1	0	0	0	0	1	1	1	1
x_2	0	0	1	1	0	0	1	1
x_3	0	1	0	1	0	1	0	1

∵ the condition for hypothesis to be true, $p \leq x < q$

i.e if we assume, $x_1 < x_2 < x_3$

then any $\{x_1, x_3\}$ will always contain $\underline{\underline{x_2}}$, hence not all the dichotomies can be realized as marked above

$$\boxed{\text{Hence the VC dimension for the given hypothesis is } 2}$$

4. Regularizer:

4. Given: D dimensional data $\{x_1, \dots, x_D\}$

Linear model: $y(x,\omega) = \omega_0 + \sum_{k=1}^{D} \omega_k x_k$

N such data samples

MSE: $E(\omega) = \frac{1}{2} \sum_{i=1}^{N} \left(y(x_i,\omega) - t_i\right)^2$

Suppose Gaussian Noise $N(0,\sigma^2)$ is added to each input variable x_k

To find: Relation b/w minimizing MSE averaged over noisy data & minimizing standard MSE averaged over noise free data with a L_2 regularization term, with bias w_0 omitted

Solution: Let the model with noisy inputs be defined as:

$$\boxed{\hat{y}'(x,\omega) = w_0 + \sum_{k=1}^{D} w_k (x_{nk} + e_{nk})}$$

where $e_{nk} \to$ noise added to input except bias $\to N(0,\sigma^2)$

Reducing the above eqn

$$\hat{y}(x,\omega) = w_0 + \sum_{k=1}^{D} \omega_k x_{nk} + \sum_{k=1}^{D} e_{nk}$$

$$\hat{y}(x,\omega) = y(x,\omega) + \sum_{k=1}^{D} e_{nk}$$

Substituting this in MSE for noisy input:

$$\hat{E} = \frac{1}{2} \sum_{n=1}^{N} (\hat{y}_n - t_n)^2$$

$$= \frac{1}{2} \sum_{n=1}^{N} \left\{(\hat{y}_n)^2 - 2y_n t_n + (t_n)^2\right\}$$

$$\hat{E} = \frac{1}{2}\sum_{n=1}^{N}\left\{\left(y_n + \sum_{k=1}^{D} w_k e_{nk}\right)^2 - 2\left(y_n + \sum_{k=1}^{D} w_k e_{nk}\right)t_n + t_n^2\right\}$$

$$\hat{E} = \frac{1}{2}\sum_{n=1}^{N}\left\{ y_n^2 + 2y_n\sum_{k=1}^{D} w_k e_{nk} + \left(\sum_{k=1}^{D} w_k e_{nk}\right)^2 - 2y_n t_n - 2t_n\sum_{k=1}^{D} w_k e_{nk} + t_n^2\right\}$$

Finding Averaged MSE

$$\mathbb{E}[\hat{E}] = E\left\{\frac{1}{2}\sum_{n=1}^{N}\left(y_n^2 - 2y_n t_n + t_n^2\right)\right\} + \cancel{\mathbb{E}\left\{\sum_{n=1}^{N}\left(2y_n\sum_{k=1}^{D} w_k e_k\right)\right\}} \to 0 \quad \because e_k \text{ mean} = 0$$

$$- \cancel{E\left\{\frac{1}{2}\sum_{k=1}^{N} 2t_n\sum_{k=1}^{D} w_k e_{nk}\right\}} \to 0 \quad \because e_k \text{ mean} = 0$$

$$+ \mathbb{E}\left\{\frac{1}{2}\sum_{k=1}^{D}\left(\sum_{k=1}^{D} w_k e_k\right)^2\right\}$$

$$= \mathbb{E}\{E\} + \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}\left\{\left(\sum_{k=1}^{D} w_k e_{wk}\right)^2\right\}$$

$$= \mathbb{E}\{E\} + \frac{1}{2}\sum_{k=1}^{N}\sum_{k=1}^{D} w_k^2 \underbrace{E\left(e_{wk}^2\right)}_{\sigma^2}$$

$$= \mathbb{E}\{E\} + \frac{1}{2}\sigma^2\sum_{k=1}^{D} w_k^2$$

$\therefore$ the rel b/w noisy input MSE & std MSE is:

$$\boxed{\mathbb{E}[\hat{E}] = \frac{1}{2}\sum_{n=1}^{N}\left(y(x_n, w) - t_n\right)^2 + \frac{\sigma^2}{2}\sum_{k=1}^{D} w_k^2}$$

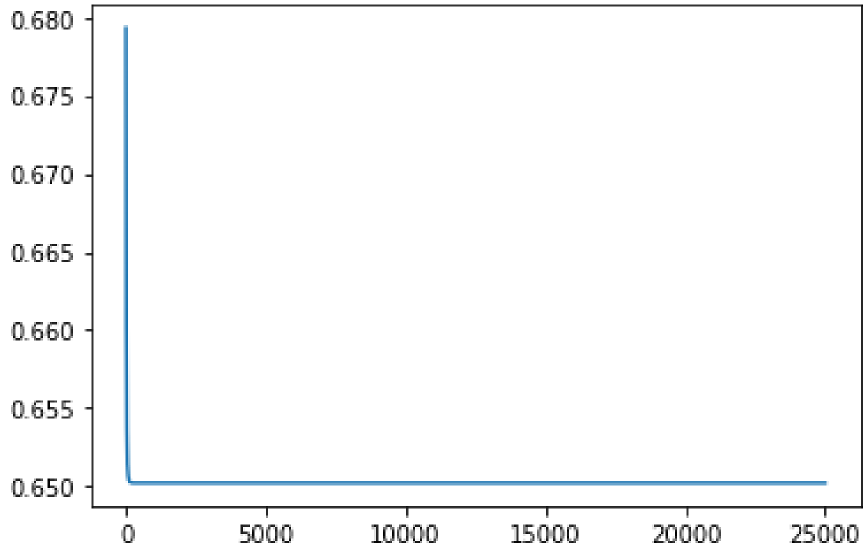
where $\hat{E}$ is MSE due to noisy input

1st term is due to std MSE

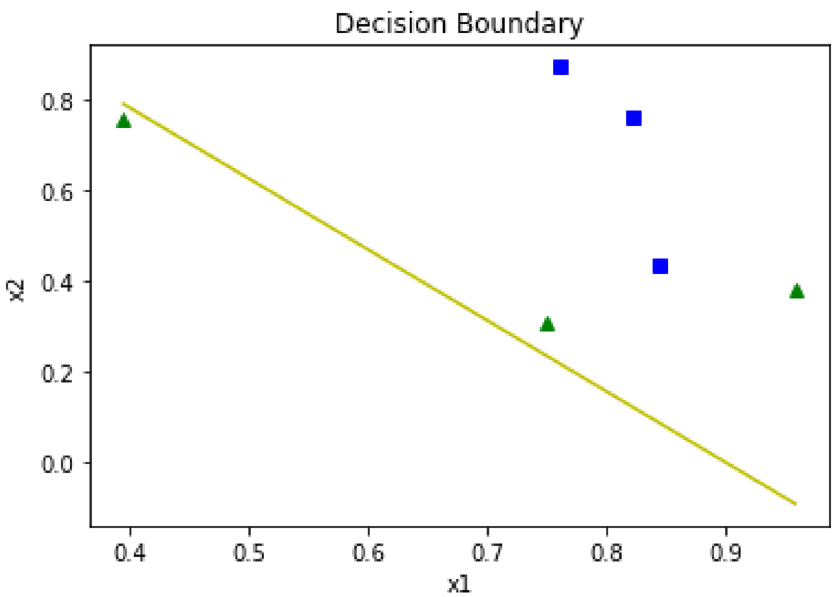
2nd term Regularizer Input

5. Logistic Regression:

```
Cost after 0 epoch is:  0.6793333930895056
Solution to 5b(i)
The updated value of w,b at the end of the epoch is:
 [1.45104757 0.49269336] -1.0073819919017042


 Solution to 5b(ii)
The logistic model P(y=1|x1,x2) is:
 0.577540312892111
Corresponding cross entropy function =
 0.6770734435545663
```



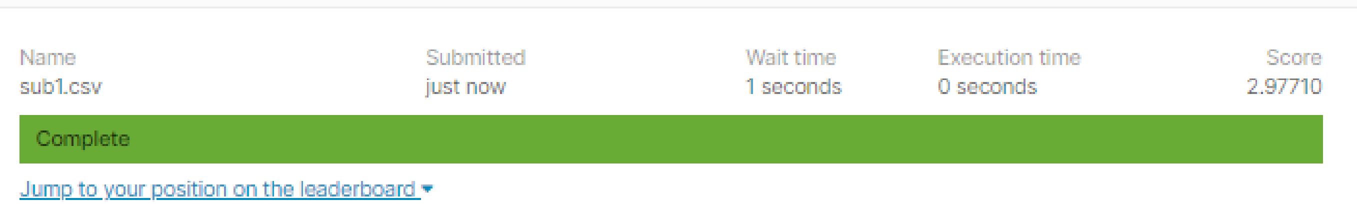

```
The solution to 5b(iii)

Accuracy:  66.66666666666667
Precision:  60.0
Recall:   100.0
```

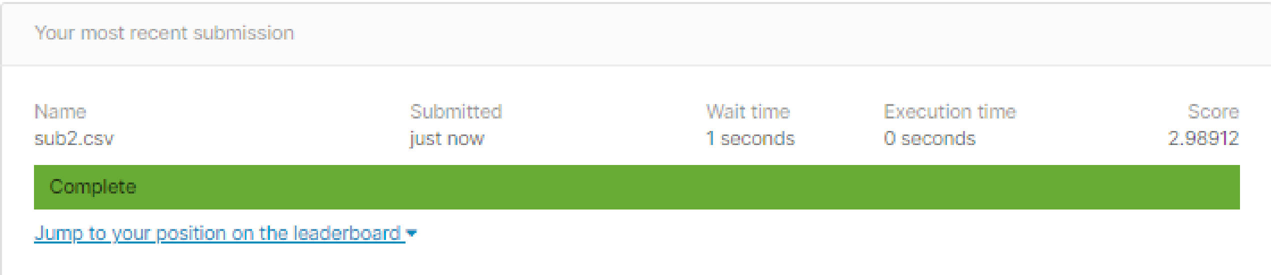


6. Kaggle - Taxi Fare Prediction:

The following two models performed best.

- An ensemble of kNN and lightGBM
- LightGBM model





A brief on the models selected and why they performed better.

1. LightGBM

   Faster training speed and higher efficiency: Light GBM use histogram based algorithm i.e binning information Lower memory usage: Replaces continuous values to discrete bins which result in lower memory.

2. Ensemble

   There are two main reasons to use an ensemble over a single model, and they are related; they are: Performance: An ensemble can make better predictions and achieve better performance than any single contributing model. Robustness: An ensemble reduces the spread or dispersion of the predictions and model performance