TRIBHUVAN UNIVERSITY

SCHOOL OF MATHEMATICAL SCIENCES

Balkhu, Kathmandu

Internal Assessment Answer Sheet

28

Level : Masters

Semaster : Second

Name : Arpan Sapkota

Roll No. : 07

Subject : Applied Machine Learning

Date : 2080 / 06 / 20

Marks Obtained	
Q. No. 1	
Q. No. 2	
Q. No. 3	
Q. No. 4	
Q. No. 5	
Q. No. 6	
Q. No. 7	
Q. No. 8	
Q. No. 9	
Q. No. 10	
TOTAL	

Group 'A'

2.

Batch Gradient Descent	Stochastic Gradient Descent
(i) In Batch gradient descent, sample of training data are taken.	(i) In Stochastic Gradient Descent individual data are taken.
(ii) It gives the average effect.	(ii) It gives the individual effect.
(iii) It converges slowly.	(iii) Convergence is faster as compared to Batch Gradient Descent.
(iv) Gives Optimal solution.	(iv) Give solution but not optimal.
(v) It uses, $w_j = w_j + \alpha(y + h(x))x_j$	(v) It uses: $B_j = B_j + \alpha(y - h_\beta(x))x_j$

3. Precision is the one of the evalution metric for the classification task, where it gives the ratio of positive prediction over the positive values.

let us consider a confusion matrix.

		Actual	
		T	F
Predicted	T	TP	FP
	F	FN	TN

from the above confusion matrix,

$$\text{Precision} = \frac{TP}{TP+FP}.$$

Hence, Precision gives the true positive value to sum of True positive and false positive ratio. It give more insight to the Algorithm since, accuracy will not be efficient alone to evaluate the classification models.

Similarly,

$$\text{Recall} = \frac{TP}{TP+FN}.$$

Recall is also another evaluation metrics for the classification algorithm, where it gives the ratio of True positive value to the sum of true positive and false negative value. Also, Recall gives the sensitivity of the classification algorithm.

4. K-means clustering is the Unsupervised Machine learning Algorithms, in which the k-numbers of clusters will be formed from the dataset which undergoes the k-means clustering.

For k-means clustering, we have the following procedure :.

1. Initialize the initial clusters $c_1, c_2, \ldots c_k$, where k is the number of clusters.

2. Get the distance of each data point with the clusters $c_1, c_2, \ldots c_k$
Here, we can use any distance metrics like :-
   - Euclidean distance $\Rightarrow d = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$ (mostly used)
   - Manhattan distance
   - Cosine Similarity measure
   - Jaccard distance, etc.

3. Compare the distance with the clusters $c_1, c_2, \ldots c_k$ and get the minimum distance from them in which the cluster lie between $c_1, c_2 \ldots c_k$

4. Repeat the step 3 for all the datapoints.

5. From the new assigned clusters, get the centers by taking the averages, which will be the next clusters for next iteration.
i.e. $c_1', c_2', \ldots c_k'$.

6. Repeat the process until we get the datapoints in the same clusters repeatedly.

7. Finally, we get the clusters :-
$C^A = \{d_1, d_2, \ldots d_m\}$.
$C^B = \{d_1', d_2', \ldots d_m'\}$.
$\vdots$
$C^k = \{d_1'', d_2'', \ldots d_m''\}$.
Compute the center of the obtained clusters.

1. Supervised learning is a costly approach compared to unsupervised learning. In Supervised learning, there are labelled data which undergoes the model training process. In Unsupervised learning there are unlabelled data which undergoes the model training process. During this process, supervision is required for the labelled data, where in Unsupervised learning, supervision is not required. Due to this, the computational cost for training the Unsupervised machine learning algorithms will be lower as compared to the supervised Machine learning algorithm. Extra resources like memory, processor, etc are required in order to train the supervised machine learning models. In Unsupervised machine learning there the model itself learns the pattern and performs the prediction, No any extra input should be given except the training data. However in supervised Machine learning Algorithm, we have to give the labelled data and supervise the model to learn the pattern as per the given data.

Thus, supervised machine learning is costly approach compared to the unsupervised machine learning.

5. Logestic Regression and Softmax regression both are the regression algorithm. where In Logestic regression we perform the regression of binary class i.e. (two classes). where as In Softmax regression we perform the regression of multi class i.e. greater than two.

In logestic regression,

$$h(x) = \frac{1}{1+e^{-z}}$$; sigmoid function

is used.

But In Softmax regression.

$$h(x) = \frac{e^z}{\sum e^z}$$; Softmax function

is used.

In logestic regression, we get the values in between 0 & 1 where as in softmax regression, we get the value between positive infinity and negative infinity.

Example:

If we want to do the ~~class~~ decision in two class like Yes & No, True or false then we use logestic regression.

If we want to get the decision in multiclass value like between 1 to 100 or more for any kind of prediction, then we use Softmax regression.

6. Logestic Regression for classification tasks.

In the linear regression, we use, $h(x) = \hat{y} = w_1x_1 + w_2x_2 + \ldots + w_nx_n = w^Tx$. But in logestic regression we use probability P.

Where $x_1, x_2 \ldots x_n$ are the inputs and $w_1, w_2 \ldots w_n$ are the weights.

While using the probability values, our range is limited to 0 and 1., so, we use odds,

i.e.

$$\text{odds} = \frac{p}{1-p}.$$

It is nothing but the ratio of probability to the probability value which is complementing value to the P. Here, the range will be 0 to positive infinite, so we are still restricted to the range.

To extend the range we take log,

$\log\left(\frac{p}{1-p}\right)$, So, our range becomes positive infinite to negative infinite.

Now,

$$\log\left(\frac{p}{1-p}\right) = w^Tx.$$

Taking exponential on both sides, we get.

$$\exp\left(\log\left(\frac{p}{1-p}\right)\right) = \exp(w^Tx)$$

$$\Rightarrow \frac{p}{1-p} = \exp(w^Tx)$$

$\Rightarrow p = \exp(w^T x) - p.\exp(w^T x).$

$\Rightarrow p + p\exp(w^T x) = \exp(w^T x)$

$\Rightarrow p(1 + \exp(w^T x)) = \exp(w^T x)$

$\Rightarrow p = \dfrac{\exp(w^T x)}{1 + \exp(w^T x)}$

Dividing by $\exp(w^T x)$

$\Rightarrow p = \dfrac{\exp(w^T x) / \exp(w^T x)}{(1 + \exp(w^T x)) / \exp(w^T x)}$

$= \dfrac{1}{\frac{1}{\exp(w^T x)} + \frac{\exp(w^T x)}{\exp(w^T x)}}$

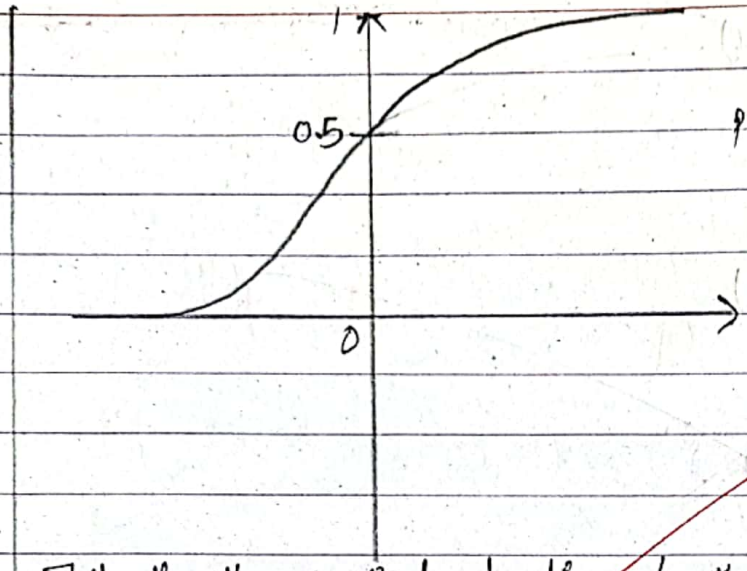
$= \dfrac{1}{\exp(-w^T x) + 1.}$

$= \dfrac{1}{1 + \exp(-w^T x)}$

$\therefore p = \dfrac{1}{1 + \exp(-z)}$

Where $z = w^T x$

Here, $f = \dfrac{1}{1+\exp(-z)}$ is the sigmoid function.



This is the required function for the logestic regression for classification task. It will takes the values between $-\infty$ to $+\infty$ and classify between the 0 and 1 as shown in the graph above.

7. The various fam g evaluation matrices used for regression are:

(i) Mean Square Error (MSE)
(ii) Mean ~~Absolute~~ Error (MAE)
(iii) Root Mean Square Error (RMSE)
(iv) R-square Error (R-error)

(i) Mean Square Error (MSE)

In the regression algorithm, we have,

error = $Y_{actual} - Y_{predicted}$.

The sum g total error, $e = e_1 + e_2 + \ldots + e_n$.

Mean Square Error is given by,

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (\hat{y} - y)^2$$

where, m is the number of data.
$\hat{y}$ = predicted value.
y = actual value.

It is the sum of squared of the errors.

(ii) Mean Absolute Error (MAE)

In MSE, we take square, but taking square is not only a approach, we can take the absolute value for predicted and actual value i.e. $|\hat{y} - y|$

This will also help us to avoid the error nullification problem.

So,

Mean Absolute Error (MAE) = $\frac{1}{m} \sum_{i=1}^{m} |\hat{y} - y|$

(iii) Root Mean Square Error (RMSE).

In this root mean square error, we take the root value of Mean Square Error (MSE). This is done to scale out the error,

i.e.

$$RMSE = \sqrt{\frac{1}{2m} \sum_{i=1}^{m} (\hat{y} - y)^2}$$

In order to scale out the scaling factor $1/2$, root is taken.

(11) R-Square Error (R^2- Error)

In this error, variance of model and the variance of the average is taken.

i.e.

$$R^2\text{- Error} = 1 - \frac{\text{Variance(model)}}{\text{Variance(average)}}$$

where, $\text{Variance(model)} = \sum_{i=1}^{m} (\hat{y} - y)$

and, $\text{variance(average)} = \sum_{i=1}^{m} (\bar{y} - y)$

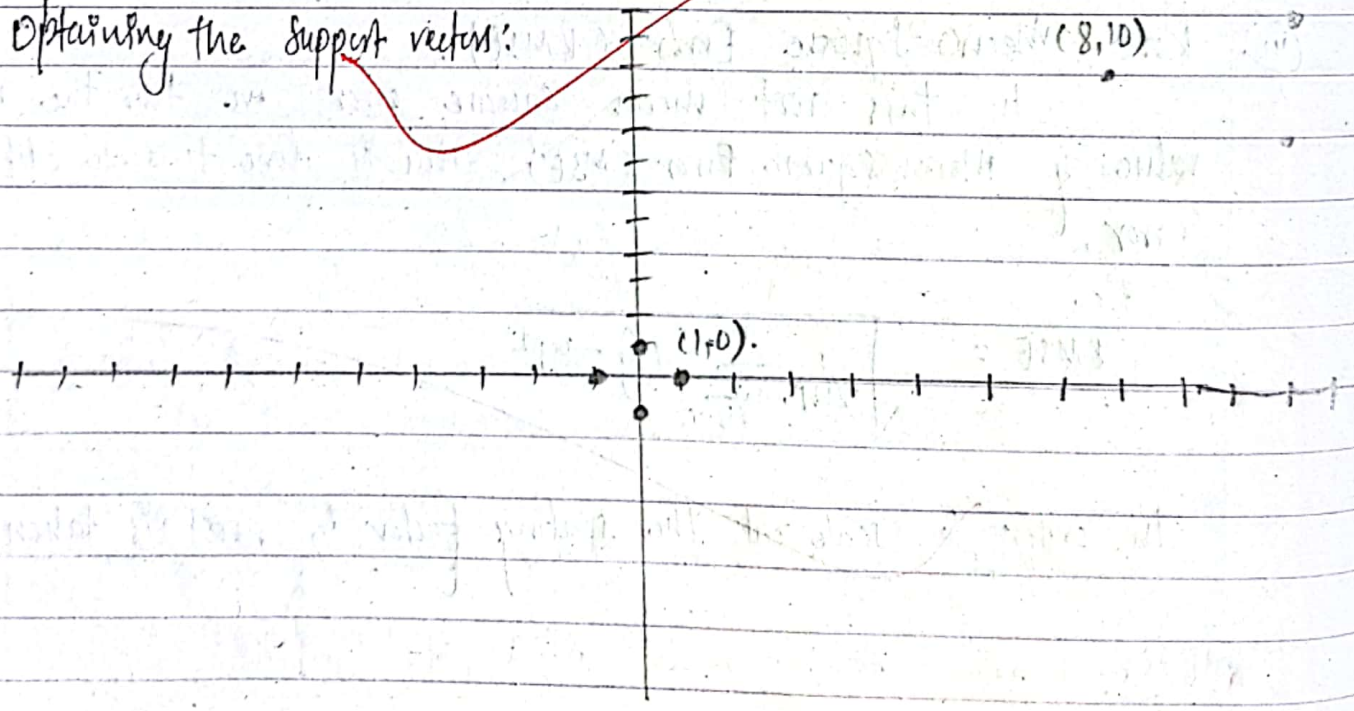
---

8. Given,

Negatively labelled data: (1,0), (0,1), (-1,0), (0,-1)

positively labelled data: (8,10), (10,8), (12,10), (10,12)

Obtaining the support vectors:



From the graph, we can see, the two support vectors.

$S_1 : (1,0)$ ⇒ Negatively Labelled.

$S_2 : (8,10)$ ⇒ positively Labelled.

We take two variables i.e. α, β then, the equation becomes,

$$\alpha S_1 S_1 + \beta S_1 S_2 = -1$$
$$\alpha S_2 S_1 + \beta S_2 S_2 = 1.$$

Adding the bias 1 to the support vectors,

i.e. $S_1 = (1,0,1)$

$S_2 = (8, 10, 1)$.

Performing the dot product for the linear kernel SVM,

$S_1 S_2 = (1,0,1).(8,10,1) = 8+0+1 = 9$

$S_1 S_1 = (1,0,1).(1,0,1) = 1+1 = 2.$

$S_2 . S_2 = (8,10,1).(8,10,1) = 64+100+1 = 165$

Now the equation becomes,

$$2\alpha + 9\beta = -1$$
$$9\alpha + 165\beta = 1$$

On solving, we get

$\alpha = -0.69$

$\beta = 0.044$

Now, the weight becomes.

$w = \alpha S_1 + \beta S_2$

$= -0.69 (1,0,1) + 0.044(8,10,1)$

$w = (-0.338, 0.44, -0.646)$

Here,

$w_1 = -0.338$, $w_2 = 0.44$ & $w_0 = -0.646$.

Now,

$y = w_1x_1 + w_2x_2 + w_0$.

$y = -0.338x_1 + 0.44x_2 + (-0.646)$. required hyperplane equation.

In order to predict the class (8,6).

$y = -0.338 \times 8 + 0.44 \times 6 - 0.646$.

$y = -0.71$.

Thus, it falls under negatively labelled data.

10.

Points.	1	2	3	4	5.
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	(2)	5.	0

point 3 & 5 ; i.e the minimum so, we combine it as cluster

Points.	1	2	{3,5}	4.
1	0			
2	9	0		
{3,5}	(3)	7	0	
4	6	5	9	0.

point {3,5} and 1 is the minimum, ; we combine it as cluster

So,

Point	{1,3,5}	2	4
{1,3,5}	0		
2	9	0	
4	6	(5)	0

Point 2 4 4 are minimum so, we combine it as cluster.

Point	{1,3,5}	{2,4}
{1,3,5}	0	
{2,4}	(6)	0

Finally,

Point	{1,3,5,2,4}
{1,3,5,2,4}	0

Hence the cluster formed.

We draw the dendogram as:

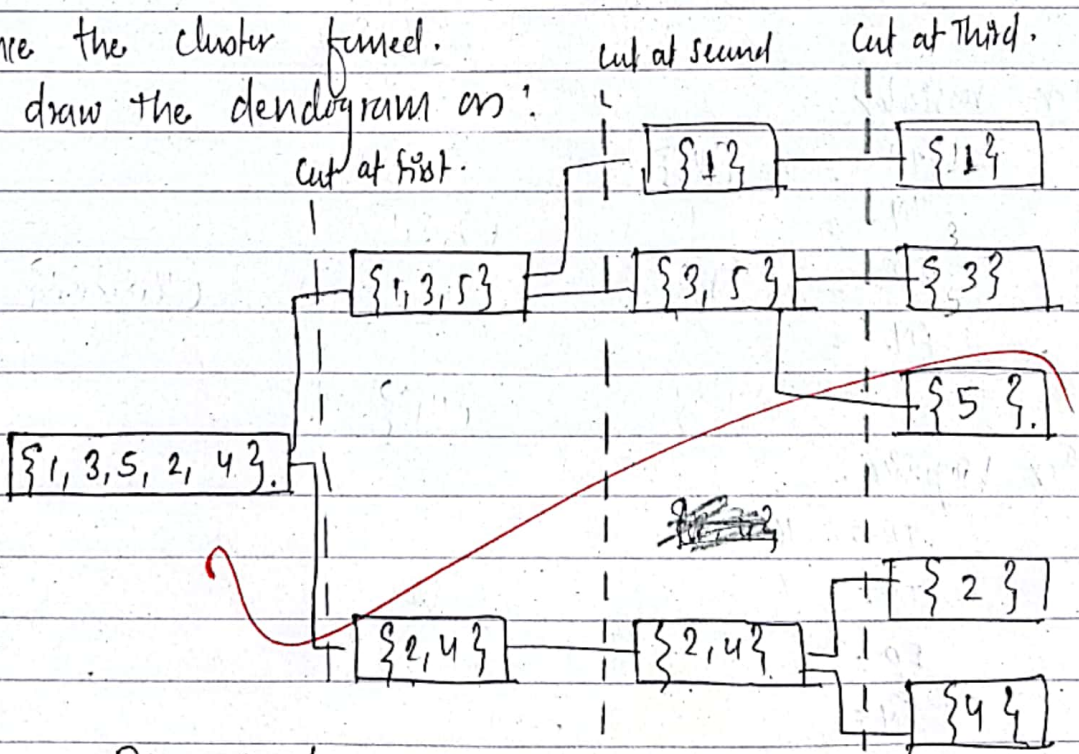


Fig: Dendogram.

In this cut we determine the value of k for K-means clustering.

9. Given, Confusion Matrix.

Actual	Predicted: Setosa	versicolor	Virginica
Setosa	21	2	0
versicolor	12	4	12
Virginica	5	3	16.

For, Setosa,

TP = 21

TN = 2+0 = 2.

FP = 2+0 = 2.

FN. =

A \ P	T	F
T	TP	FP
F	FN	TN.

For versicolor.

TP = 4

TN =

FP =

FN. =

For Virginica.

TP = 16.

TN =

FP =

FN =

In order to compute the F_1 score,

$$F_1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision is the total actual and predicted positive i.e True positive to the ratio of True and False positive.

Recall is the 1 - sensitivity.

Hence, F_1 - Score gives the Harmonic Mean of the confusion matrix.