Omkar Gurav
BE IT
8048

Rajdhani
DATE / /

MLA Theory Assignment 2

Q.1. Consider following data where
X_i = Rating for movie "Bahubali-part 1" by the person.
Y_i = Rating for movie "Bahubali-part 2" by the person.
Where rating is to done on the scale of 1 to 5 & 1 is lowest rating & 5 is highest rating.

i. Find values of β_0 & β_1 w.r.t. linear regression model which best fits given data.
ii. Interpret & explain equation of regression line.
iii. If new person votes "Bahubali part-1" as 3 then predict the rating of same person for "Bahubali part-2".

Person	X_i = Rating for movie "Bahubali-part 1" by ith the person	Y_i = Rating for movie "Bahubali-part 2" by ith the person
1	4	3
2	2	4
3	3	2
4	5	5
5	1	3
6	3	1

→

Rajdhani
DATE / /

i.

~~X_i Y_i~~

$$\overline{X} = \frac{4+2+3+5+1+3}{6} = \frac{18}{6} = 3$$

$$\overline{Y} = \frac{3+4+2+5+3+1}{6} = \frac{18}{6} = 3$$

X_i	Y_i	$X_i-\overline{X}$	$(X_i-\overline{X})^2$	$(Y_i-\overline{Y})$	$(X_i-\overline{X})(Y_i-\overline{Y})$
4	3	1	1	0	0
2	4	-1	1	1	-1
3	2	0	0	-1	0
5	5	2	4	2	4
1	3	-2	4	0	0
3	1	0	0	-2	0
			$\Sigma = 10$		$\Sigma = 3$

$$B_1 = \frac{\Sigma(X_i-\overline{X})(Y_i-\overline{Y})}{\Sigma(X_i-\overline{X})^2} = \frac{3}{10} = 0.3$$

$$B_0 = \overline{Y} - B_1\overline{X}$$
$$= 3 - 0.3 \times 3$$
$$= 3 - 0.9$$
$$B_0 = 2.1$$

ii. Equation of regression line is $Y = 2.1 + 0.3X$

Scanned with CamScanner

Interpretations:-

① For increase in value of x by 1 unit there is increase in value of y in 0.3 units.

② Even if $x=0$ value of independent variable, it is expected that value of Y is 2.1.

iii. We know, eqⁿ of regression line is $Y = 2.1 + 0.3X$.
Using this we can predict rating ~~of~~ for "Bahubali-part 2" if we know rating for "Bahubali-part 1".

Here, $X = 3$

$$Y = 2.1 + 0.3 \times 3$$
$$= 2.1 + 0.9$$
$$Y = 3$$

Thus, the rating prediction of same person for "Bahubali-part 2" is 3.

Q.2. Define Regularized regression. Explain Lasso & Ridge regression.

→ • The least-squares regression can turn out to be unstable as it is highly reliant on the training data. Instability is a pointer of a tendency to over fit. Regularization is a broad-spectrum method to avoid such over fitting

Rajdhani
DATE / /

by applying supplementary constraints to the weight vector.

- A general approach is to construct sure the weights are, on average, small in magnitude: this is referred to as shrinkage. This regularized problem still has a closed-form solution:

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

where I stands for the identity matrix with 1's on the diagonal & 0's everywhere else.

Lasso :-

- It is substitute form of regularized regression which means 'least absolute shrinkage & selection operator. It replaces the ridge regularization term $\Sigma_i w_i^2$ with the sum of absolute weights $\Sigma_i |w_i|$.

- It uses L1 regularization technique.

- It is normally used with more number of features, for the reason that it automatically does feature selection.

- It does not work well when features are highly correlated.

Rajdhani
DATE / /

Ridge Regression:-

- The regularization amounts to adding λ to the diagonal of X^TX, a renowned trick to get better the numerical stability of matrix inversion. This form of least-squares regression is known as ridge regression.

- It shrinks the parameters, so it is normally used to prevent multicollinearity.

- It reduces the model complexity by coefficient shrinkage.

- It uses L2 regularization technique.

- It is majorly used to prevent overfitting.

- It is not useful in case of high no. of features.

- It works well when features are highly correlated.

Q.3. Consider the following dataset consisting of the scores of two variables on each of seven individuals:

Rajdhani
DATE / /

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Consider subject 1 & subject 4 as initial centroids. Apply k-means algorithm & show step-by-step generation of clusters. Write the centroids final clusters.

→ $M_1 = (1.0, 1.0)$ $M_2 = (5.0, 7.0)$

Step 1:

	Subject	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Euclidian distance $= \sqrt{|M_{1x} - x_i|^2 + |M_{2y} - y_i|^2}$

$d(M_1, 1) = \sqrt{|1.0 - 1.0|^2 + |1.0 - 1.0|^2} = 0$

$d(M_2, 1) = \sqrt{|5.0 - 1.0|^2 + |7.0 - 1.0|^2} = 7.21$

~~$d(M_3, 2) =$~~

$d(M_1, 2) = \sqrt{|1.0 - 1.5|^2 + |1.0 - 2.0|^2} = 1.12$

Rajdhani DATE / /

$$d(M_2,2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} = 6.10$$

$$d(M_1,3) = \sqrt{|1.0-3.0|^2 + |1.0-4.0|^2} = 3.61$$

$$d(M_2,3) = \sqrt{|5.0-3.0|^2 + |7.0-4.0|^2} = 3.61$$

$$d(M_1,4) = \sqrt{|1.0-5.0|^2 + |1.0-7.0|^2} = 7.21$$

$$d(M_2,4) = \sqrt{|5.0-5.0|^2 + |7.0-7.0|^2} = 0$$

$$d(M_1,5) = \sqrt{|1.0-3.5|^2 + |1.0-5.0|^2} = 4.72$$

$$d(M_2,5) = \sqrt{|5.0-3.5|^2 + |7.0-5.0|^2} = 2.50$$

$$d(M_1,6) = \sqrt{|1.0-4.5|^2 + |1.0-5.0|^2} = 5.32$$

$$d(M_2,6) = \sqrt{|5.0-4.5|^2 + |7.0-5.0|^2} = 2.06$$

$$d(M_1,7) = \sqrt{|1.0-3.5|^2 + |1.0-4.5|^2} = 4.30$$

$$d(M_2,7) = \sqrt{|5.0-3.5|^2 + |7.0-4.5|^2} = 2.92$$

Step 2 : Thus, we obtain 2 clusters containing $\{1,2,3\}$ & $\{4,5,6,7\}$ as $d(M_1,i) \leq d(M_2,i)$ for subjects $\{1,2,3\}$.

Their new centroids are:

$$M_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$M_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$M_2 = (4.12, 5.38)$$

Subject	A	B	$d(M_1,S_i)$	$d(M_2,S_i)$
1	1.0	1.0	1.57	5.38
2	1.5	2.0	0.47	4.28
3	3.0	4.0	2.04	1.78
4	5.0	7.0	5.64	1.84
5	3.5	5.0	3.15	0.73
6	4.5	5.0	3.78	0.54
7	3.5	4.5	2.74	1.08

step 3: We obtain 2 new clusters $\{1,2\}$ & $\{3,4,5,6,7\}$ as $d(M_1,i) < d(M_2,i)$ for subjects 1 & 2.

New centroids are:

$$M_1 = \left(\frac{1}{2}(1.0+1.5), \frac{1}{2}(1.0+2.0)\right) = (1.25, 1.5)$$

$$M_2 = \left(\frac{1}{5}(3.0+5.0+3.5+4.5+3.5), \frac{1}{5}(4.0+7.0+5.0+5.0+4.5)\right)$$

$$M_2 = (3.9, 5.1)$$

Subject	A	B	d(M₁, Sᵢ)	d(M₂, Sᵢ)
1	1.0	1.0	0.56	5.02
2	1.5	2.0	0.47	3.92
3	3.0	4.0	2.04	1.42
4	5.0	7.0	5.64	2.20
5	3.5	5.0	3.15	0.41
6	4.5	5.0	3.78	0.61
7	3.5	4.5	2.74	0.72

Step 4: Since, there is no change in the cluster, the k-means algorithm comes to a halt here & final result consists of 2 clusters {1, 2} & {3, 4, 5, 6, 7}.

Q.4. For the following transactions data set generate association rules using Apriori algorithm. Consider minimum support as 50% & confidence as 75%.

Transaction ID	Items
1	Bread, Cheese, Eggs, Juice
2	Bread, Cheese, Juice
3	Bread, Milk, Yogurt
4	Bread, Juice, Milk
5	Cheese, Juice, Milk

→

Step 1:

support (item) = Frequency of item / Number of transactions

Item	Frequency	Support (in %)
Bread	4	4/5 = 80%
Cheese	3	3/5 = 60%
Eggs	1	1/5 = 20%
Juice	4	4/5 = 80%
Milk	3	3/5 = 60%
Yogurt	1	1/5 = 20%

Step 2: Remove all the items whose support is below given minimum support i.e. 50%

Item	Frequency	Support (in %)
Bread	4	80%
Cheese	3	60%
Juice	4	80%
Milk	3	60%

Step 3: Now, Form the 2 items candidate set & their Frequencies.

Items Pair	Frequency	Support (in %)
Bread, Cheese	2	2/5 = 40%
Bread, Juice	3	3/5 = 60%
Bread, Milk	2	2/5 = 40%
Cheese, Juice	3	3/5 = 60%
Cheese, Milk	1	1/5 = 20%
Juice, Milk	2	2/5 = 40%

Step 4: Remove all the items whose support is below given minimum support i.e. 50%.

Items Pair	Frequency	Support (in %)
Bread, Juice	3	60%
Cheese, Juice	3	60%

Step 5: Now, form 3 items candiate set & Frequency.

Items Pair	Frequency	Support (in %)
Bread, Juice, Cheese	2	2/5 = 40%

The 2 item subsets are {Bread, Juice}, {Juice, Cheese} & {Bread, Cheese}. But {Bread, Cheese} is not a member of table in step 4. & hence it is not Frequent & it is violating Apripri

Rajdhani
DATE / /

property. Thus {Bread, Juice, Cheese} is not considered.

Step 6: Generate Rules.

For rules, we consider item pairs:

① {Bread, Juice} Bread → Juice & Juice → Bread

② {Cheese, Juice} Cheese → Juice & Juice → Cheese

Confidence(A→B) = Support(A∪B) / Support(A)

Thus,

① Confidence(Bread → Juice) = Support(Bread ∪ Juice) / Support(Bread)
= 60/80
= 75%

② Confidence(Juice → Bread) = Support(Juice ∪ Bread) / Support(Juice)
= 60/80
= 75%

③ Confidence(Cheese → Juice) = Support(Cheese ∪ Juice) / Support(Cheese)
= 60/60
= 100%

Scanned with CamScanner

Rajdhani
DATE / /

④ Confidence (Juice → Cheese) = Support(Juice ∪ Cheese) / Support(Juice)

= 60/80

= 75%

All the above rules are valid because the confidence of each rule is greater than or equal to the confidence given i.e. 75%.