Assignment 4

Due on 2015-08-23, 23:55 IST

Submitted assignment

1) Ridge or Lasso regression is used to: *1 point*

- Shrink the co-efficients of the fitted model as compared to the regular OLS multiple regression.
- Select the appropriate subset of the input variables to include in a regression model.
- To compare different regression fits and select the best one.
- Transform the input variables such that they is no correlation between them (remove multicollinearity).

2) If you have 5 input variables. A best subset selection would evaluate how many unique models? *1 point*

- 25
- 32
- 5
- 1

3) In Decision Trees, There can be only 1 decision split/ attribute. Say True or False. *1 point*

- True
- False

4) For the following Data, *1 point*

A	B	Y
-1	-1	3
-1	1	3
1	-1	-1
1	1	7

Which of the following two prediction equations better describe the data?

y =3 + 4B --- model (i)

y = 4 + AB --- model(ii)

Using a squared error loss which model do you prefer?

- Model (i).
- Model (ii).
- Indifferent between Model(i) and Model (ii).
- There is insufficient data to answer this question.

5) For the same data and models in Question 4. Which model do you prefer if you were using mean absolute deviation? ***1 point***

- Model (i)
- Model (ii)
- Indifferent between Model (i) and Model (ii)
- There is insufficient data to answer this question

6) Questions 6 through 8 pertains to the following description: ***1 point***

A leading fashion store chooses to predict the Willingness of a customer to buy a shirt of a particular price category based on the customers data. The company strongly believes that the willingness of a customer to buy depends on 3 factors essentially. They

are the Gender of the customer(Male/ Female), The type of car used by the customer(Sports/ Family) and the type of shirt price category(Cheap/ Expensive).

From the past history, The company has got the data of 12 customers (gender, The type of the car used by the customer and The Shirt price category) along with the data of whether they bought the shirt of that category.

Customer ID	Gender	Car Type	Shirt Price Category	Will Buy?
1	Male	Sports	Cheap	no
2	Male	Sports	Expensive	yes
3	Male	Family	Cheap	yes
4	Male	Family	Expensive	no
5	Male	Sports	Cheap	yes
6	Male	Sports	Expensive	yes
7	Male	Family	Cheap	yes
8	Male	Family	Expensive	no
9	Female	Sports	Cheap	no
10	Female	Family	Cheap	no
11	Female	Sports	Expensive	no
12	Female	Family	Expensive	no

What is the best feature to split at the root level, If the splitting criterion is Entropy?

- Car Type
- Gender
- Shirt Price Category
- Gender or Car Type.

7) Assume that you stop growing the tree, when there are 2 or fewer data points in a leaf node. If you use Entropy for the splitting criterion, then What is the prediction accuracy? ***1 point***

- 83.33%
- 100%
- 91.67%
- 75%

8) Question 9 and 10 pertains to the following description. ***1 point***

A University chooses to predict the first semester GPA of a student based on his performance in the various rounds of the MBA Admission. The MBA Admission process of the University takes into consideration 3 scores essentially. They are the CAT(Common Admission Test) score, Written exam score and the Interview Score. The Scores are given in percentile.

From the previous batch, The University has got the scores(CAT, Written Test and Interview) of 10 students along with the GPA that they scored at the end of their first semester.

Student ID	CAT Score	Written Test	Interview	GPA
1	88	87	90	9.7
2	88	87	94	9.6
3	84	87	85	9.6
4	86	85	87	7.9
5	88	91	96	9.4
6	82	89	90	7.8
7	84	92	86	9.5
8	86	89	84	6.8
9	86	89	88	9
10	87	89	88	9.2

9. Perform a K-NN prediction at the point [CAT Score= 86, Written exam Score = 89, Interview Score = 88]. use a K value of 3.

To determine the nearest neighbors use the concept of Euclidean distance. Hint: In a 3-dimensional input space Euclidean distance between two points is determined in the following way: If point 1 has co-ordinates [a1, b1 and c1], and point 2 has co-ordinates [a2, b2 and c2], then the distance between these two points is $(a_2 - a_1)^2+(b_2-b_1)^2+(c_2-c_1)^2$

- 9
- 8.85
- 8.3
- 9.3

9) Perform a K-NN prediction at the point [CAT Score= 86, Written exam Score = 89, Interview Score = 88]. use a K value of 3. ***1 point***

To determine the nearest neighbors use the concept of Manhattan distance. Hint: In a 3-dimensional input space Manhattan distance between two points is determined in the following way: If point 1 has co-ordinates [a1, b1 and c1], and point 2 has co-ordinates [a2, b2 and c2], then the distance between these two points is |a1-a2|+|b1-b2|+|c1-c2|

- 9
- 8.85
- 8.3
- 9.3