Data Science Workshop-2 (CSE 2196)

ASSIGNMENT-5

1. We have seen in linear regression when the cost function is

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T x_{(i)} - y^{(i)})^2,$$

the closed form formula for θ is given by

$$\theta = (X^T X)^{-1} X^T y.$$

Now consider the cost function

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T x_{(i)} - y^{(i)})^2 + \frac{\lambda}{2}||\theta||^2 = \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) + \frac{\lambda}{2}||\theta||^2,$$

and find the closed form expression for θ.

2. Explain Gini impurity with an example.
3. Explain Entropy of a dataset with an example.
4. Is a Node's Entropy is generally higher than it's parent's entropy? Justify your answer.
5. Difference between the followings
   - Bagging and Boosting
   - Random Forest and Bagged decision tree
   - Splitting and pruning
   - Hard Voting and soft voting
   - K Means clustering and K nearest neighbors
6. What is Information gain in Decision tree? How is it related to Entropy?
7. What are the disadvantages of K-means clustering over other unsupervised algorithms?
8. Write an python program to calculate the entropy of a dataset if the dataset, a particular feature in the dataset and a value in that feature is given to you.
9. Do you think decision tree model can lead to over fitting? If yes, then how to handle over fitting? If no, why?
10. What do you mean by Ensemble learning? What are the benefits of ensemble learning? Discuss three other ensemble learning techniques other than Random forest.
11. Using the following dataset, find the final Cluster with initial centroid [1, 1] and [3, 3]

x_1	x_2
1	1
1	2
2	2
2	3
3	4
4	4
5	1
5	2
5	3

12. Explain the following clustering algorithms.
    - DB Scan
    - K-modes clustering
13. Consider the following dataset.

CGPA	Communication	Aptitude	Programming skill	Job offered
High	Good	High	Good	Yes
Medium	Good	High	Good	Yes
Low	Bad	Low	Good	Yes
Low	Good	Low	Bad	No
High	Good	High	Bad	Yes
High	Good	High	Good	Yes
Medium	Bad	Low	Bad	No
Medium	Bad	Low	Good	No
High	Good	High	Good	Yes
High	Good	High	Good	No
High	Good	Low	Good	Yes

    Using the given dataset predict whether a B. Tech CSE student will get the job or not using Naive Bays algorithm, If the evaluation parameters are as follows: CGPA= High; Communication= Bad; Aptitude= High; Programming skills= Bad
14. Which attribute will be selected as a root node, while constructing a Decision Tree. If attribute selection measure (ASM) is Entropy.
15. Which attribute will be selected as a root node, while constructing a Decision Tree. If attribute selection measure (ASM) is Gini Index.
16. Draw the decision tree by considering $max\_depth = 3$.
16. If the data points are: 5, 10, 15, 20, 25, 30, 35. Assume K= 2 and initial centroid as 15, 32. Create two clusters with the given set of centroids and calculate SSE.
17. Write the python code for finding closest centroid to a sample in k-means clustering.
18. What are the basic assumptions in the case of the Naive Bayes classifier?