Machine Learning – ES 654

Spring 2019

IITGN

1Ans.

a) In the case when date is included as an input feature, it will have an information gain of 0.94 which is highest with respect to all the other attributes (outlook-0.240, humidity-0.15, wind-0.042, temparature-0.030). This is because, date would perfectly predict the target attribute over the training data.

Choosing this attribute is not a good choice, as date is not a useful predictor despite being it being perfectly separates the training data. Hence, this tree would poorly classify on new examples/over unseen instances.

b) Yes, we could still learn the decision tree. By splitting the samples based on the probability values for each class at that depth of the decision tree

2Ans.

- **a)** Here is the link of the secret gist
- **b)** Here is the learnt decision tree on IRIS dataset. The tree is stored in dictionary format. This tree has test accuracy =93.33%. The above link contains the code for the same.

```
{'f3': {'<2.45': 0.0,
        '>2.45': {'f4': {'<1.65': {'f2': {'<2.8':{'f1': {'<6.0': 1.0,
'>6.0': 1.0}}, '>2.8': 1.0}},
                         '>1.65': {'f1': {'<5.9': 2.0,
                                          '>5.9': 2.0}}}}}}
```

Here

- f1:Sepal length
- f2:Sepal width
- f3:Petal length
- f4:Petal width
- class – 0: Setosa
- class – 1: Versicolar
- class - 2: Virginica

4Ans.

Here is the link for the code written using sklearn library. It gives an accuracy of 95.55%.

7Ans.

Here is the code for the figure-1. It depicts the variation of train time with change in N and M values.

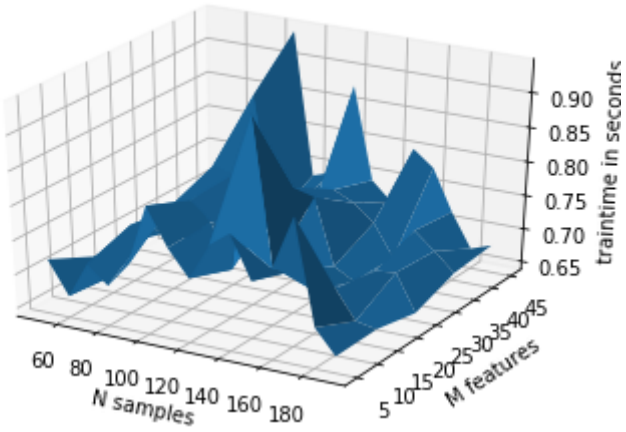

Figure-1

Here is the code for the figure-2. It depicts the variation of test time with change in N and M values.

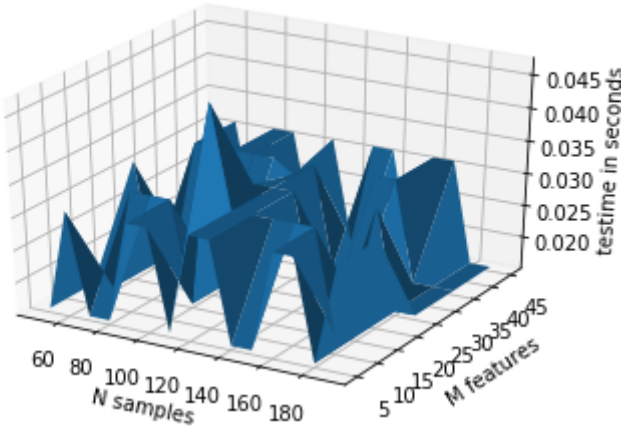

Figure-2

In the both the graphs we can see that, the values on Z-axis increases with increase in values of X and Y axes. The theoretical train and test times are of order O(M**2), which nearly approximates the above two curves.

References:

- Lecture notes
- https://scikit-learn.org/stable/modules/tree.html