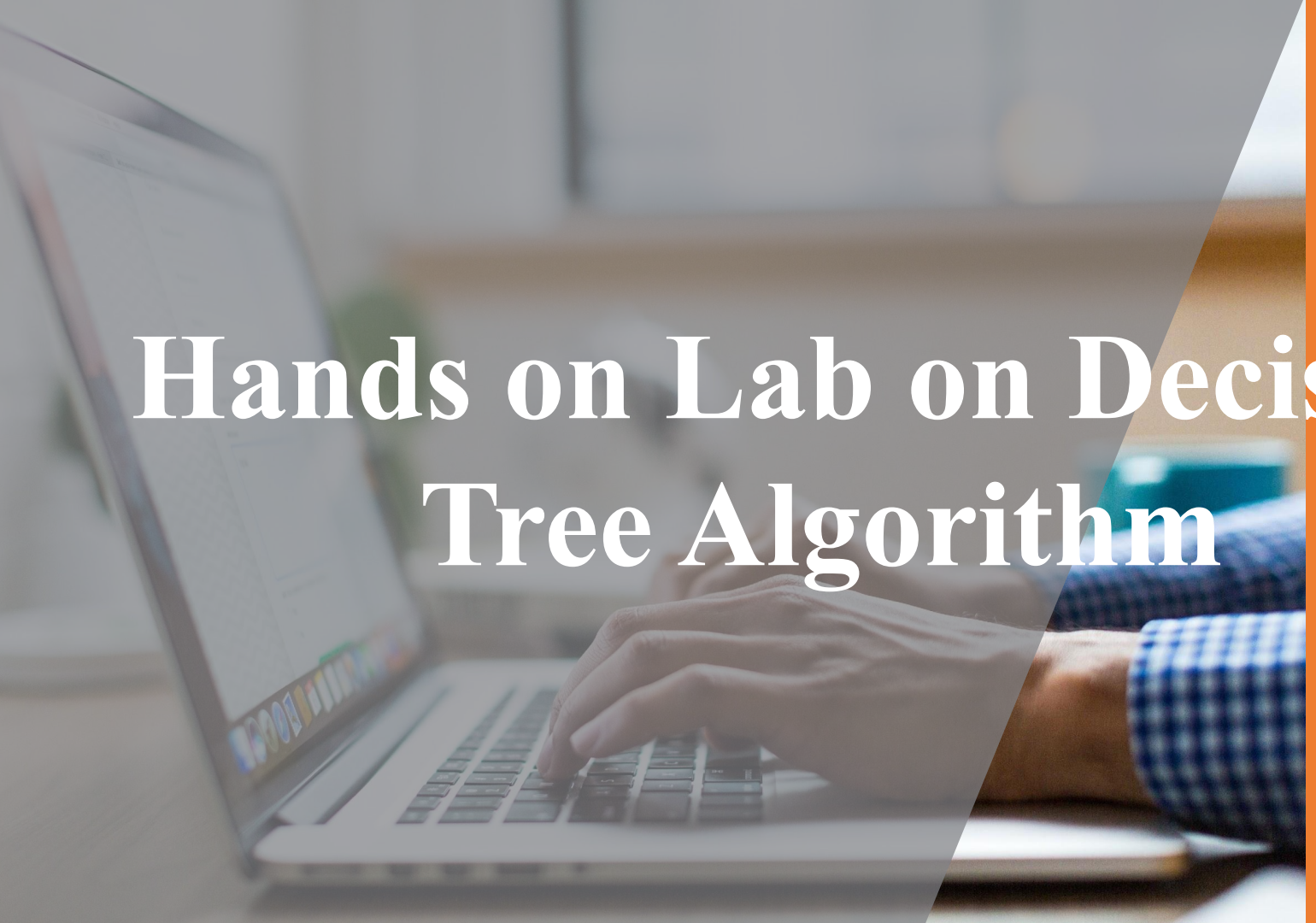

Hands on Lab on Decision Tree Algorithm

AGENDA OF THE VIDEO

- Should I go out playing or not?
- How does Decision Tree Algorithm Work?
- Two Important Questions while constructing the best decision tree.
- Information Gain for Deciding which feature to split the data on.
- Measure of Impurity of a split (Node).
- Information Gain as a Reduction in Impurity of a split.
- Decision Tree algorithm
- Pseudocode of Decision Tree Algorithm
- Coding a Decision Tree where all the features are categorical.
- Looking at the class Decision Tree Classifier of the sklearn
- Task For you….

Should I go out playing or not?

Outlook	Temperature	Humidity	Windy	Go Out
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

How does Decision Tree Algorithm Work?

Decision Tree algorithm is one of the most simplest algorithms to understand intuitively and it is also one of the algorithms with the least amount of math involved.

This algorithm works by asking a series of binary questions and with each question it splits the data in two parts. Then, for each of the resulting two datasets, we repeat the above steps recursively until some sufficient criteria gets satisfied.

Now, it can also be extended to cases where at each question, it splits the data into more than two parts. But for the sake of understanding, let us just stick to two.

Two important Questions…..

1. If i have n featues say X1, X2, X3,....,Xn, then which is the best feature to make the split on.
2. Once, I have decided that say Xi is the best feature to split our data on, the next obvious question is what value should we make a split on? This question only arises when the type of the variable of Xi is continuous as for continuous there can be many values on which we can split our dataset on. For categorical features, this issue is easiliy tackled.

What is Information Gain?

Information Gain is a metric for deciding which feature should we choose to make a split on.

Higher the information gain of a particular feature, higher its chance of getting chosen in final split step.

To understand how does Information Gain works, we need to understand about concept of impurity.

Concept of Impurity

Let us say I have two sets of data each of them is classification data with two classes A and B:

For set 1, no of rows belonging to class A = 100
no of rows belonging to class A = 100

For set 2, no of rows belonging to class A = 100
no of rows belonging to class A = 10

What do you think about each of these sets of data? Let us say to predict the class of a row (example), then without doing any machine learning, for which set of data, this prediction will be easier.

Concept of Impurity (Continued)

Obviously, it will be second set because in this set, most of the examples, are already belonging to the class A and hence this set of data will be more pure than the set 1. The set 1 will have high impurity as we are not sure, there is 50% chance of a example belonging to either of the clas.

Now, this impurity can be computed by various means. We will use ginni impurity in this video. The formula of ginni impurity is:

$$I_G(t) = \sum_{i=1}^{c} p(i|t)\left(1 - p(i|t)\right) = 1 - \sum_{i=1}^{c} p(i|t)^2$$

Concept of Impurity (Continued)

For set 1,

Impurity = 1 - (100/200)^2 - (100/200)^2= 0.50

For set 2,

Impurity = 1 - (100/110)^2 - (10/100)^2 = 0.165

We can clearly see that for set 2, the impurity is less than the set 1.

Information Gain in terms of Impurity

Let us say we have a data, D.
Let us say we want to split this data on feature, Xi.

Then, let us feature Xi has 4 unique values. So, After splitting the data D, we get four subsets D1, D2, D3 and D4.

Now, we can computer impurity for D, D1,D2,D3 and D4.

IG(Xi) = Imp(D) - sum of weighted impurities of D1,D2,D3 and D4
Now, if we choose that feature for which this information gain is maximum.
Hence, we can view information gain as a most reduction in the impurity.

Decision Tree Algorithm

We have X1,X2,..,Xn feature set and a variable to predict, say y.

1. Loop through all the features. For each feature, do the following..
   a. Compute its Information Gain
   b. Store this Information Gain in a List
2. Choose that feature for splitting for which the information gain is maximum.
3. Let the above feature splits the data into k sub nodes. So, we will have k sub datasets. For each of those datasets, call the steps 1 - 2 recursively until we get a perfect split.

Pseudocode

```
# Step.1 Make a function for computing ginni impurity for a dataset
def ginni_inpurity(dataset):
  get first class p1
  get second class p2
  return 1 - pow(p1,2) - pow(p2,2)
# Step.2 Make a function to calculate which feature to split on
def which_feature_to_split_on(data):
  infor_gain = []
  ontain the impurity of data without split
  loop through the columns
    for each unique value of column:
      compute the inpurity
    then compute the info gain and append it to infor_gain
  return that feature for which info gain was maximum

# Step.3 run_decision tree
Finally, call the which feature to split on function, recursively.
```

Coding the Decision Tree Algorithm in Python

Decision Tree Classifier class in Sklearn

Task For You...

Go to kaggle.com and search for real or fake job posting dataset and download the corresponding dataset.

Your task is to predict the whether a job posted is real or fake using the decision tree classifier used in sklearn.

If you have difficulty finding the right data, visit
https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction