

CS 7320/5320

Assignment DT (DecisionTrees)

Due : Sunday Oct 27

Consider the following data set (differs from data set in handouts)

Day	Outlook	Temperature	Humidity	Wind	Play Golf
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	No
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	No
D14	Rain	Mild	High	Strong	No

- (A) Use ID3 to MANUALLY determine which attribute should be used as the first high level factor in the resultant decision tree. Show your work including entropy values and Gain.
- (B) Use the Python Scikit DecisionTreeClassifier to build a DecisionTree model based on the above data. (see section 14.1) To do this you will need to convert the above labeled data into numeric data
(e.g. Sunny=1, Overcast=2, Rain=3, etc.)

What is the outcome for

D15 Overcast Mild Normal Weak

D16 Sunny Hot Normal Strong

- (C) Use the graphing technique in section 14.3 to draw the decision tree. The top level should match your results from part A.

Submit to Canvas:

- PDF or word doc of your results. If you use a Jupyter Notebook save it as a PDF so it can be easily read.

Code from 14.1 with ‘entropy’ option for Decision Tree.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn import datasets

# load data
iris = datasets.load_iris()
features = iris.data
print (features.shape)
print(features[0])

target = iris.target
print ( target.shape )
print (target[0] )

# create DT
decisiontree = DecisionTreeClassifier(criterion='entropy', random_state=0)

# train
model = decisiontree.fit(features, target)

# new observation
observation = [[5,4,3,2]]

# predict
print (model.predict(observation) )
```

Output:

(150, 4)

[5.1 3.5 1.4 0.2]

(150,)

0

[1]