DecisionTreeClassifier

April 24, 2018

0.1 Decision Tree Classifier

```
In [1]: from sklearn import tree
        from sklearn.tree import DecisionTreeClassifier
        from sklearn.preprocessing import OneHotEncoder, LabelEncoder
        import matplotlib.pyplot as plt
        from sklearn.model_selection import train_test_split, cross_val_score
        import pandas as pd
```

```
In [2]: indexes = ['age','workclass','fnlwgt','education','education-num','marital-status','occu
        'relationship','race','sex','capital-gain','capital-loss','hours-per-week','native-count
```

```
In [16]: # retrieve dataset => to predict that person earns <=50k or >50k
         df = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult
```

```
In [4]: df.tail()
```

```
Out[4]:        age      workclass  fnlwgt    education  education-num  \
        32556   27        Private  257302   Assoc-acdm             12
        32557   40        Private  154374      HS-grad              9
        32558   58        Private  151910      HS-grad              9
        32559   22        Private  201490      HS-grad              9
        32560   52   Self-emp-inc  287927      HS-grad              9

                    marital-status           occupation relationship    race      sex  \
        32556   Married-civ-spouse         Tech-support         Wife   White   Female
        32557   Married-civ-spouse    Machine-op-inspct      Husband   White     Male
        32558              Widowed         Adm-clerical    Unmarried   White   Female
        32559        Never-married         Adm-clerical    Own-child   White     Male
        32560   Married-civ-spouse      Exec-managerial         Wife   White   Female

               capital-gain  capital-loss  hours-per-week  native-country  earned
        32556             0             0              38   United-States   <=50K
        32557             0             0              40   United-States    >50K
        32558             0             0              40   United-States   <=50K
        32559             0             0              20   United-States   <=50K
        32560         15024             0              40   United-States    >50K
```

```
In [5]: Y = pd.DataFrame(df['earned'], columns=['earned'])
        # del df['earned']
```

```
In [6]: X = df[indexes[:-1]]
```

```
In [7]: X.head()
        Y.head()
```

```
Out[7]:    earned
        0   <=50K
        1   <=50K
        2   <=50K
        3   <=50K
        4   <=50K
```

```
In [8]: x_train,x_test,y_train,y_test = train_test_split(X,Y)
```

```
In [9]: # decision tree dt object
        dt = DecisionTreeClassifier()
        # label and onehot encoder object
        le = LabelEncoder()
        enc = OneHotEncoder()
```

```
In [10]: # x_train.shape
         # y_train.shape
         x_train.head()
```

```
Out[10]:        age         workclass  fnlwgt   education  education-num  \
         28861   19           Private  283945        10th              6
         12041   25           Private  248313   Assoc-voc             11
         20338   53         Local-gov  188772     HS-grad              9
         4076    39  Self-emp-not-inc  211785     HS-grad              9
         14049   38           Private  117528   Bachelors             13

                marital-status         occupation    relationship   race     sex  \
         28861   Never-married  Handlers-cleaners  Other-relative  White    Male
         12041   Never-married       Adm-clerical   Not-in-family  White  Female
         20338         Widowed      Other-service   Not-in-family  White  Female
         4076    Never-married       Craft-repair       Own-child  Black  Female
         14049   Never-married      Other-service  Other-relative  White  Female

                capital-gain  capital-loss  hours-per-week  native-country
         28861             0          1602              45   United-States
         12041             0             0              40   United-States
         20338             0             0              30   United-States
         4076              0             0              20   United-States
         14049             0             0              45   United-States
```

```
In [11]: x_train = x_train.apply(le.fit_transform)
         enc.fit(x_train)
         onehotlables = enc.transform(x_train).toarray()
         y_train = y_train.apply(le.fit_transform)
         enc.fit(y_train)
         onehotlables = enc.transform(y_train).toarray()
```

```
In [12]: dt.fit(x_train,y_train)
```

```
In [13]: x_test = x_test.apply(le.fit_transform)
         enc.fit(x_test)
         onehotlables = enc.transform(x_test).toarray()
         predicted = dt.predict(x_test)
```

```
In [15]: print(predicted)
         print(y_test)
```

```
[0 0 0 ... 0 1 0]
       earned
18868    >50K
1782    <=50K
8819    <=50K
5959    <=50K
1725    <=50K
19317   <=50K
10598   <=50K
8668     >50K
19505   <=50K
30593   <=50K
16103   <=50K
13533   <=50K
5960    <=50K
21987   <=50K
30782   <=50K
32190   <=50K
1896    <=50K
29561   <=50K
2720    <=50K
4037    <=50K
29856   <=50K
19290   <=50K
27554   <=50K
2021    <=50K
11689   <=50K
29789    >50K
22878   <=50K
17867    >50K
17064   <=50K
12404   <=50K
...       ...
26785   <=50K
12073   <=50K
7377     >50K
28279   <=50K
18441   <=50K
```

```
25951   <=50K
22249   <=50K
4281    <=50K
13871   <=50K
30279   <=50K
5429     >50K
24126   <=50K
8295    <=50K
1420    <=50K
26865    >50K
18083    >50K
16731   <=50K
27452   <=50K
22705   <=50K
5426    <=50K
28372   <=50K
12020   <=50K
14809   <=50K
27990   <=50K
21823   <=50K
20915   <=50K
13651    >50K
11466    >50K
22991    >50K
27175   <=50K

[8141 rows x 1 columns]
```