Lab 4 - Decision Tree

July 21, 2022

[30]:
```
# Decision Tree Classifier Building in Scikit-learn
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

[31]:
```
# Load the dataset. It consists of 5 features, UserID, Gender, Age,␣
 ↪EstimatedSalary and Purchased.
data = pd.read_csv('Social_Network_Ads.xls')
data.head()
```

[31]:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

[32]:
```
# We will take only Age and EstimatedSalary as our independent variables X␣
 ↪because of other features
# like Gender and User ID are irrelevant and have no effect on the purchasing␣
 ↪capacity of a person.
# Purchased is our dependent variable y.
feature_cols = ['Age','EstimatedSalary']
X = data.iloc[:,[2,3]].values
y = data.iloc[:,4].values
```

[33]:
```
# The next step is to split the dataset into training and test.
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test =  train_test_split(X,y,test_size = 0.25,␣
 ↪random_state= 0)
```

[34]:
```
#feature scaling
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

[35]:
```
# Fit the model in the Decision Tree classifier.
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion='entropy', random_state=0)
classifier = classifier.fit(X_train,y_train)
```

[36]:
```
# Make predictions and check accuracy.
#prediction
y_pred = classifier.predict(X_test)#Accuracy
from sklearn import metrics
print('Accuracy Score:', metrics.accuracy_score(y_test,y_pred))
```

```
Accuracy Score: 0.91
```

[37]:
```
# Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
```

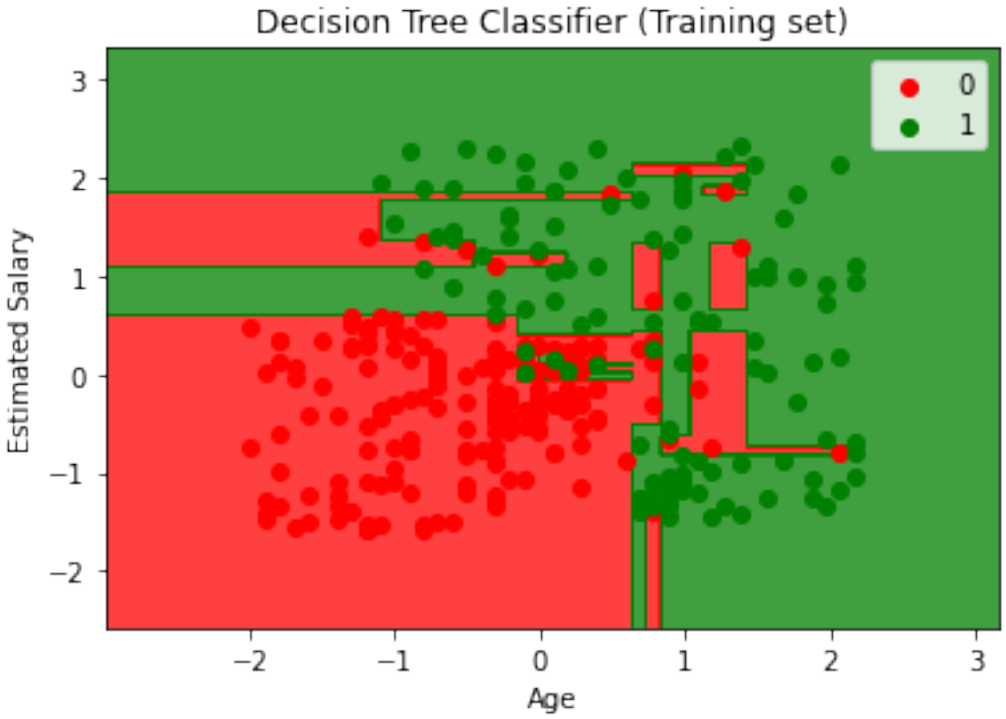
[38]:
```
print (cm)
```

```
[[62  6]
 [ 3 29]]
```

[39]:
```
# Let us first visualize the model prediction results.
from matplotlib.colors import ListedColormap
X_set, y_set = X_train, y_train
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:,␣
↪0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:,␣
↪1].max() + 1, step = 0.01))
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).
↪reshape(X1.shape),
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Decision Tree Classifier (Training set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()
```

```
*c* argument looks like a single numeric RGB or RGBA sequence, which should be
avoided as value-mapping will have precedence in case its length matches with
*x* & *y*.  Please use the *color* keyword-argument or provide a 2D array with a
single row if you intend to specify the same RGB or RGBA value for all points.
```

```
*c* argument looks like a single numeric RGB or RGBA sequence, which should be
avoided as value-mapping will have precedence in case its length matches with
*x* & *y*.  Please use the *color* keyword-argument or provide a 2D array with a
single row if you intend to specify the same RGB or RGBA value for all points.
```



[49]:
```
from sklearn import tree
```

[53]:
```
pip install graphviz
```

```
Collecting graphviz
  Downloading graphviz-0.20-py3-none-any.whl (46 kB)
Installing collected packages: graphviz
Successfully installed graphviz-0.20
Note: you may need to restart the kernel to use updated packages.
```

[63]:
```
text_representation = tree.export_text(clf)
print(text_representation)
```

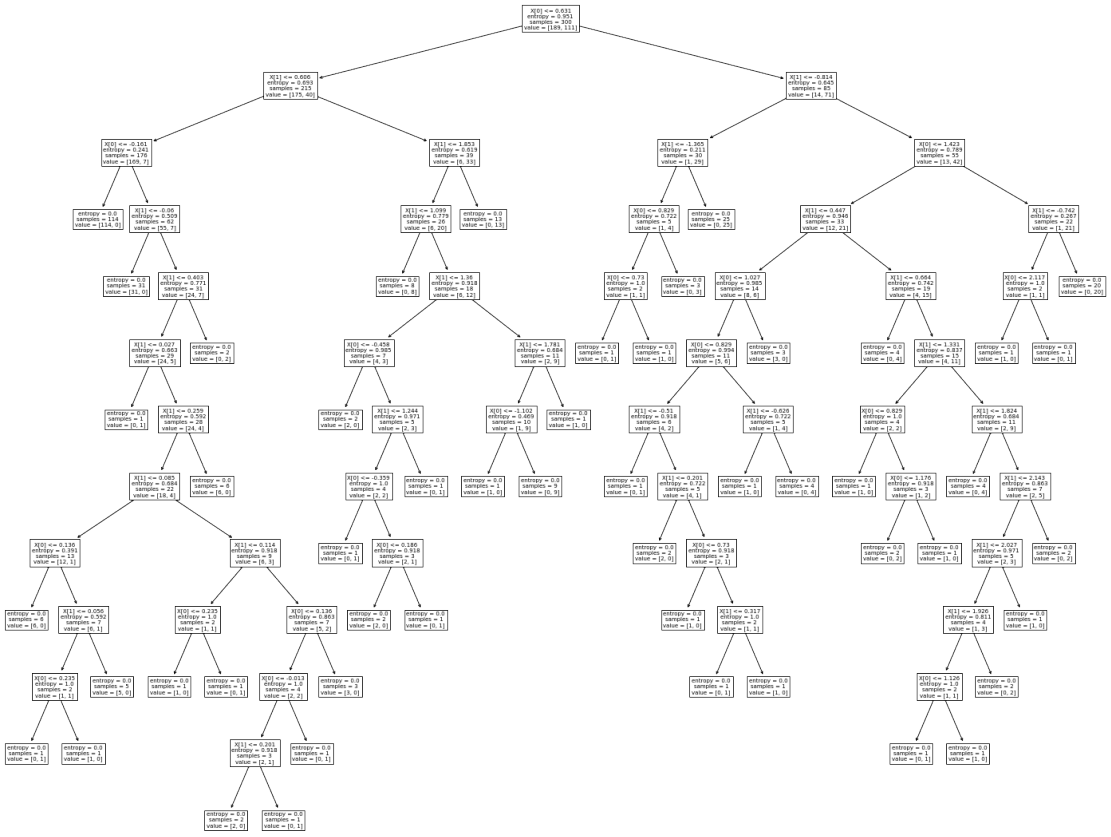
```
|--- feature_0 <= 0.63
|   |--- feature_1 <= 0.61
|   |   |--- feature_0 <= -0.16
|   |   |   |--- class: 0
|   |   |--- feature_0 >  -0.16
|   |   |   |--- feature_1 <= -0.06
```

```
|   |   |   |   |--- class: 0
|   |   |   |--- feature_1 >  -0.06
|   |   |   |   |--- feature_1 <= 0.40
|   |   |   |   |   |--- feature_1 <= 0.03
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- feature_1 >  0.03
|   |   |   |   |   |   |--- feature_1 <= 0.26
|   |   |   |   |   |   |   |--- feature_1 <= 0.08
|   |   |   |   |   |   |   |   |--- feature_0 <= 0.14
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_0 >  0.14
|   |   |   |   |   |   |   |   |   |--- feature_1 <= 0.06
|   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 0.24
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |   |--- feature_0 >  0.24
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |--- feature_1 >  0.06
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_1 >  0.08
|   |   |   |   |   |   |   |   |   |--- feature_1 <= 0.11
|   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 0.24
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_0 >  0.24
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_1 >  0.11
|   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 0.14
|   |   |   |   |   |   |   |   |   |   |   |--- feature_0 <= -0.01
|   |   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |   |   |--- feature_0 >  -0.01
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_0 >  0.14
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_1 >  0.26
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_1 >  0.40
|   |   |   |   |   |   |   |--- class: 1
|   |   |--- feature_1 >  0.61
|   |   |   |--- feature_1 <= 1.85
|   |   |   |   |--- feature_1 <= 1.10
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- feature_1 >  1.10
|   |   |   |   |   |--- feature_1 <= 1.36
|   |   |   |   |   |   |--- feature_0 <= -0.46
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_0 >  -0.46
|   |   |   |   |   |   |   |--- feature_1 <= 1.24
|   |   |   |   |   |   |   |   |--- feature_0 <= -0.36
|   |   |   |   |   |   |   |   |   |--- class: 1
```

```
|   |   |   |   |   |   |   |--- feature_0 >  -0.36
|   |   |   |   |   |   |   |   |--- feature_0 <= 0.19
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_0 >  0.19
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- feature_1 >  1.24
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- feature_1 >  1.36
|   |   |   |   |   |   |--- feature_1 <= 1.78
|   |   |   |   |   |   |   |--- feature_0 <= -1.10
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_0 >  -1.10
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- feature_1 >  1.78
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |--- feature_1 >  1.85
|   |   |   |   |--- class: 1
|--- feature_0 >  0.63
|   |--- feature_1 <= -0.81
|   |   |--- feature_1 <= -1.37
|   |   |   |--- feature_0 <= 0.83
|   |   |   |   |--- feature_0 <= 0.73
|   |   |   |   |   |--- class: 1
|   |   |   |   |--- feature_0 >  0.73
|   |   |   |   |   |--- class: 0
|   |   |   |--- feature_0 >  0.83
|   |   |   |   |--- class: 1
|   |   |--- feature_1 >  -1.37
|   |   |   |--- class: 1
|   |--- feature_1 >  -0.81
|   |   |--- feature_0 <= 1.42
|   |   |   |--- feature_1 <= 0.45
|   |   |   |   |--- feature_0 <= 1.03
|   |   |   |   |   |--- feature_0 <= 0.83
|   |   |   |   |   |   |--- feature_1 <= -0.51
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |--- feature_1 >  -0.51
|   |   |   |   |   |   |   |--- feature_1 <= 0.20
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_1 >  0.20
|   |   |   |   |   |   |   |   |--- feature_0 <= 0.73
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_0 >  0.73
|   |   |   |   |   |   |   |   |--- feature_1 <= 0.32
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- feature_1 >  0.32
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- feature_0 >  0.83
```

```
|   |   |   |   |   |   |--- feature_1 <= -0.63
|   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_1 >  -0.63
|   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- feature_0 >  1.03
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |--- feature_1 >  0.45
|   |   |   |   |   |--- feature_1 <= 0.66
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |--- feature_1 >  0.66
|   |   |   |   |   |   |--- feature_1 <= 1.33
|   |   |   |   |   |   |   |--- feature_0 <= 0.83
|   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |--- feature_0 >  0.83
|   |   |   |   |   |   |   |   |--- feature_0 <= 1.18
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |--- feature_0 >  1.18
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- feature_1 >  1.33
|   |   |   |   |   |   |   |--- feature_1 <= 1.82
|   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |--- feature_1 >  1.82
|   |   |   |   |   |   |   |   |--- feature_1 <= 2.14
|   |   |   |   |   |   |   |   |   |--- feature_1 <= 2.03
|   |   |   |   |   |   |   |   |   |   |--- feature_1 <= 1.93
|   |   |   |   |   |   |   |   |   |   |   |--- feature_0 <= 1.13
|   |   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |   |   |--- feature_0 >  1.13
|   |   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |   |   |--- feature_1 >  1.93
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |   |   |   |   |   |   |--- feature_1 >  2.03
|   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- feature_1 >  2.14
|   |   |   |   |   |   |   |   |   |--- class: 1
|   |   |   |--- feature_0 >  1.42
|   |   |   |   |--- feature_1 <= -0.74
|   |   |   |   |   |--- feature_0 <= 2.12
|   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |--- feature_0 >  2.12
|   |   |   |   |   |   |--- class: 1
|   |   |   |   |--- feature_1 >  -0.74
|   |   |   |   |   |--- class: 1
```

[71]:
```
fig = plt.figure(figsize=(25,20))
_ = tree.plot_tree(clf)
```



[70]:
```
iris = datasets.load_iris()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-70-f811bec3b759> in <module>
----> 1 iris = datasets.load_iris()

NameError: name 'datasets' is not defined
```