03-simple-classification

April 22, 2025

1 Topic 2: Simple Classification Exercise

I'm following along the **Decision Trees** implementation from this YouTube.

This notebook can be accessed via GitHub repository here.

About the Dataset Dataset used in this exercise is using built-in data set from scikit-learn.

The list of the dataset can be accessed here.

In this exercise, I'll be using breast cancer data which also will be used in my next exercise too.

Import necessary libraries & dataset

[1]:
```
import pandas as pd
from sklearn.datasets import load_breast_cancer # the dataset
```

Load & display the dataset

[2]:
```
data = load_breast_cancer()
dataset = pd.DataFrame(data=data['data'], columns=data['feature_names']) # what␣
↪makes this dataset is easy because it's been cleaned
dataset
```

[2]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	\
0	17.99	10.38	122.80	1001.0	0.11840	
1	20.57	17.77	132.90	1326.0	0.08474	
2	19.69	21.25	130.00	1203.0	0.10960	
3	11.42	20.38	77.58	386.1	0.14250	
4	20.29	14.34	135.10	1297.0	0.10030	
..	…	…	…	…	…	
564	21.56	22.39	142.00	1479.0	0.11100	
565	20.13	28.25	131.20	1261.0	0.09780	
566	16.60	28.08	108.30	858.1	0.08455	
567	20.60	29.33	140.10	1265.0	0.11780	
568	7.76	24.54	47.92	181.0	0.05263	

	mean compactness	mean concavity	mean concave points	mean symmetry	\
0	0.27760	0.30010	0.14710	0.2419	
1	0.07864	0.08690	0.07017	0.1812	
2	0.15990	0.19740	0.12790	0.2069	

```
3            0.28390         0.24140              0.10520         0.2597
4            0.13280         0.19800              0.10430         0.1809
..               …               …                    …              …
564          0.11590         0.24390              0.13890         0.1726
565          0.10340         0.14400              0.09791         0.1752
566          0.10230         0.09251              0.05302         0.1590
567          0.27700         0.35140              0.15200         0.2397
568          0.04362         0.00000              0.00000         0.1587

     mean fractal dimension  …  worst radius  worst texture  \
0                   0.07871  …        25.380          17.33
1                   0.05667  …        24.990          23.41
2                   0.05999  …        23.570          25.53
3                   0.09744  …        14.910          26.50
4                   0.05883  …        22.540          16.67
..                      …  …           …              …
564                 0.05623  …        25.450          26.40
565                 0.05533  …        23.690          38.25
566                 0.05648  …        18.980          34.12
567                 0.07016  …        25.740          39.42
568                 0.05884  …         9.456          30.37

     worst perimeter  worst area  worst smoothness  worst compactness  \
0             184.60      2019.0           0.16220            0.66560
1             158.80      1956.0           0.12380            0.18660
2             152.50      1709.0           0.14440            0.42450
3              98.87       567.7           0.20980            0.86630
4             152.20      1575.0           0.13740            0.20500
..               …           …                …                 …
564           166.10      2027.0           0.14100            0.21130
565           155.00      1731.0           0.11660            0.19220
566           126.70      1124.0           0.11390            0.30940
567           184.60      1821.0           0.16500            0.86810
568            59.16       268.6           0.08996            0.06444

     worst concavity  worst concave points  worst symmetry  \
0             0.7119                0.2654          0.4601
1             0.2416                0.1860          0.2750
2             0.4504                0.2430          0.3613
3             0.6869                0.2575          0.6638
4             0.4000                0.1625          0.2364
..               …                     …               …
564           0.4107                0.2216          0.2060
565           0.3215                0.1628          0.2572
566           0.3403                0.1418          0.2218
567           0.9387                0.2650          0.4087
568           0.0000                0.0000          0.2871
```

```
     worst fractal dimension
0                    0.11890
1                    0.08902
2                    0.08758
3                    0.17300
4                    0.07678
..                       …
564                  0.07115
565                  0.06637
566                  0.07820
567                  0.12400
568                  0.07039

[569 rows x 30 columns]
```

[3]:
```
dataset.shape
```

[3]:
```
(569, 30)
```

[4]:
```
dataset.describe()
```

[4]:
```
       mean radius  mean texture  mean perimeter    mean area  \
count   569.000000    569.000000      569.000000   569.000000
mean     14.127292     19.289649       91.969033   654.889104
std       3.524049      4.301036       24.298981   351.914129
min       6.981000      9.710000       43.790000   143.500000
25%      11.700000     16.170000       75.170000   420.300000
50%      13.370000     18.840000       86.240000   551.100000
75%      15.780000     21.800000      104.100000   782.700000
max      28.110000     39.280000      188.500000  2501.000000

       mean smoothness  mean compactness  mean concavity  mean concave points  \
count       569.000000        569.000000      569.000000           569.000000
mean          0.096360          0.104341        0.088799             0.048919
std           0.014064          0.052813        0.079720             0.038803
min           0.052630          0.019380        0.000000             0.000000
25%           0.086370          0.064920        0.029560             0.020310
50%           0.095870          0.092630        0.061540             0.033500
75%           0.105300          0.130400        0.130700             0.074000
max           0.163400          0.345400        0.426800             0.201200

       mean symmetry  mean fractal dimension  …  worst radius  \
count     569.000000              569.000000  …    569.000000
mean        0.181162                0.062798  …     16.269190
std         0.027414                0.007060  …      4.833242
min         0.106000                0.049960  …      7.930000
```

```
25%         0.161900                0.057700  …     13.010000
50%         0.179200                0.061540  …     14.970000
75%         0.195700                0.066120  …     18.790000
max         0.304000                0.097440  …     36.040000

       worst texture  worst perimeter   worst area  worst smoothness  \
count     569.000000       569.000000   569.000000        569.000000
mean       25.677223       107.261213   880.583128          0.132369
std         6.146258        33.602542   569.356993          0.022832
min        12.020000        50.410000   185.200000          0.071170
25%        21.080000        84.110000   515.300000          0.116600
50%        25.410000        97.660000   686.500000          0.131300
75%        29.720000       125.400000  1084.000000          0.146000
max        49.540000       251.200000  4254.000000          0.222600

       worst compactness  worst concavity  worst concave points  \
count         569.000000       569.000000            569.000000
mean            0.254265         0.272188              0.114606
std             0.157336         0.208624              0.065732
min             0.027290         0.000000              0.000000
25%             0.147200         0.114500              0.064930
50%             0.211900         0.226700              0.099930
75%             0.339100         0.382900              0.161400
max             1.058000         1.252000              0.291000

       worst symmetry  worst fractal dimension
count      569.000000               569.000000
mean         0.290076                 0.083946
std          0.061867                 0.018061
min          0.156500                 0.055040
25%          0.250400                 0.071460
50%          0.282200                 0.080040
75%          0.317900                 0.092080
max          0.663800                 0.207500

[8 rows x 30 columns]
```

Split the dataset into training and testing

[5]:
```
from sklearn.model_selection import train_test_split
X = dataset.copy()
y = data['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

Train the Model

[6]:
```
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier()
clf = clf.fit(X_train, y_train)
```

[7]:
```
# identifying the parameters and it's in default mode as we're not specifying␣
↪any parameters in the clf above

clf.get_params()
```

[7]:
```
{'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': None,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'monotonic_cst': None,
 'random_state': None,
 'splitter': 'best'}
```

[8]:
```
X_test
```

[8]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	\
354	11.140	14.07	71.24	384.6	0.07274	
42	19.070	24.81	128.30	1104.0	0.09081	
31	11.840	18.70	77.93	440.6	0.11090	
233	20.510	27.81	134.40	1319.0	0.09159	
125	13.850	17.21	88.44	588.7	0.08785	
..	…	…	…	…	…	
192	9.720	18.22	60.73	288.1	0.06950	
110	9.777	16.99	62.50	290.2	0.10370	
81	13.340	15.86	86.49	520.0	0.10780	
197	18.080	21.84	117.40	1024.0	0.07371	
241	12.420	15.04	78.61	476.5	0.07926	

	mean compactness	mean concavity	mean concave points	mean symmetry	\
354	0.06064	0.04505	0.01471	0.1690	
42	0.21900	0.21070	0.09961	0.2310	
31	0.15160	0.12180	0.05182	0.2301	
233	0.10740	0.15540	0.08340	0.1448	
125	0.06136	0.01420	0.01141	0.1614	
..	…	…	…	…	
192	0.02344	0.00000	0.00000	0.1653	
110	0.08404	0.04334	0.01778	0.1584	
81	0.15350	0.11690	0.06987	0.1942	
197	0.08642	0.11030	0.05778	0.1770	
241	0.03393	0.01053	0.01108	0.1546	

```
     mean fractal dimension  …  worst radius  worst texture  \
354                 0.06083  …        12.120          15.82
42                  0.06343  …        24.090          33.17
31                  0.07799  …        16.820          28.12
233                 0.05592  …        24.470          37.38
125                 0.05890  …        15.490          23.58
..                      …  …           …              …
192                 0.06447  …         9.968          20.83
110                 0.07065  …        11.050          21.47
81                  0.06902  …        15.530          23.19
197                 0.05340  …        19.760          24.70
241                 0.05754  …        13.200          20.37

     worst perimeter  worst area  worst smoothness  worst compactness  \
354            79.62       453.5           0.08864            0.12560
42            177.40      1651.0           0.12470            0.74440
31            119.40       888.7           0.16370            0.57750
233           162.70      1872.0           0.12230            0.27610
125           100.30       725.9           0.11570            0.13500
..               …         …                …                 …
192            62.25       303.8           0.07117            0.02729
110            71.68       367.0           0.14670            0.17650
81             96.66       614.9           0.15360            0.47910
197           129.10      1228.0           0.08822            0.19630
241            83.85       543.4           0.10370            0.07776

     worst concavity  worst concave points  worst symmetry  \
354          0.12010               0.03922          0.2576
42           0.72420               0.24930          0.4670
31           0.69560               0.15460          0.4761
233          0.41460               0.15630          0.2437
125          0.08115               0.05104          0.2364
..               …                     …               …
192          0.00000               0.00000          0.1909
110          0.13000               0.05334          0.2533
81           0.48580               0.17080          0.3527
197          0.25350               0.09181          0.2369
241          0.06243               0.04052          0.2901

     worst fractal dimension
354                  0.07018
42                   0.10380
31                   0.14020
233                  0.08328
125                  0.07182
..                       …
192                  0.06559
```

```
110                  0.08468
81                   0.10160
197                  0.06558
241                  0.06783

[188 rows x 30 columns]
```

Test the dataset & make prediction

[9]:
```
# 0 is malignant
# 1 is benign

predictions = clf.predict(X_test)
predictions
```

[9]:
```
array([1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1,
       1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1,
       1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1,
       1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1,
       1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0,
       0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1,
       0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1,
       0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0,
       1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1])
```

Evaluate the performance There's already built-in accuracy scores from the scikit-learn library. It can be accessed here.

[10]:
```
#accuracy score

from sklearn.metrics import accuracy_score
accuracy_score(y_test, predictions)
```

[10]:
```
0.9468085106382979
```

[11]:
```
# confusion matrix

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, predictions, labels=[0,1])
cm
```

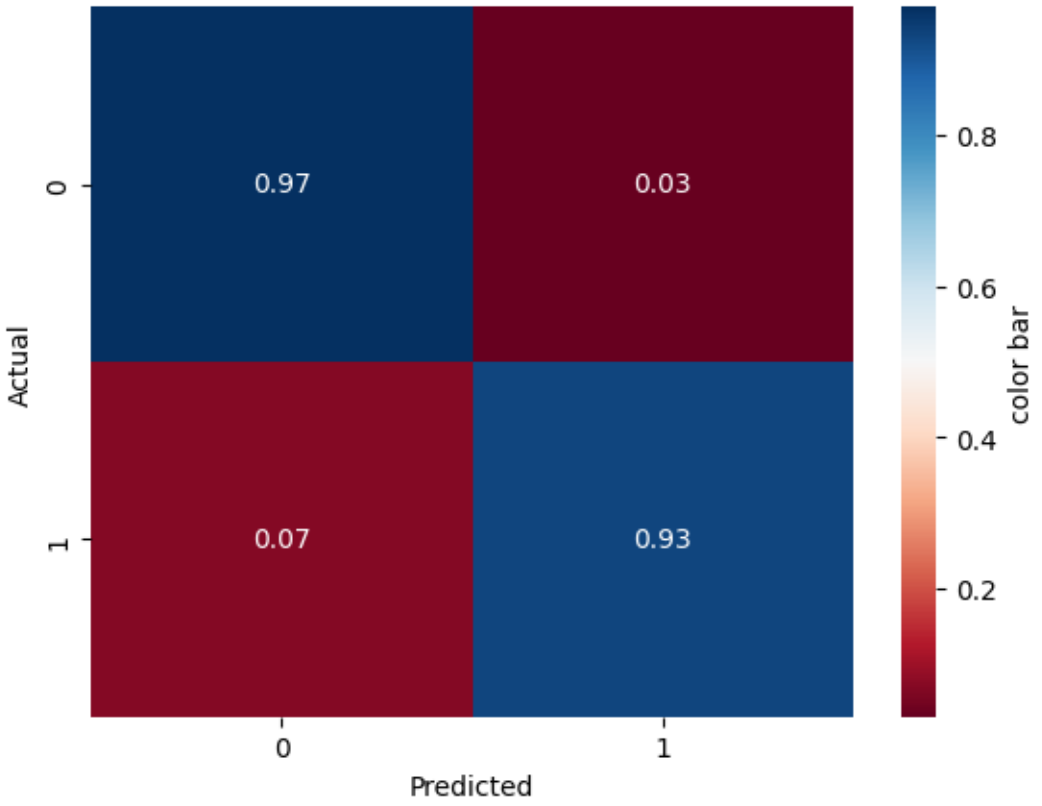
[11]:
```
array([[ 64,   2],
       [  8, 114]])
```

[12]:
```
# Let's visualize a heatmap for our confusion matrix using seaborn & matplotlib
# We also want to normalize the numbers to be between -1 to 1 using numpy

import numpy as np
```

```
cm_normalized = np.round(cm/np.sum(cm, axis=1).reshape(-1,1),2)

import seaborn as sns
import matplotlib.pyplot as plt
sns.heatmap(cm_normalized, cmap="RdBu", annot=True,
            cbar_kws={"orientation":"vertical","label":"color bar"},
            xticklabels=[0,1], yticklabels=[0,1]
            )
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```



[13]:
```
# precision score

from sklearn.metrics import precision_score
precision_score(y_test, predictions)
```

[13]: 0.9827586206896551

[14]:
```
# recall

from sklearn.metrics import recall_score
recall_score(y_test, predictions)
```

[14]: 0.9344262295081968

[15]:
```
# classification report (this is an extensive report from the library)

from sklearn.metrics import classification_report
print(classification_report(y_test, predictions, target_names
 ↪=['malignant','benign']))
```

```
              precision    recall  f1-score   support

   malignant       0.89      0.97      0.93        66
      benign       0.98      0.93      0.96       122

    accuracy                           0.95       188
   macro avg       0.94      0.95      0.94       188
weighted avg       0.95      0.95      0.95       188
```