Machine Learning - 酒類品種區分

Stella, Jing

January, 04, 2019

簡介

- 程式功能：以機器學習的模型解析各欄位資料以分辨酒的種類
- 執行步驟：
- ✓ 了解資料欄位意義
- ✓ 以圖表表示資料各欄位與 label 間的相關性
- ✓ 分別以 SVC, Linear SVC, Decision Tree, Random Forest 四個分類器進行預測檢視其 accuracies

資料欄位

- 資料由 Institute of Pharmaceutical and Food Analysis and Technologies 提供，以科學分析種植於義大利相同地區的三種葡萄人工栽培品種所釀造出的酒，確認 13 種判斷酒質與種類的資訊
- 欄位：
  - Alcohol,Malic acid,Ash,Alcalinity of ash,Magnesium,Total phenols,Flavanoids,Nonflavanoid phenols,Proanthocyanins,Color intensity,Hue,OD280/OD315 of diluted wines,Proline

資料匯入

- 自 sciki-learn 的 datasets 匯入酒類資料的各項資料數值 -wine.data
- 自 sciki-learn 的 datasets 匯入 label-wine.target
- 觀察資料結構

```
wine = datasets.load_wine()
```

```
features = wine.data
labels = wine.target
df = pd.DataFrame(wine['data'], columns = wine['feature_names'])
df["target"] = wine["target"]

print(df.shape)
df.info()
df.head()
```

```
(178, 14)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 14 columns):
alcohol                         178 non-null float64
malic_acid                      178 non-null float64
ash                             178 non-null float64
alcalinity_of_ash               178 non-null float64
```

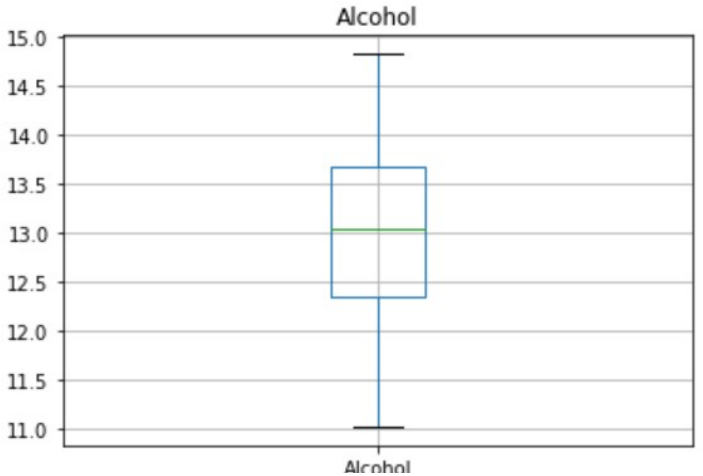
Exploratory Data Analysis

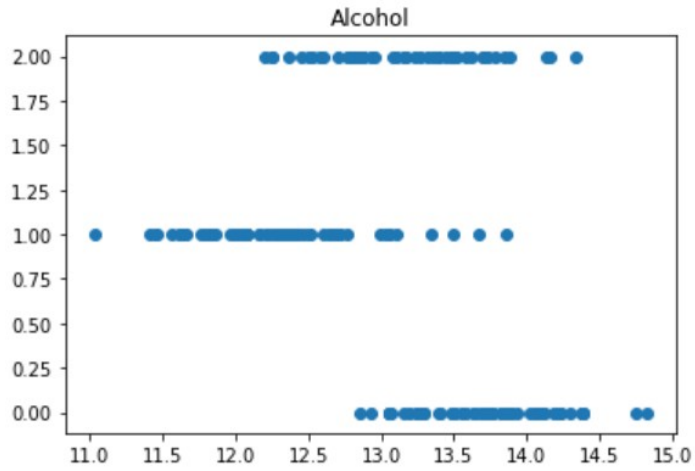
- 檢視各欄位的統計性數據
- 檢視各欄位的盒鬚圖、與 Target 間的點狀分布、直線圖

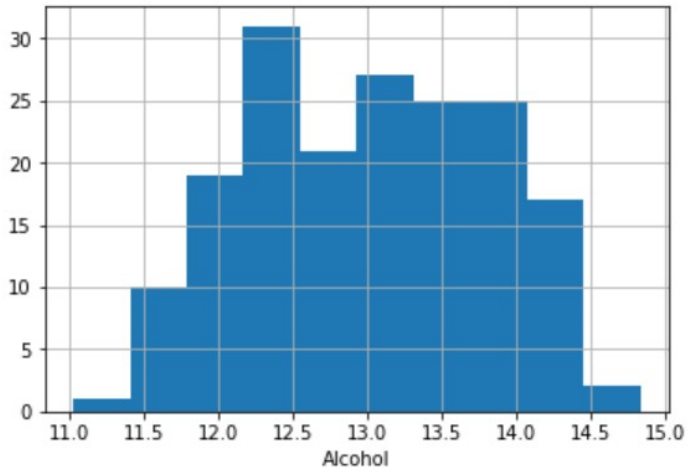
```
plt.figure(figsize=(14,35))

for col in df.columns:
#       print(df[col].describe())
    plt.subplot(nrows,ncols,a)
    df.boxplot(col)
    plt.ylabel(col)
    plt.title(col)

    plt.subplot(nrows,ncols,a+1)
    plt.scatter(df[col], df['target'])
    plt.xlabel(col)
    plt.ylabel('target')
#     plt.title(col)

    plt.subplot(nrows,ncols,a+2)
    df[col].hist()
    plt.xlabel(col)
    plt.ylabel('count')
#     plt.title(col)

    a+=3
plt.tight_layout()
plt.show()
```

Exploratory Data Analysis

```
count    178.000000
mean      13.000618
std        0.811827
min       11.030000
25%       12.362500
50%       13.050000
75%       13.677500
max       14.830000
Name: Alcohol, dtype: float64
```



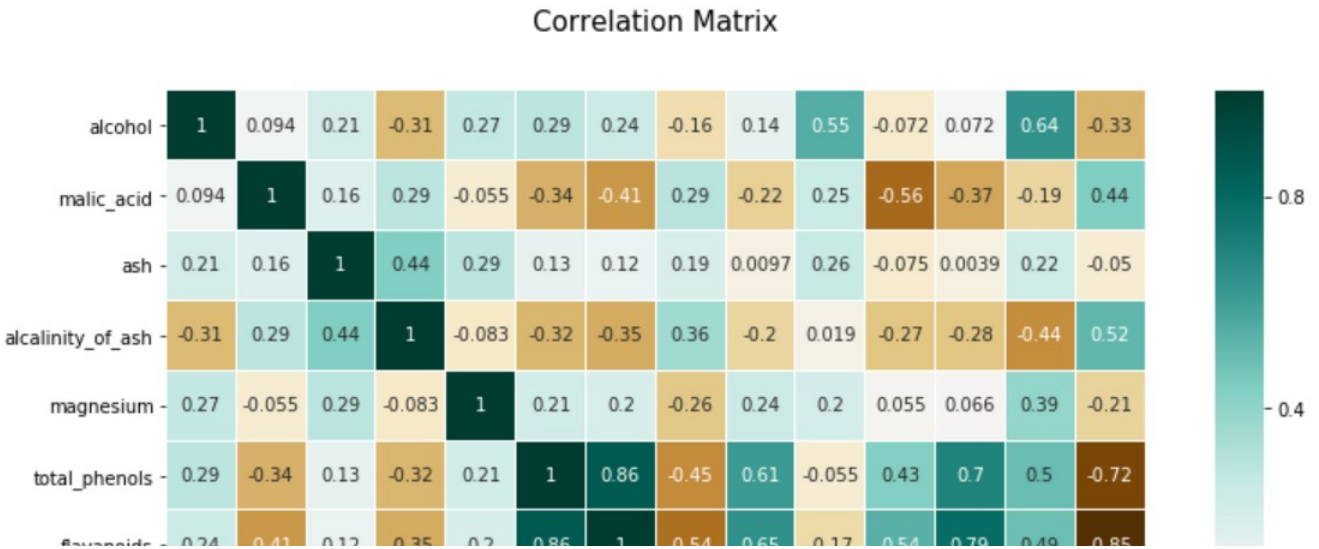




Exploratory Data Analysis

- 檢視各欄位的相關係數

```
: plt.figure(figsize=(14,10))
  plt.title('Correlation Matrix', y=1.05, size=15)
  sns.heatmap(df.astype(float).corr(), cmap = "BrBG",linewidths=0.1, square=True, linecolor='white',annot=True)

: <matplotlib.axes._subplots.AxesSubplot at 0x1313c8d59b0>
```



特徵縮放與拆分訓練集與測試集

- 將資料以 1:5 拆分訓練集與測試集
- 用 Scale 進行特徵縮放 (能提升 SVC 的準確度)

```
train_feats, test_feats, train_labels, test_labels = tts(features, labels, test_size=0.2,random_state=21)
train_feats = scale(train_feats)
test_feats = scale(test_feats)
```

以四個分類器產生模型

- 使用 SVC, Linear SVC, Decision Tree, Random Forest
- 用 Training dataset 產生模型

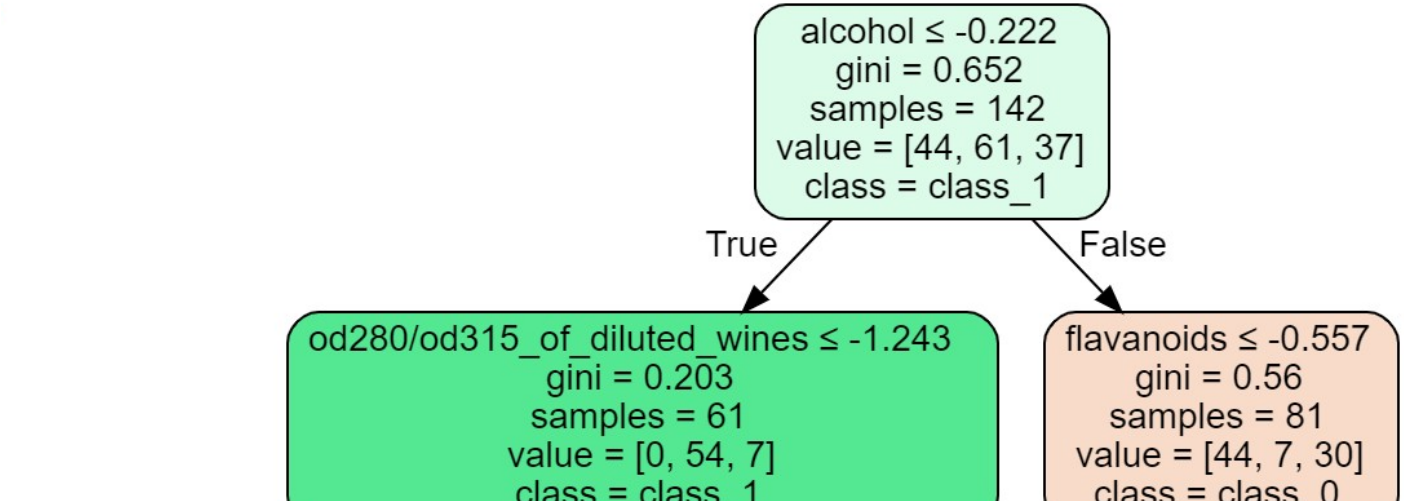
```
clf1 = svm.SVC()
clf2 = svm.SVC(kernel='linear')
clf3 = tree.DecisionTreeClassifier()
clf4 = RandomForestClassifier()
```

```
# training
clf1.fit(train_feats, train_labels)
clf2.fit(train_feats, train_labels)
clf3.fit(train_feats, train_labels)
clf4.fit(train_feats, train_labels)
```

以四個分類器產生模型

- 將 Decision Tree 模型圖像化

```
dot_data3 = export_graphviz(clf3, out_file=None,feature_names=wine.feature_names,class_names=w
graph = graphviz.Source(dot_data3)
graph
```



以四個分類器產生模型

- 以 Testing dataset 進行預測

```
# predictions
prediction1 = clf1.predict(test_feats)
print("\nPrediction1:", prediction1)
prediction2 = clf2.predict(test_feats)
print("\nPrediction2:", prediction2)
prediction3 = clf3.predict(test_feats)
print("\nPrediction3:", prediction3)
prediction4 = clf4.predict(test_feats)
print("\nPrediction4:", prediction4)
```

以四個分類器產生模型

- 計算各模型的 Accuracy

```
# Accuracy
print("SVC Accuracy:",accuracy_score(test_labels, prediction1)*100, "%")
print("Linear Accuracy:",accuracy_score(test_labels, prediction2)*100, "%")
print("DecisionTree Accuracy:",accuracy_score(test_labels, prediction3)*100, "%")
print("RandomForest Accuracy:",accuracy_score(test_labels, prediction4)*100, "%")
```

```
SVC Accuracy: 36.11111111111111 %
Linear Accuracy: 94.44444444444444 %
DecisionTree Accuracy: 91.66666666666666 %
RandomForest Accuracy: 91.66666666666666 %
```

THANK YOU