IRIS_Python

May 25, 2020

1 Plotly Example

Scott Campit

1.1 Preliminaries: Installing Essential Packages

First we need to install some Python packages

[1]:
```
import sys
!{sys.executable} -m pip install plotly pandas numpy scipy sklearn
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: plotly in
/home/scampit/.local/lib/python3.7/site-packages (4.5.4)
Requirement already satisfied: pandas in
/home/scampit/.local/lib/python3.7/site-packages (0.25.3)
Requirement already satisfied: numpy in /home/scampit/.local/lib/python3.7/site-
packages (1.18.1)
Requirement already satisfied: scipy in /home/scampit/.local/lib/python3.7/site-
packages (1.4.1)
Requirement already satisfied: sklearn in
/home/scampit/.local/lib/python3.7/site-packages (0.0)
Requirement already satisfied: six in /home/scampit/.local/lib/python3.7/site-
packages (from plotly) (1.14.0)
Requirement already satisfied: retrying>=1.3.3 in
/home/scampit/.local/lib/python3.7/site-packages (from plotly) (1.3.3)
Requirement already satisfied: pytz>=2017.2 in
/home/scampit/.local/lib/python3.7/site-packages (from pandas) (2019.3)
Requirement already satisfied: python-dateutil>=2.6.1 in
/home/scampit/.local/lib/python3.7/site-packages (from pandas) (2.8.1)
Requirement already satisfied: scikit-learn in
/home/scampit/.local/lib/python3.7/site-packages (from sklearn) (0.22.2.post1)
Requirement already satisfied: joblib>=0.11 in
/home/scampit/.local/lib/python3.7/site-packages (from scikit-learn->sklearn)
(0.14.1)
```

```
WARNING: You are using pip version 20.1; however, version 20.1.1 is
available.
You should consider upgrading via the '/usr/bin/python3 -m pip install --upgrade
pip' command.
```

1.2 Exploratory data analysis

Now that we have installed some libraries, let's make some plots of the Iris dataset. This dataset is used for so many machine learning tutorials because of its simplicity. The dataset contains 150 data points, corresponding to three species of Irises (Setosa, Versicolor, and Virginica). There are 4 attributes (or **features**) associated with the Iris that have been traditionally used to classify these three species: Sepal Length, Sepal Width, Petal Length, and Petal Width.

[2]:
```
from sklearn import datasets
import pandas as pd
import numpy as np
import graphobjs

iris = datasets.load_iris()

data = pd.DataFrame(data=np.c_[iris['data'], iris['target']],
                    columns=iris['feature_names']+['Species'])
print("First glance at the Iris dataset")
print(data.head(10))
```

```
First glance at the Iris dataset
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)  \
0                5.1               3.5                1.4               0.2
1                4.9               3.0                1.4               0.2
2                4.7               3.2                1.3               0.2
3                4.6               3.1                1.5               0.2
4                5.0               3.6                1.4               0.2
5                5.4               3.9                1.7               0.4
6                4.6               3.4                1.4               0.3
7                5.0               3.4                1.5               0.2
8                4.4               2.9                1.4               0.2
9                4.9               3.1                1.5               0.1

   Species
0      0.0
1      0.0
2      0.0
3      0.0
4      0.0
5      0.0
6      0.0
7      0.0
```

```
8      0.0
9      0.0
```

2 Create pairwise coordinate plot object

[3]:
```
pxplt = graphobjs.parallel(data)
pxplt.show()
```

3 Reformat dataset for other plots

[4]:
```
targets = {0.0:'setosa',
           1.0:'versicolor',
           2.0:'virginica'}
data['Species'] = data['Species'].replace(targets)
print("Data with new label encoding")
print(data.head(10))
```

```
Data with new label encoding
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)  \
0                5.1               3.5                1.4               0.2
1                4.9               3.0                1.4               0.2
2                4.7               3.2                1.3               0.2
3                4.6               3.1                1.5               0.2
4                5.0               3.6                1.4               0.2
5                5.4               3.9                1.7               0.4
6                4.6               3.4                1.4               0.3
7                5.0               3.4                1.5               0.2
8                4.4               2.9                1.4               0.2
9                4.9               3.1                1.5               0.1

  Species
0  setosa
1  setosa
2  setosa
3  setosa
4  setosa
5  setosa
6  setosa
7  setosa
8  setosa
9  setosa
```

[5]:
```
melted_data = pd.melt(data,
                      id_vars=['Species'],
                      value_vars=iris.feature_names)
```

```
print("Melted dataframe")
print(melted_data.head(10))
```

```
Melted dataframe
  Species           variable  value
0  setosa  sepal length (cm)    5.1
1  setosa  sepal length (cm)    4.9
2  setosa  sepal length (cm)    4.7
3  setosa  sepal length (cm)    4.6
4  setosa  sepal length (cm)    5.0
5  setosa  sepal length (cm)    5.4
6  setosa  sepal length (cm)    4.6
7  setosa  sepal length (cm)    5.0
8  setosa  sepal length (cm)    4.4
9  setosa  sepal length (cm)    4.9
```

4 Create other plots

[6]:
```
# Histograms
hist = graphobjs.histograms(melted_data)
hist.show()
```

[7]:
```
# Boxplots
boxplots = graphobjs.boxplots(melted_data)
boxplots.show()
```

[8]:
```
# Scatterplot matrices
splom = graphobjs.splom(data)
splom.show()
```