Module 6 | Python

November 30, 2021

1 Module 6 | Python

1.0.1 Ryan S. Dunn

1.1 Data Science Using Python and R: Chapter 10 - Page 149: Questions #11, 12, 13, & 14

[5]:

```
#import libraries for KMeans algorithm
import pandas as pd
from scipy import stats
from sklearn.cluster import KMeans
```

[4]:

```
#import the white wine training and test
wine_train = pd.read_csv("/Users/ryan_s_dunn/Documents/USD MS-ADS/Applied Data␣
 ↪Mining 502/Module 6/datasets/white_wine_training", header = 0)
wine_test = pd.read_csv("/Users/ryan_s_dunn/Documents/USD MS-ADS/Applied Data␣
 ↪Mining 502/Module 6/datasets/white_wine_test", header = 0)
```

1.1.1 11. Input and standardize both the training and test data sets.

[14]:

```
#standardize the training and test data sets (z-score)
wine_train_z = pd.DataFrame(stats.zscore(wine_train), columns =␣
 ↪['alcohol','quality','sugar'])
wine_test_z = pd.DataFrame(stats.zscore(wine_test), columns =␣
 ↪['alcohol','quality','sugar'])
```

1.1.2 12. Run k-means clustering on the training data set, using two clusters.

[41]:

```
#run the k-means clustering algorithm on the training data
kmeans01 = KMeans(n_clusters = 2).fit(wine_train_z)
```

[10]:

```
#identify the cluster memebership
cluster = kmeans01.labels_
```

[11]:

```
#seperate the records into two groups based on cluster memebership
Cluster1 = wine_train_z.loc[cluster ==0]
Cluster2 = wine_train_z.loc[cluster ==1]
```

1.1.3 13. Give the mean of each variable within each cluster and use the means to identify a “Dry wines” and a “Sweet wines” cluster.

[12]:

```
#compute summary statistics of cluster 1
Cluster1.describe()
```

[12]:

```
          alcohol     quality       sugar
count  992.000000  992.000000  992.000000
mean    -0.689756   -0.553389    0.424439
std      0.560951    0.778523    1.068893
min     -1.826971   -3.252193   -1.122791
25%     -1.096280   -0.958094   -0.609069
50%     -0.824881   -0.958094    0.396970
75%     -0.323836    0.188956    1.210364
max      1.847359    2.483055    5.512788
```

[29]:

```
Cluster1.mean()
```

[29]:

```
alcohol   -0.689756
quality   -0.553389
sugar      0.424439
dtype: float64
```

Cluster1 is a sweet wine - notice the mean of the sugar attribute.

[13]:

```
#compute summary statistics of cluster 2
Cluster2.describe()
```

[13]:

```
          alcohol     quality       sugar
count  817.000000  817.000000  817.000000
mean     0.837501    0.671924   -0.515353
std      0.744389    0.810249    0.586883
min     -1.075403   -2.105143   -1.101386
25%      0.344224    0.188956   -0.940848
50%      0.761762    0.188956   -0.780310
75%      1.429822    1.336005   -0.223777
max      2.891203    3.630104    2.066568
```

[30]:

```
Cluster2.mean()
```

[30]:

```
alcohol    0.837501
quality    0.671924
sugar     -0.515353
dtype: float64
```

Cluster 2 is a dry wine - notice the mean of the sugar attribute

1.1.4 14. Validate the clustering results by running k-means clustering on the test data set, using two clusters, and identifying a “Dry wines” and a “Sweet wines” cluster.

[15]:

```
#run the k-means clustering algorithm on the test data
kmeans_test = KMeans(n_clusters = 2).fit(wine_test_z)
```

[16]:

```
#identify the cluster memebership
cluster_test = kmeans_test.labels_
```

[17]:

```
#seperate the records into two groups based on cluster memebership
cluster1_test = wine_test_z.loc[cluster_test==0]
cluster2_test = wine_test_z.loc[cluster_test==1]
```

[27]:

```
#compute summary statistics of test cluster 1
cluster1_test.describe() #dry wine - notice sugar mean
```

[27]:

```
          alcohol     quality       sugar
count  868.000000  868.000000  868.000000
mean     0.756414    0.590031   -0.532449
std      0.777203    0.831249    0.552722
min     -1.432916   -2.063322   -1.068851
25%      0.186001    0.139557   -0.937516
50%      0.671676    0.139557   -0.780428
75%      1.400189    1.240997   -0.244785
max      2.776268    3.443877    1.877186
```

[31]:

```
cluster1_test.mean()
```

[31]:

```
alcohol    0.756414
quality    0.590031
sugar     -0.532449
dtype: float64
```

cluster1_test is a dry wine - notice the mean of the sugar attribute.

[40]:

```
#compute summary statistics of test cluster 2
cluster2_test.describe() #sweet wine - notice sugar mean
```

[40]:

```
          alcohol     quality       sugar
count  892.000000  892.000000  892.000000
mean    -0.736062   -0.574156    0.518123
std      0.536420    0.796097    1.064472
min     -2.080483   -3.164762   -1.089453
25%     -1.109133   -0.961882   -0.285988
50%     -0.866295   -0.961882    0.414468
75%     -0.380620    0.139557    1.341542
max      1.643026    2.342437    3.298700
```

[32]:

```
cluster2_test.mean()
```

[32]:

```
alcohol   -0.736062
quality   -0.574156
sugar      0.518123
dtype: float64
```

cluster2_test is a sweet wine - notice the mean of the sugar attribute.