Lab 15 10-11-2022

November 10, 2022

1 K-Means Clustering

Program to implement k-means clustering technique using any standard dataset available in the public domain

Ajay Jeevan Jose

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

1.1 Loading Dataset

[2]:
```
df = pd.read_csv('Mall_Customers.csv')
df.head()
```

[2]:
```
   CustomerID   Genre  Age  Annual_Income_(k$)  Spending_Score
0           1    Male   19                  15              39
1           2    Male   21                  15              81
2           3  Female   20                  16               6
3           4  Female   23                  16              77
4           5  Female   31                  17              40
```

1.2 Dataset Preprocessing

1.2.1 Removing duplicates

[3]:
```
df.drop_duplicates(inplace=True)
df.describe()
```

[3]:
```
       CustomerID         Age  Annual_Income_(k$)  Spending_Score
count  200.000000  200.000000          200.000000      200.000000
mean   100.500000   38.850000           60.560000       50.200000
std     57.879185   13.969007           26.264721       25.823522
min      1.000000   18.000000           15.000000        1.000000
25%     50.750000   28.750000           41.500000       34.750000
50%    100.500000   36.000000           61.500000       50.000000
```

```
75%     150.250000   49.000000             78.000000         73.000000
max     200.000000   70.000000            137.000000         99.000000
```

1.2.2 Extracting Independent variables

selecting only annual income and spending score

```
[4]: x = df.iloc[:,[3,4]].values
     print(x)
```

```
[[ 15  39]
 [ 15  81]
 [ 16   6]
 [ 16  77]
 [ 17  40]
 [ 17  76]
 [ 18   6]
 [ 18  94]
 [ 19   3]
 [ 19  72]
 [ 19  14]
 [ 19  99]
 [ 20  15]
 [ 20  77]
 [ 20  13]
 [ 20  79]
 [ 21  35]
 [ 21  66]
 [ 23  29]
 [ 23  98]
 [ 24  35]
 [ 24  73]
 [ 25   5]
 [ 25  73]
 [ 28  14]
 [ 28  82]
 [ 28  32]
 [ 28  61]
 [ 29  31]
 [ 29  87]
 [ 30   4]
 [ 30  73]
 [ 33   4]
 [ 33  92]
 [ 33  14]
 [ 33  81]
 [ 34  17]
 [ 34  73]
 [ 37  26]
 [ 37  75]
 [ 38  35]
 [ 38  92]
 [ 39  36]
 [ 39  61]
 [ 39  28]
 [ 39  65]
 [ 40  55]
 [ 40  47]
 [ 40  42]
 [ 40  42]
 [ 42  52]
 [ 42  60]
 [ 43  54]
 [ 43  60]
 [ 43  45]
 [ 43  41]
 [ 44  50]
 [ 44  46]
 [ 46  51]
 [ 46  46]
 [ 46  56]
 [ 46  55]
 [ 47  52]
 [ 47  59]
 [ 48  51]
 [ 48  59]
 [ 48  50]
 [ 48  48]
 [ 48  59]
 [ 48  47]
 [ 49  55]
 [ 49  42]
 [ 50  49]
 [ 50  56]
 [ 54  47]
 [ 54  54]
 [ 54  53]
 [ 54  48]
 [ 54  52]
 [ 54  42]
 [ 54  51]
 [ 54  55]
 [ 54  41]
 [ 54  44]
 [ 54  57]
 [ 54  46]
 [ 57  58]
 [ 57  55]
 [ 58  60]
 [ 58  46]
 [ 59  55]
 [ 59  41]
 [ 60  49]
 [ 60  40]
 [ 60  42]
 [ 60  52]
 [ 60  47]
 [ 60  50]
 [ 61  42]
 [ 61  49]
 [ 62  41]
 [ 62  48]
 [ 62  59]
 [ 62  55]
 [ 62  56]
 [ 62  42]
 [ 63  50]
 [ 63  46]
 [ 63  43]
 [ 63  48]
 [ 63  52]
 [ 63  54]
 [ 64  42]
 [ 64  46]
 [ 65  48]
 [ 65  50]
 [ 65  43]
 [ 65  59]
 [ 67  43]
 [ 67  57]
 [ 67  56]
 [ 67  40]
 [ 69  58]
 [ 69  91]
 [ 70  29]
 [ 70  77]
 [ 71  35]
 [ 71  95]
 [ 71  11]
 [ 71  75]
 [ 71   9]
 [ 71  75]
 [ 72  34]
 [ 72  71]
 [ 73   5]
 [ 73  88]
 [ 73   7]
 [ 73  73]
 [ 74  10]
 [ 74  72]
 [ 75   5]
 [ 75  93]
 [ 76  40]
 [ 76  87]
 [ 77  12]
 [ 77  97]
 [ 77  36]
 [ 77  74]
 [ 78  22]
 [ 78  90]
 [ 78  17]
 [ 78  88]
 [ 78  20]
 [ 78  76]
 [ 78  16]
 [ 78  89]
 [ 78   1]
 [ 78  78]
 [ 78   1]
 [ 78  73]
 [ 79  35]
 [ 79  83]
 [ 81   5]
 [ 81  93]
 [ 85  26]
 [ 85  75]
 [ 86  20]
 [ 86  95]
 [ 87  27]
 [ 87  63]
 [ 87  13]
 [ 87  75]
 [ 87  10]
 [ 87  92]
 [ 88  13]
 [ 88  86]
 [ 88  15]
 [ 88  69]
 [ 93  14]
 [ 93  90]
 [ 97  32]
 [ 97  86]
```

```
 [ 98  15]
 [ 98  88]
 [ 99  39]
 [ 99  97]
 [101  24]
 [101  68]
 [103  17]
 [103  85]
 [103  23]
 [103  69]
 [113   8]
 [113  91]
 [120  16]
 [120  79]
 [126  28]
 [126  74]
 [137  18]
 [137  83]]
```

1.3 Fitting data using ELBOW Method

[5]:

```python
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(x)
    wcss.append(kmeans.inertia_)

print(wcss)
```

```
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f4373f27ee0>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
```

```
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f4373f27ee0>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f4373f27ee0>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f4373f27ee0>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
```

```
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f436e33fc10>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f436e33fc10>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f436e33fc10>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
```

```
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f436e33fc10>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f436e33fc10>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f436e33fc10>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
```

```
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
Exception ignored on calling ctypes callback function: <function _ThreadpoolInfo
._find_modules_with_dl_iterate_phdr.<locals>.match_module_callback at
0x7f436e33fc10>
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 400,
in match_module_callback
    self._make_module_from_path(filepath)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 515,
in _make_module_from_path
    module = module_class(filepath, prefix, user_api, internal_api)
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 606,
in __init__
    self.version = self.get_version()
  File "/opt/anaconda3/lib/python3.9/site-packages/threadpoolctl.py", line 646,
in get_version
    config = get_config().split()
AttributeError: 'NoneType' object has no attribute 'split'
```

```
[269981.28000000014, 181363.59595959607, 106348.37306211119, 73679.78903948837,
44448.45544793369, 37233.81451071002, 30259.657207285458, 25011.839349156595,
21850.16528258562, 19672.07284901432]
```

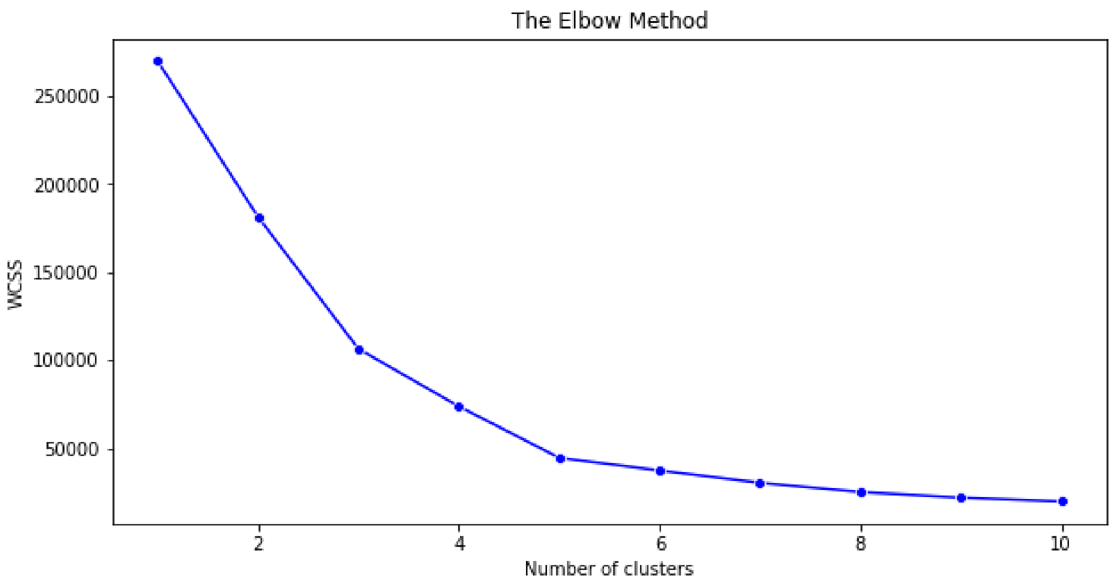
1.4 Visualisation

1.4.1 Lineplot

[6]:

```
plt.figure(figsize=(10,5))
sns.lineplot(range(1,11), wcss,marker='o',color='b')
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

```
/opt/anaconda3/lib/python3.9/site-packages/seaborn/_decorators.py:36:
FutureWarning: Pass the following variables as keyword args: x, y. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
```



1.5 Fitting K-Means to the dataset

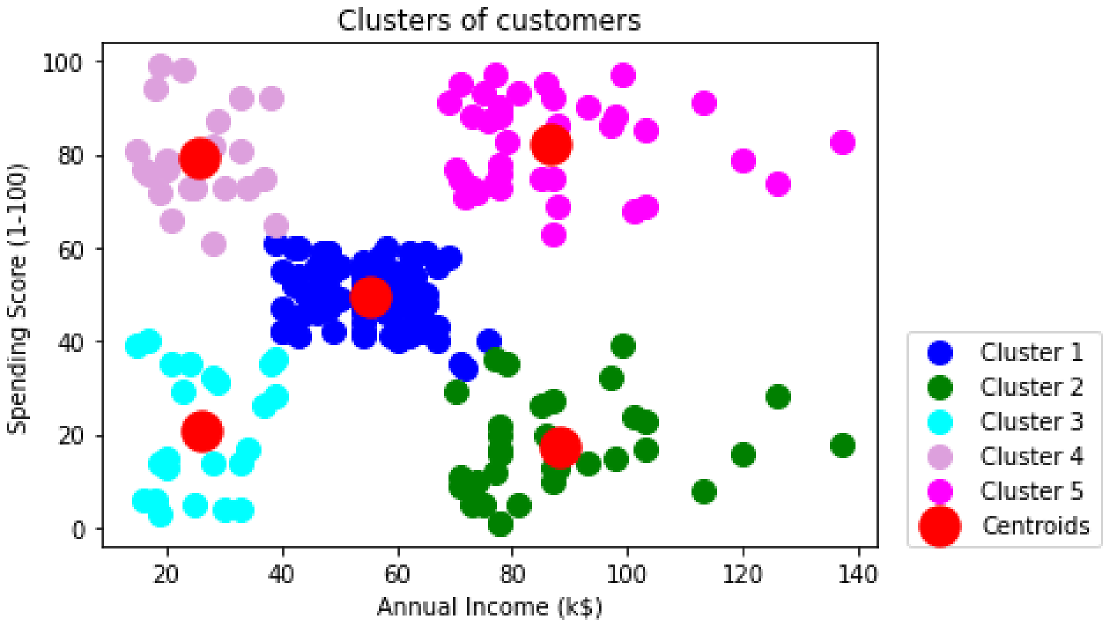
```
[7]: kmeans = KMeans(n_clusters = 5, init = 'k-means++', random_state = 42)
     y_pred = kmeans.fit_predict(x)
     print(y_pred)
```

```
[2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2
 3 2 3 2 3 2 0 2 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 4 1 4 0 4 1 4 1 4 0 4 1 4 1 4 1 4 1 4 0 4 1 4 1 4
 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1
 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4]
```

1.6 Visualisation of clusters

```
[21]: plt.scatter(x[y_pred == 0, 0], x[y_pred == 0, 1], s = 100, c = 'b', label =␣
        ↪'Cluster 1')
      plt.scatter(x[y_pred == 1, 0], x[y_pred == 1, 1], s = 100, c = 'g', label =␣
        ↪'Cluster 2')
      plt.scatter(x[y_pred == 2, 0], x[y_pred == 2, 1], s = 100, c = 'cyan', label =␣
        ↪'Cluster 3')
      plt.scatter(x[y_pred == 3, 0], x[y_pred == 3, 1], s = 100, c = 'plum', label =␣
        ↪'Cluster 4')
      plt.scatter(x[y_pred == 4, 0], x[y_pred == 4, 1], s = 100, c = 'magenta', label␣
        ↪= 'Cluster 5')
      plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s =␣
        ↪300, c = 'r', label = 'Centroids')
```

```
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend(loc=(1.04, 0))
plt.show()
```


Clusters of customers
Spending Score (1-100)
Annual Income (k$)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Centroids