Sane's STATS Academy of Statistics

Cluster Analysis

Need for Clustering

- Groups available for grouping Customers
  - Male and female
  - High Income, Medium Income, Low Income
  - Having average invoice amounts as < 1Lac, between 5 Lac and 10 Lac etc.
  - Customers making Cheque, Cash or card payments
  - Any other type of grouping
- As per our convenience we categorize the entities based on some attributes

Need for Clustering

- There may be a business need of grouping the business entities on some variables.
- Such as we may group the products not just on the profit margin but also considering the variables such as quantity sold, average shelf life, maintenance cost etc.

Why is Clustering done?

- Cluster analysis is used to form groups or clusters of similar records based on several measurements made on these records.

Clustering in Marketing

- Market Segmentation: Customers are segmented based on demographic and transaction history information and a marketing strategy is tailored for each segment.
- Market structure analysis: Identifying groups of similar products according to competitive measures of similarity.

Clustering in Finance

- Creating balanced portfolios: Given data on a variety of investment opportunities (e.g., stocks), one may find clusters based on financial performance variables such as return (daily, weekly, or monthly), volatility and other characteristics, such as industry and market capitalization.

Sane's STATS Academy of Statistics

Clustering in Pure Sciences

- Biologists have made extensive use of classes and subclasses to organize species.
- In chemistry, Mendeleeyev’s periodic table of the elements.

Distance Method

- For record i we have the vector of p measurements (xi1, xi2, . . . , xip), while for record j we have the vector of measurements (xj1, xj2, . . . , xjp).
- The most popular distance measure is the Euclidean distance, dij , which between two cases, i and j, is defined by

$$d_{ij} = \sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{ip}-x_{jp})^2}$$

Other Distance Measures

- Numerical Data
  - Correlation-based similarity
  - Statistical distance (also called Mahalanobis distance)
  - Manhattan distance (“city block”)
  - Maximum coordinate distance
- Categorical Data
  - Matching coefficient: (a + d)/p
  - Jaquard’s coefficient: d/(b+c+d)

Scaling the data

- If the variables in analysis vary in range, the variables with large values will have larger impact on the results.
- To avoid this undesirable, we should scale the data
- Variable can be scaled in the following ways:
  - Subtract mean from each value(centering) and divide it by its standard deviation(scaling)
  - Divide each value in the variable by maximum value of the variable
  - Subtract mean from each value(centering) and divide it by its mean deviation about mean(scaling)

Types of Clustering Methods

- Hierarchical methods
  - **Agglomerative**
  - Divisive
- Nonhierarchical methods
  - **K-Means**
  - K-Medoids

Example: Milk

- The data set contains the ingredients of mammal's milk of 25 animals.
- A data frame with 25 observations on the following 5 variables (all in percent)
  - water
  - protein
  - fat
  - lactose
  - ash
- Here, we are interested in grouping the 25 mammals based on the above given 5 nutrient measures

K-Means Clustering

K-Means

- Forming good clusters is to pre-specify a desired number of clusters, k, and to assign each case to one of k clusters so as to minimize a measure of dispersion (variation) within the clusters.
- The method divides the sample into a predetermined number k of non-overlapping clusters so that clusters are as homogeneous as possible with respect to the measurements used.

Algorithm

1. Start with k initial clusters (user chooses k). The starting points are chosen by software at random
2. At every step, each record is reassigned to the cluster with the "closest" centroid.
3. Re-compute the centroids of clusters that lost or gained a record, and repeat step 2.
4. Stop when moving any more records between clusters increases cluster dispersion.

K-means in Python

- K-means can be implemented in R by function kmeans( ) in *stat* package.

Syntax : sklearn.cluster.k_means(X, n_clusters, max_iter=300, random_state=None, …)

Where

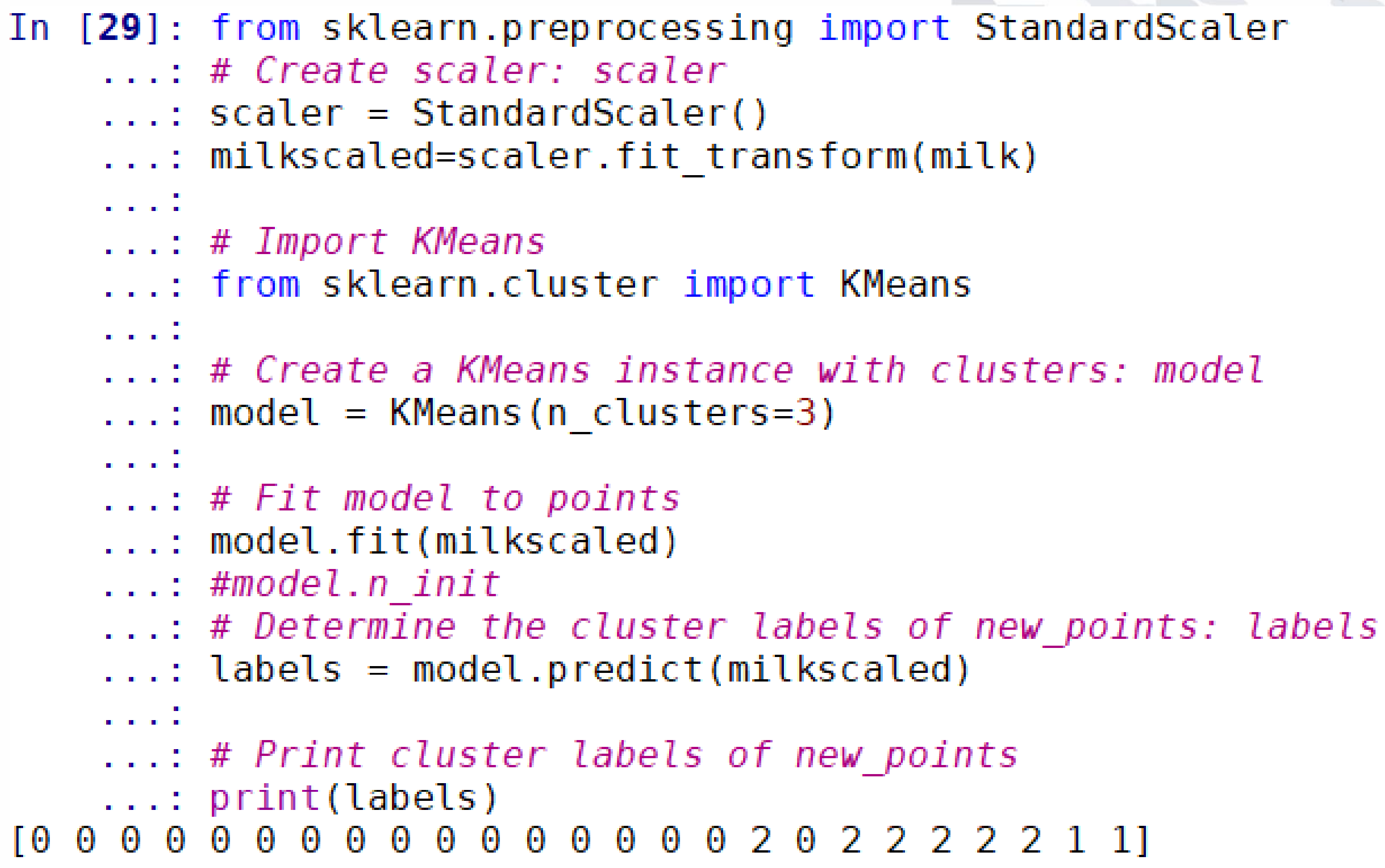
X: array-like or sparse matrix, shape

n_clusters : number of clusters

max_iter : Maximum number of iterations of the k-means algorithm to run

random_state : int, RandomState instance or None (default)

Program & Output

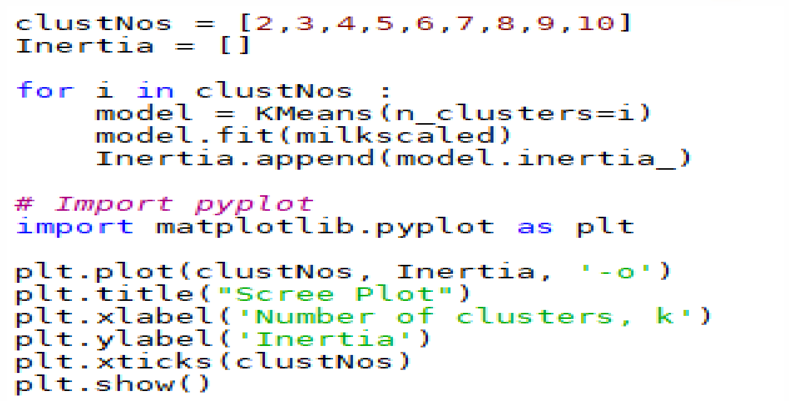


```
In [29]: from sklearn.preprocessing import StandardScaler
    ...: # Create scaler: scaler
    ...: scaler = StandardScaler()
    ...: milkscaled=scaler.fit_transform(milk)
    ...:
    ...: # Import KMeans
    ...: from sklearn.cluster import KMeans
    ...:
    ...: # Create a KMeans instance with clusters: model
    ...: model = KMeans(n_clusters=3)
    ...:
    ...: # Fit model to points
    ...: model.fit(milkscaled)
    ...: #model.n_init
    ...: # Determine the cluster labels of new_points: labels
    ...: labels = model.predict(milkscaled)
    ...:
    ...: # Print cluster labels of new_points
    ...: print(labels)
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 2 2 2 2 2 1 1]
```

Within Sum of Squares

- Within Sum of Squares indicate the variation
- We can create a function to extract and plot it.
- Lower the WSS better is the cluster variation.
- In python, **.inertia_** it can be extracted

Program & Output

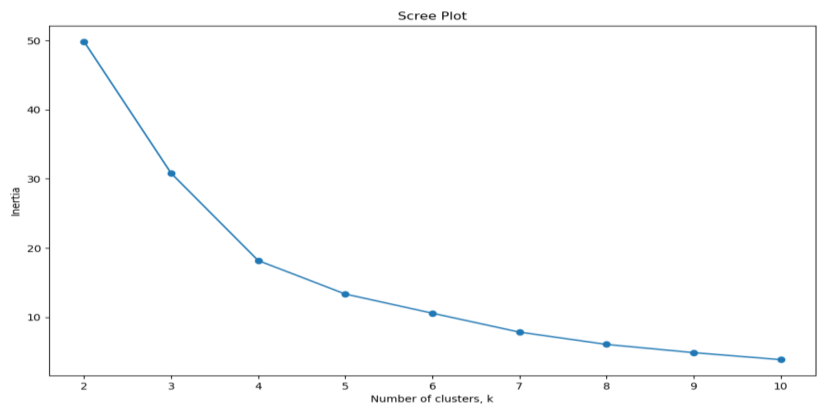


```
clustNos = [2,3,4,5,6,7,8,9,10]
Inertia = []

for i in clustNos :
    model = KMeans(n_clusters=i)
    model.fit(milkscaled)
    Inertia.append(model.inertia_)

# Import pyplot
import matplotlib.pyplot as plt

plt.plot(clustNos, Inertia, '-o')
plt.title("Scree Plot")
plt.xlabel('Number of clusters, k')
plt.ylabel('Inertia')
plt.xticks(clustNos)
plt.show()
```



Advantages of K-Means Clustering

- K-means clustering can handle larger datasets than hierarchical cluster approach.
- Observations are not permanently committed to any cluster but, they are changed at every iteration.

Limitations

- All the variables have to be continuous / numeric
- Clusters are severely affected by outliers
- Clusters are sensitive to initialization
- Clusters obtained are of differing densities