Cluster Analysis -- Basic Concepts and Algorithms

Peerapon S.

Machine Learning (2/67)

Topics

Basic K-means algorithm

Choosing number of clusters

Insights from clustering

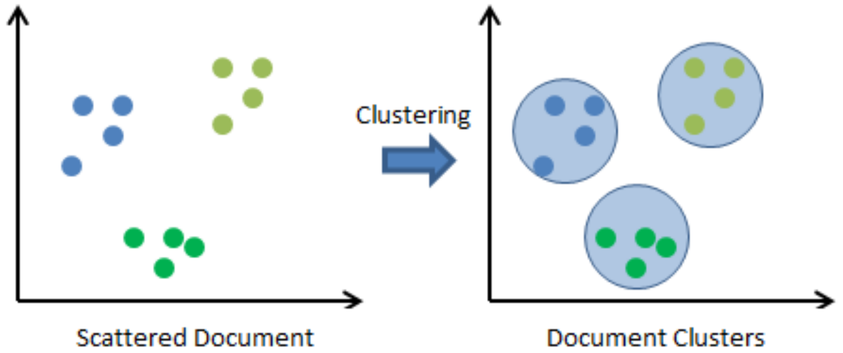
Clustering

Questions

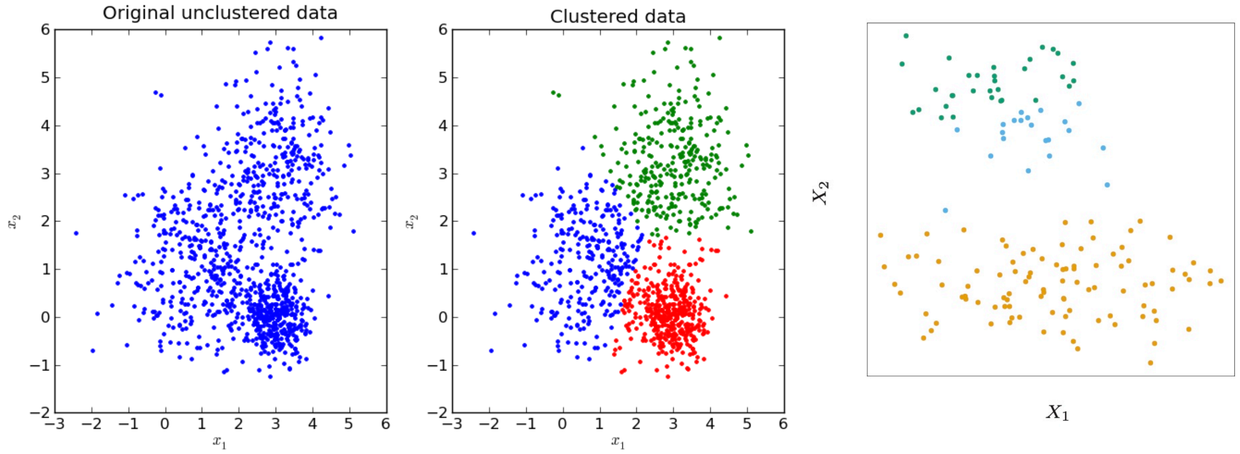
What are customer behaviors?
Who are our customers?
What are groups of species or syndromes, documents, etc. ?
No Y data !

Goal

Find natural grouping of data



A Clustering =
A set of clusters



Centroid of a Cluster

Imaginary location representing the center of a cluster.

For a d-dimensional data set,

- The centroid of a cluster is the vector with d elements.
- Average of (scaled) numeric or (encoded) ordinal variables

In representative/prototype-based clustering, objects are assigned to their closest centroids.

	Age	Height	Weight
Person 1	45	170	64
Person 2	31	172	77
Person 3	29	159	56
Person 4	64	163	72

Centroid $\boldsymbol{\mu}$ = {42.25, 166, 67.25}

K-Means Clustering

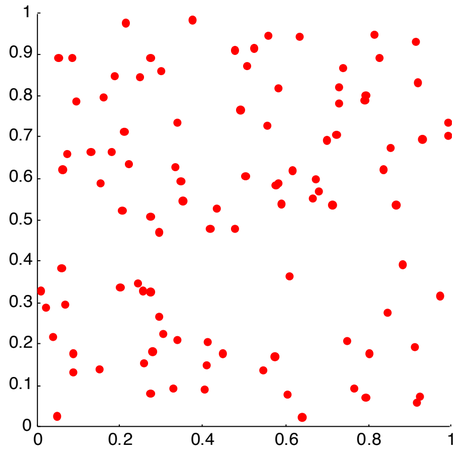
Given n data points of dimension d and a clustering $C = \{C_1, C_2, \ldots, C_K\}$ with centroids $\boldsymbol{\mu}_k$, $k = 1,2,\ldots, K$.

Define the sum of squared errors of cluster as

$$\text{SSE}(\mathcal{C}) = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_j \in C_k} \|\boldsymbol{x}_j - \boldsymbol{\mu}_k\|^2$$

Find the clustering $C^* = \{C_1, C_2, \ldots, C_K\}$ that minimizes SSE: $\mathcal{C}^* = \arg\min_{\mathcal{C}} \text{SSE}(\mathcal{C})$

Optimal clustering requires exhaustive search !

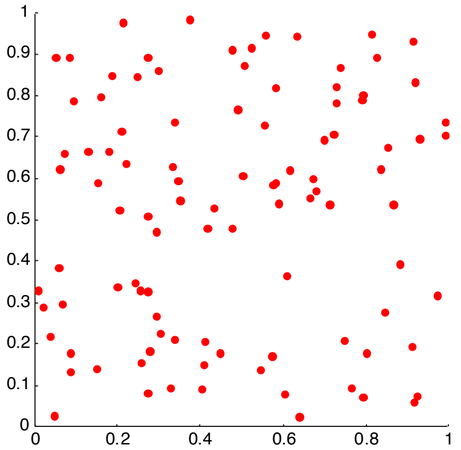


Greedy iterative approach is used to find a clustering that minimizes the SSE objective.

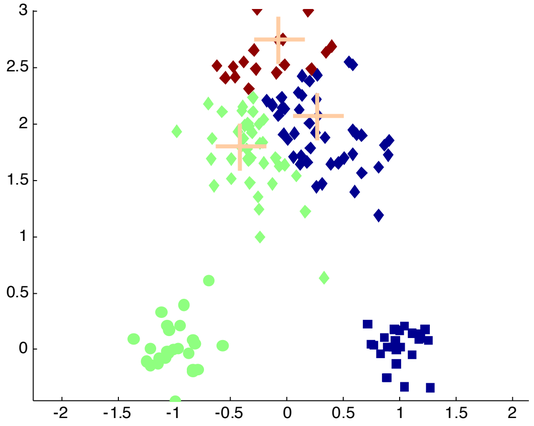
Provide good-enough solutions but not guarantee optimality.

Lloyd's K-means algorithm

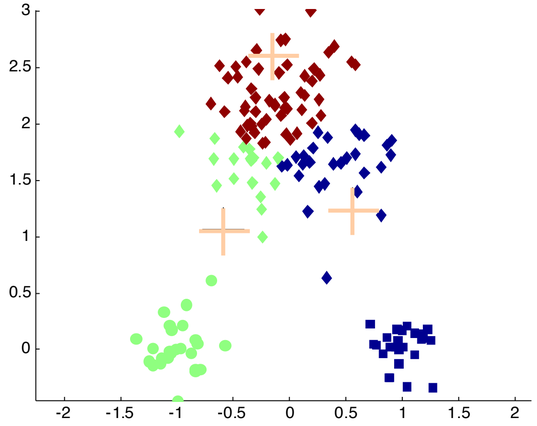
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change



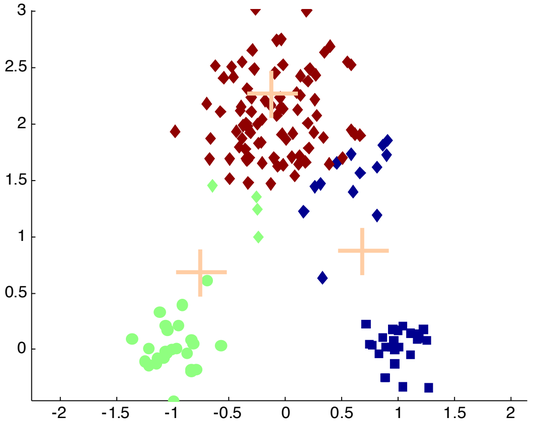
$K = 3$



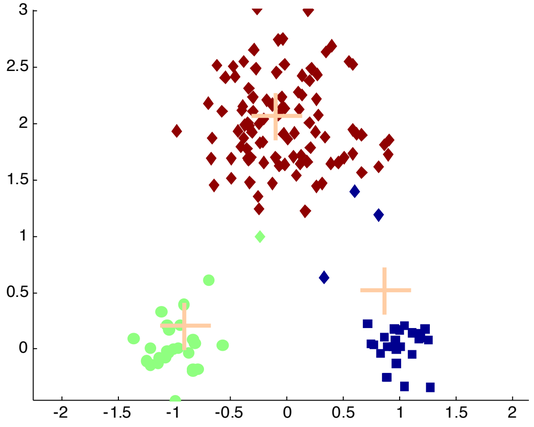
Iteration 1



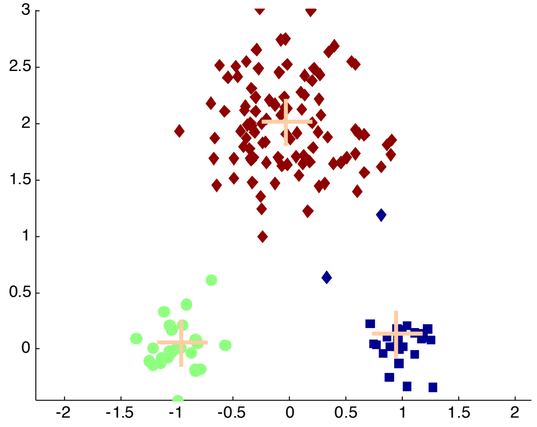
Iteration 2



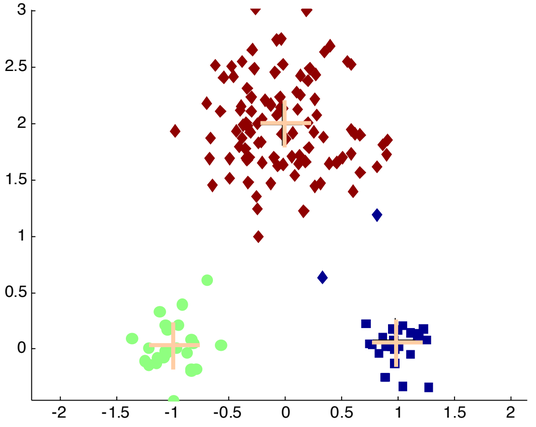
Iteration 3



Iteration 4

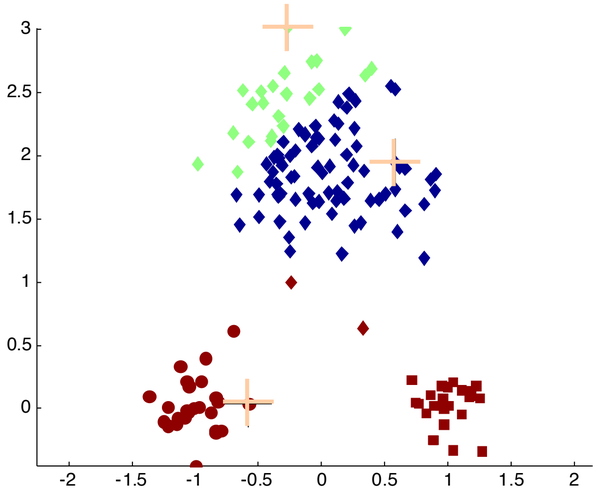


Iteration 5

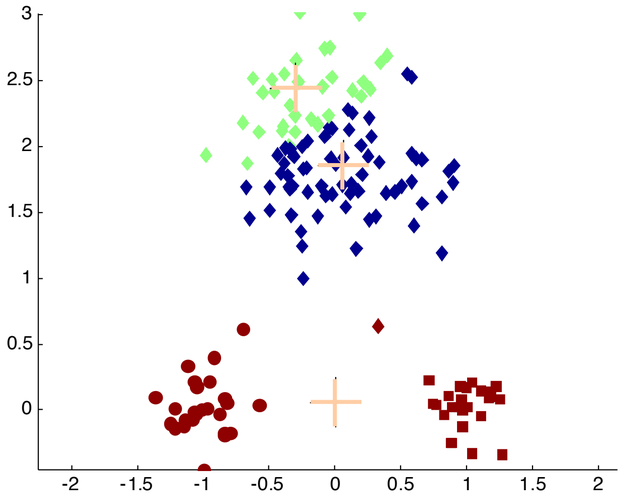


Iteration 6

Importance of Initial Conditions



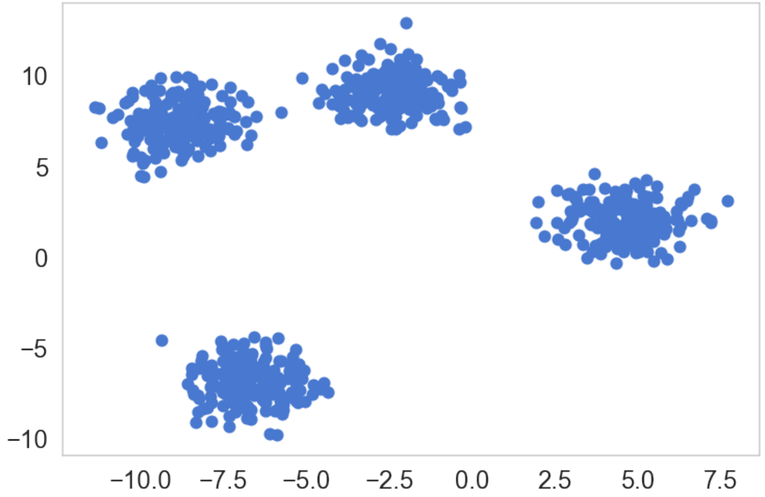
Initialization



Termination

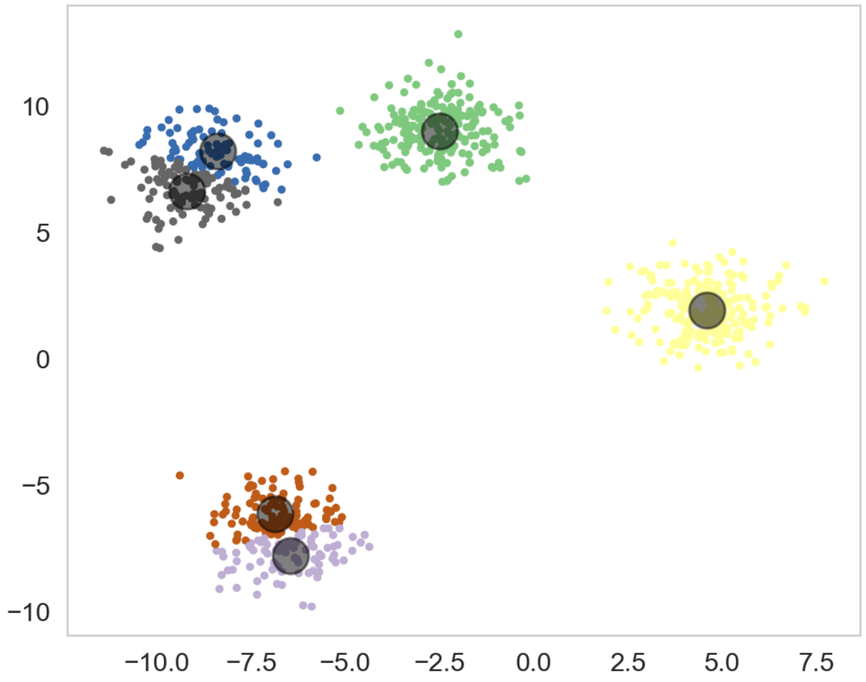
K-means++ spreads out initial centroids.

```python
from sklearn.datasets import make_blobs

X, _ = make_blobs(n_samples=800, centers=4, random_state=42)
plt.scatter(X[:, 0], X[:, 1]);

# Import the KMeans function, set K, and fit the model
from sklearn.cluster import KMeans

K = 6
kmeans = KMeans(n_clusters=K, n_init=10).fit(X)

# Get the cluster labels
labels = kmeans.predict(X) # or use kmeans.labels

# Get sum of squares distance of all points
sse = kmeans.inertia_

# Centroid values
centroids = kmeans.cluster_centers_

np.set_printoptions(precision=3)
print(f"Sum of squared errors : {sse:.3f}")
print(f"\nCentroids : \n {centroids}")
```

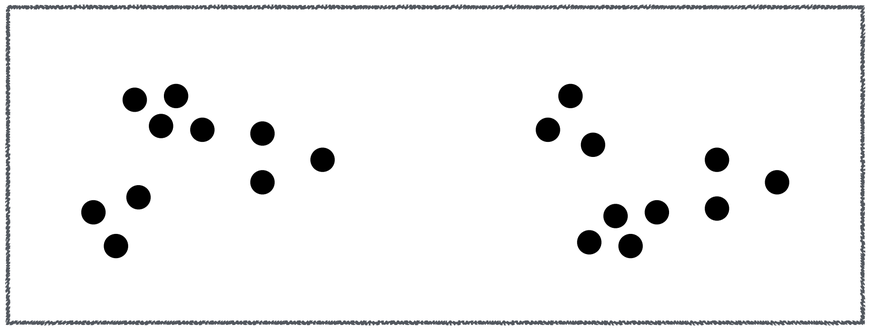


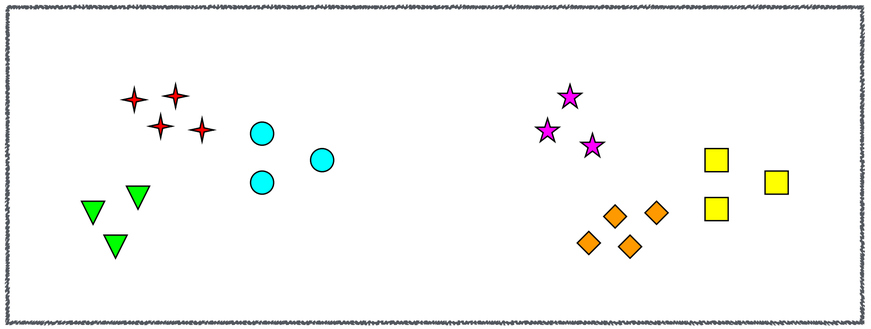

```
Sum of squared errors : 1273.890

Centroids :
 [[-8.262  8.17 ]
 [-6.682 -6.81 ]
 [ 5.041  1.4  ]
 [-2.501  9.053]
 [ 3.96   2.726]
 [-9.276  6.737]]
```

```
# Visualize the clustering

import matplotlib.pyplot as plt

plt.scatter(X[:, 0], X[:, 1], c=labels, s=5, cmap='Accent')

centers = kmeans.cluster_centers_
plt.scatter(centers[:, 0], centers[:, 1], c='black', s=200, alpha=0.5);
```

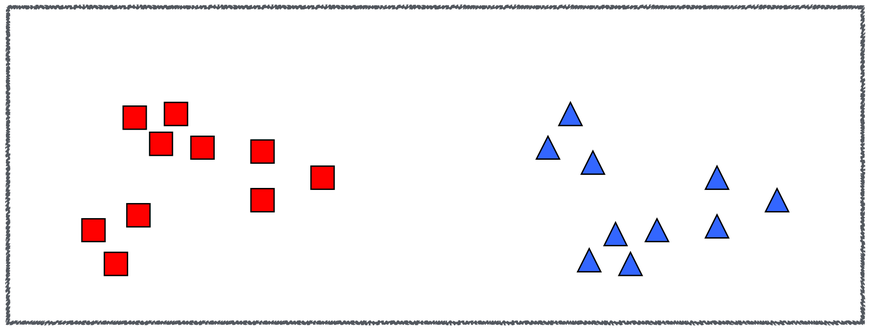


How many clusters ?

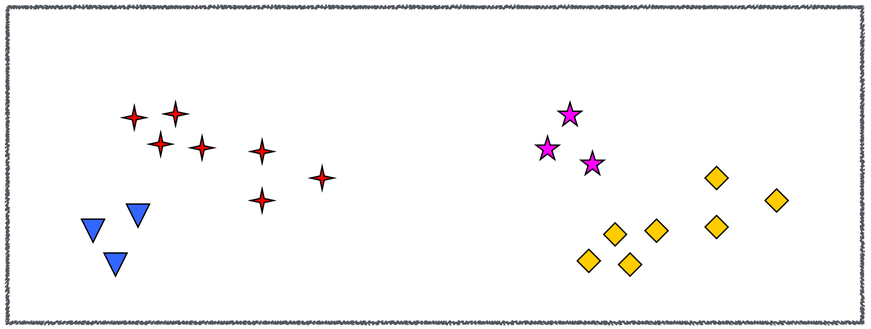




Six Clusters



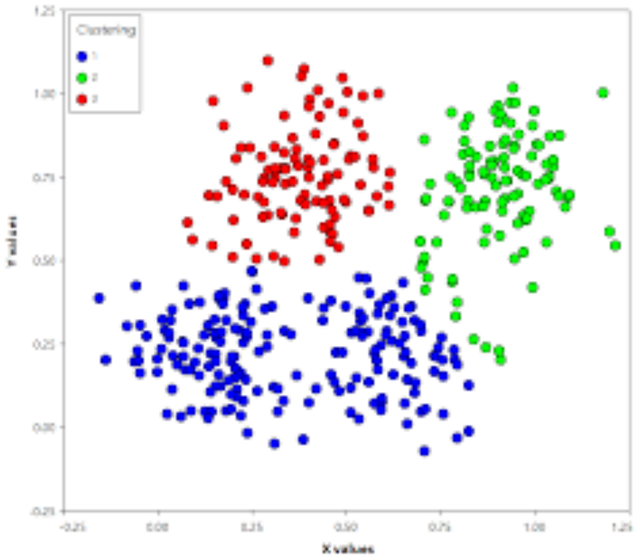
Two Clusters

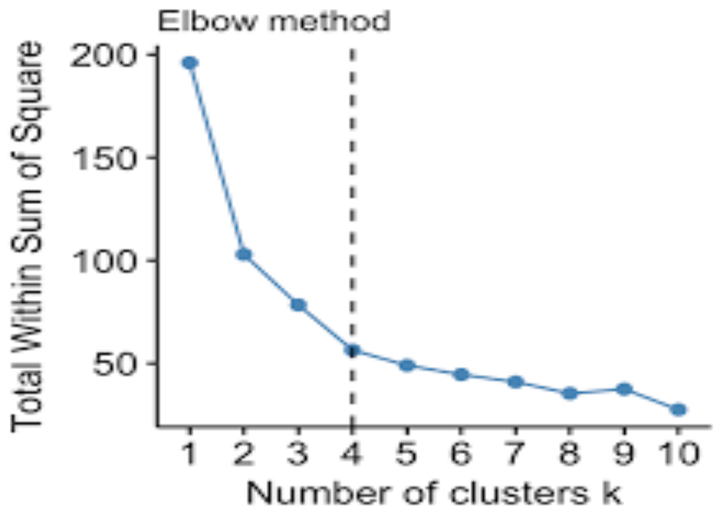


Four Clusters

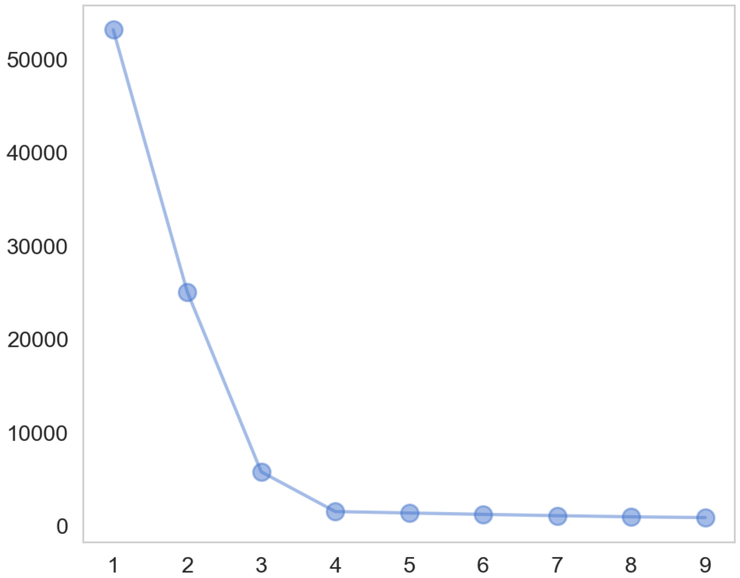
Choosing the Right K: Elbow Method

Look at the smallest possible K where SSE no longer has significant improvement, i.e., the ***knee point***.

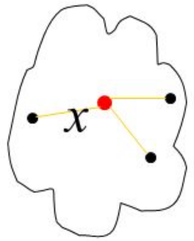




```
# Determine SSE for different K
sse = {}
for k in range(1, 10):
    kmeans = KMeans(n_clusters=k).fit(X)
    sse[k] = kmeans.inertia_

# Plot SSE vs. K
_,ax=plt.subplots(figsize=(5,4))
ax.plot(list(sse.keys()), list(sse.values()), marker='o', alpha=0.5, ms=8);
plt.tight_layout();
```

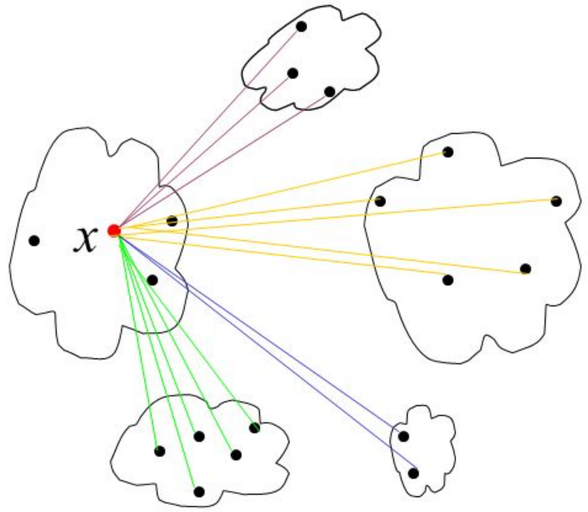


Silhouette Coefficient



Cohesion

a_i: Average distance to points within cluster



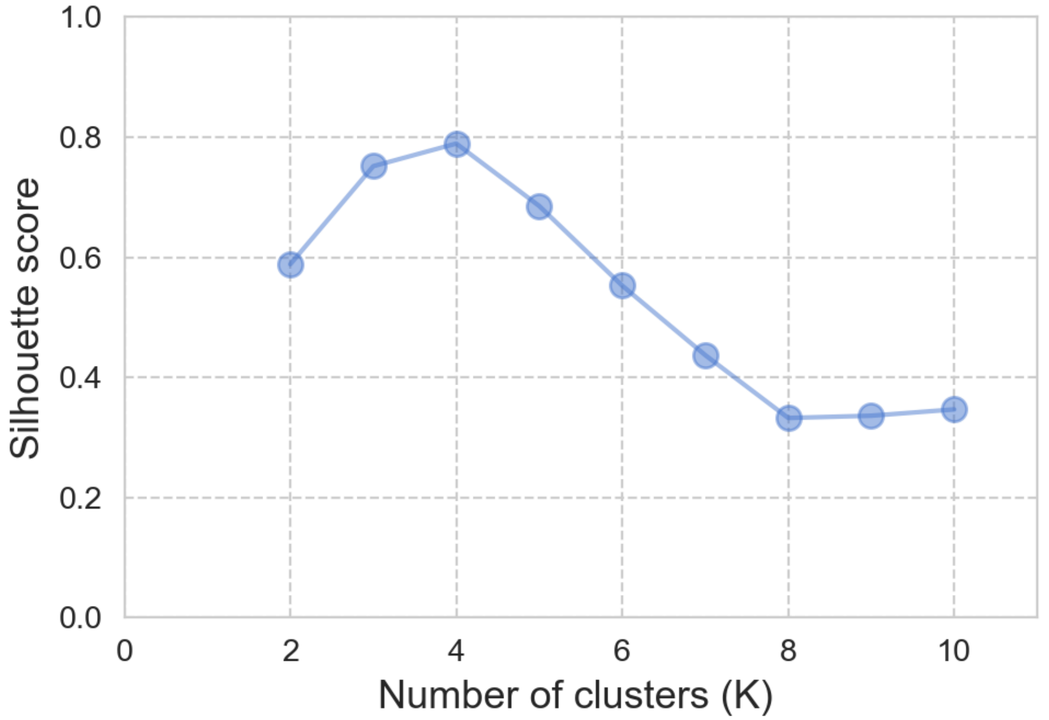
Separation

b_i: Minimum average distance to points in a different cluster

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} = \begin{cases} 1 - a_i/b_i & \text{if } a_i < b_i \\ 0 & \text{if } a_i = b_i \\ b_i/a_i - 1 & \text{if } a_i > b_i \end{cases}$$

$$\text{Silhouette coefficient (SC)} = n^{-1} \sum_{i=1}^{n} s_i$$

```
from sklearn.metrics import silhouette_score
silhouette_score(X, labels)
```



Calinski Harabasz Index

Also known as ***Variance Ratio Criterion*** (higher is better).

Ratio of sum of ***between-clusters dispersion*** and ***within-clusters dispersion*** for all clusters.

Define

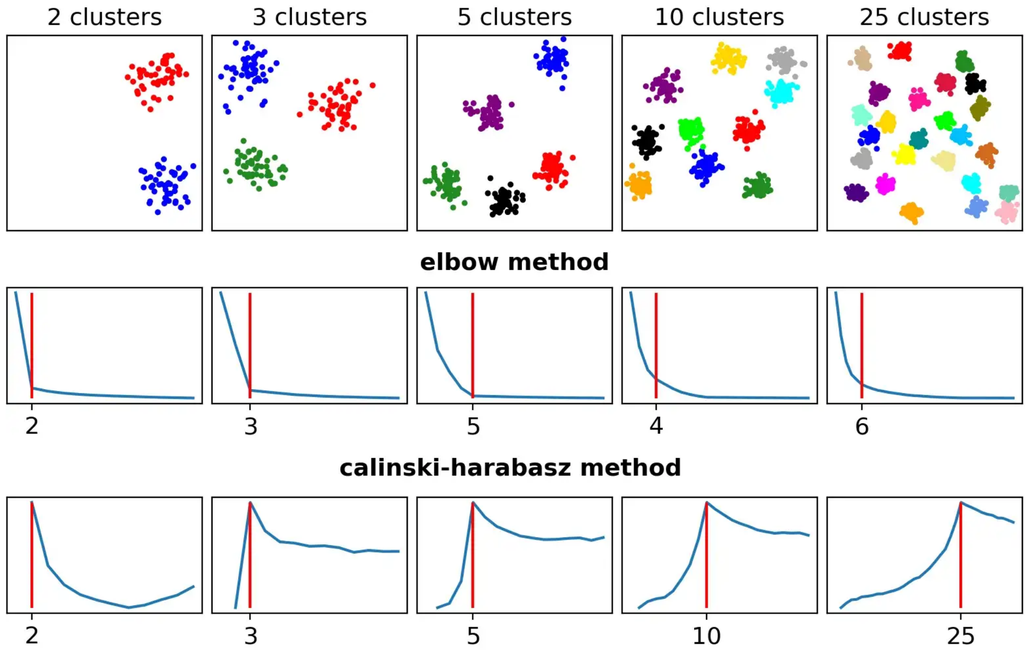
- N data points with global center $\boldsymbol{c}_G$.
- K clusters, each with N_q data points and center $\boldsymbol{c}_q$ in cluster q.

The Calinski Harabasz index is given by

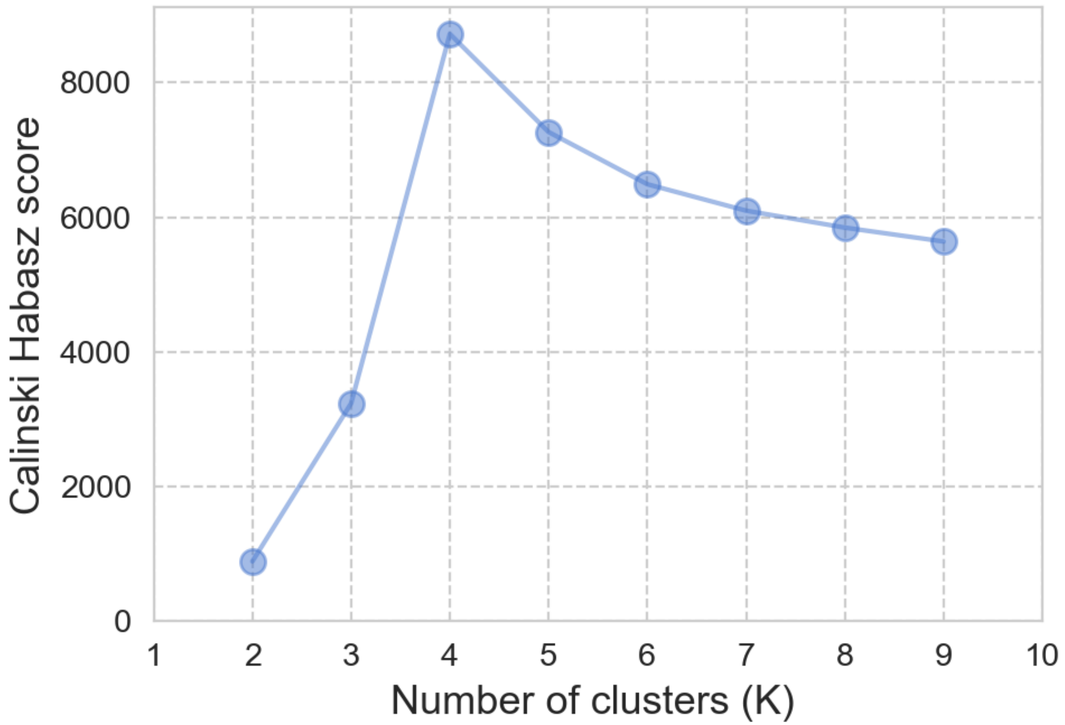
$$S = \frac{B_K/(K-1)}{W_K/(N-K)}$$

$$B_K = \sum_{q=1}^{K} N_q \|\boldsymbol{c}_q - \boldsymbol{c}_G\|^2 \quad \text{Between-cluster dispersion}$$

$$W_K = \sum_{q=1}^{K} \sum_{\boldsymbol{x} \in C_q} \|\boldsymbol{x} - \boldsymbol{c}_q\|^2 \quad \text{Within-cluster dispersion}$$



```
from sklearn.metrics import calinski_harabasz_score
calinski_harabasz_score(X, labels)
```



Davies-Bouldin Index

Signify the average ‘similarity’ between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves.

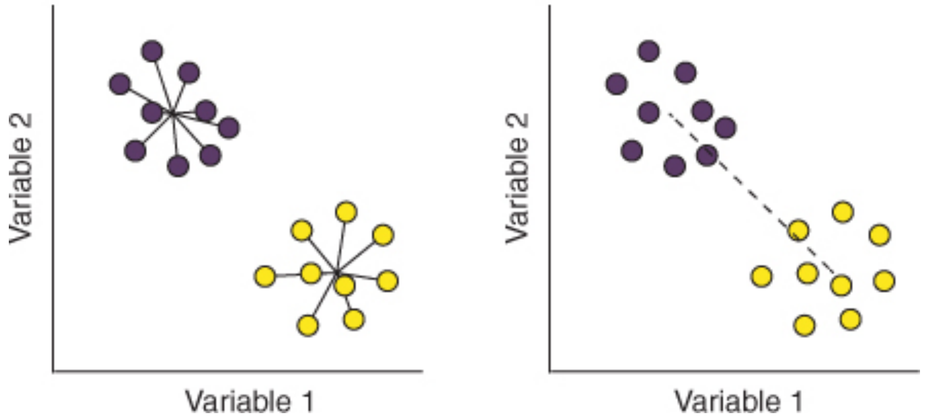
For two clusters i and j, define

- s_i = Cluster i diameter (Avg. distance of points to the centroid)
- d_{ij} = Distance between clusters i and j

The ***similarity*** of two clusters i and j is defined as $R_{ij} = \dfrac{s_i + s_j}{d_{ij}}$

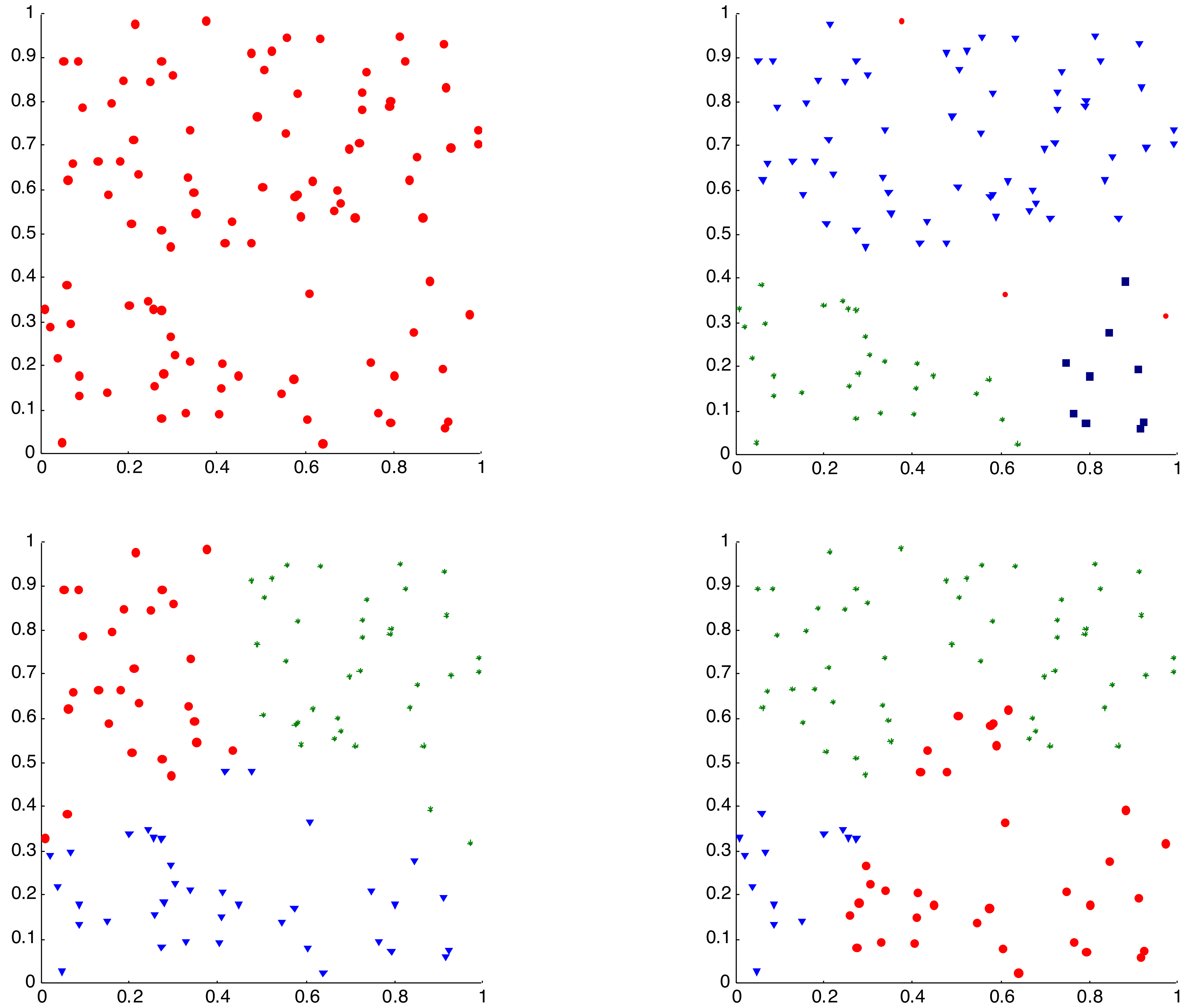
The ***Davies-Bouldin index*** is given by

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j} R_{ij}$$



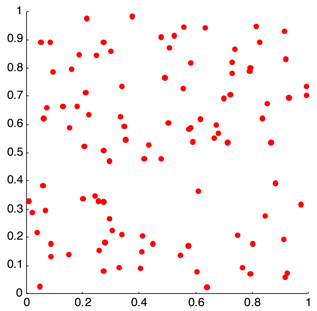
Clusters found in Random Data

Random Points



Hopskins Statistics

Random data

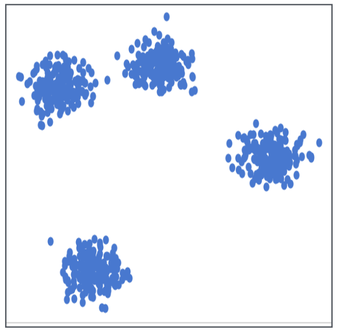


Sample p points

For each point i,

find distance u_i to its nearest neighbor in actual data

Actual data



Sample p points

For each point i,

find distance w_i to its nearest neighbor in actual data

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

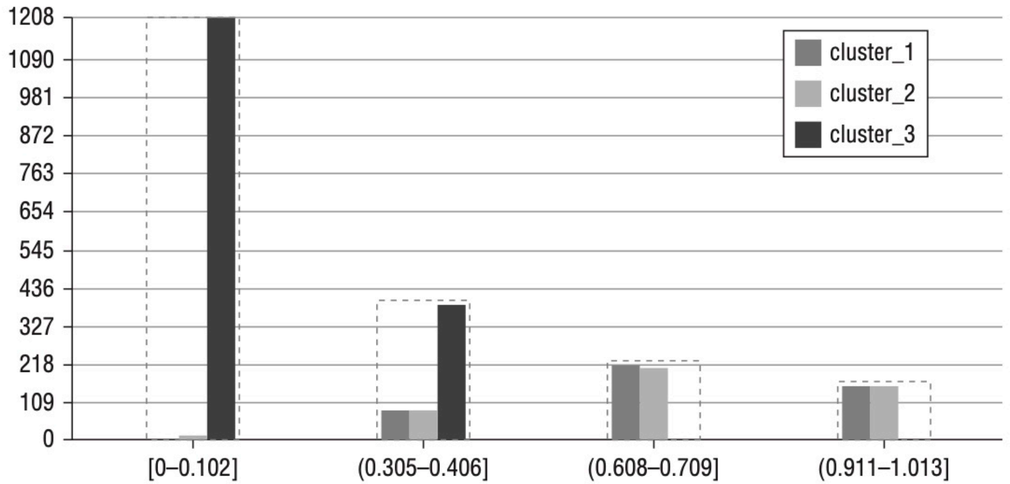
Taking Insights from Clustering

Count of records in each cluster

Summary statistics of (descaled) variables in each cluster (Mean vectors = Centroids)

ANOVA to find which variables have significant difference in means among clusters.

VARIABLE	CLUSTER 1	CLUSTER 2	CLUSTER 3	OVERALL
# Records in Cluster	8,538	8,511	30,656	47,705
LASTDATE	0.319	0.304	0.179	0.226
FISTDATE	0.886	0.885	0.908	0.900
RFA_2F	0.711	0.716	0.074	0.303
D_RFA_2A	0.382	0.390	0.300	0.331
E_RFA_2A	0.499	0.500	0.331	0.391
F_RFA_2A	0.369	0.366	0.568	0.496
DOMAIN3	0.449	0.300	0.368	0.370
DOMAIN2	0.300	0.700	0.489	0.493
DOMAIN1	0.515	0.300	0.427	0.420
NGIFTALL_log10	0.384	0.385	0.233	0.287
LASTGIFT_log10	0.348	0.343	0.430	0.400



Identifying Key Variables in Forming Clusters

May not be feasible to examine graphs/tables of every variable in every cluster.

Assign cluster labels to data and build a decision tree to find feature importance.

CLUSTER LABEL	RULE	NUMBER OF RECORDS MATCHING RULE	NUMBER OF RECORDS WITH CLUSTER LABEL IN RULE	TREE ACCURACY
Cluster 1, Rule 1	DOMAIN2 = 0 and RFA_2F > 2	6,807	6,807	100%
Cluster 1, Rule 2	DOMAIN2 = 0 and RFA_2F = 2 and E_RFA_2A = 1	1,262	1,262	100%
Cluster 2, Rule 1	DOMAIN2 = 1 and RFA_2F > 2	6,751	6,751	100%
Cluster 2, Rule 2	DOMAIN2 = 1 and RFA_2F = 2 and E_RFA_2A = 1	1,304	1,304	100%
Cluster 3, Rule 1	RFA_2F = 1 and E_RFA_2A = 0	21,415	21,415	100%
Cluster 3, Rule 2	RFA_2F = 1 and E_RFA_2A = 1	2,432	2,411	99%
Cluster 3, Rule 3	RFA_2F = 2 and E_RFA_2A = 0 and F_RFA_2A = 1	5,453	5,344	98%

Identifying Cluster Prototypes

Representative record (in natural units) of each cluster (one with closest distance to centroid)

Helpful in gaining additional insight into the meaning of each cluster.

Can examine 2nd or 3rd smallest distance as well.

FIELD	CLUSTER 1, PROTOTYPE 1	CLUSTER 2, PROTOTYPE 1	CLUSTER 3, PROTOTYPE 1
Distance	0.375	0.301	0.313
LASTDATE	9601	9601	9512
FISTDATE	9111	9109	9203
RFA_2F	3	3	1
D_RFA_2A	0	0	0
E_RFA_2A	1	1	0
F_RFA_2A	0	0	1
DOMAIN3	0	0	0
DOMAIN2	0	1	0
DOMAIN1	1	0	0
NGIFTALL	10	10	19
LASTGIFT	12	11	5

Clustering App: Feature Augmentation

$$\boldsymbol{X}_{n\times d} = \begin{bmatrix} & \boldsymbol{X}_1 & \boldsymbol{X}_2 & \cdots & \boldsymbol{X}_d \\ \boldsymbol{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \boldsymbol{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

Clustering into
$K \ll d$

Cluster encoding

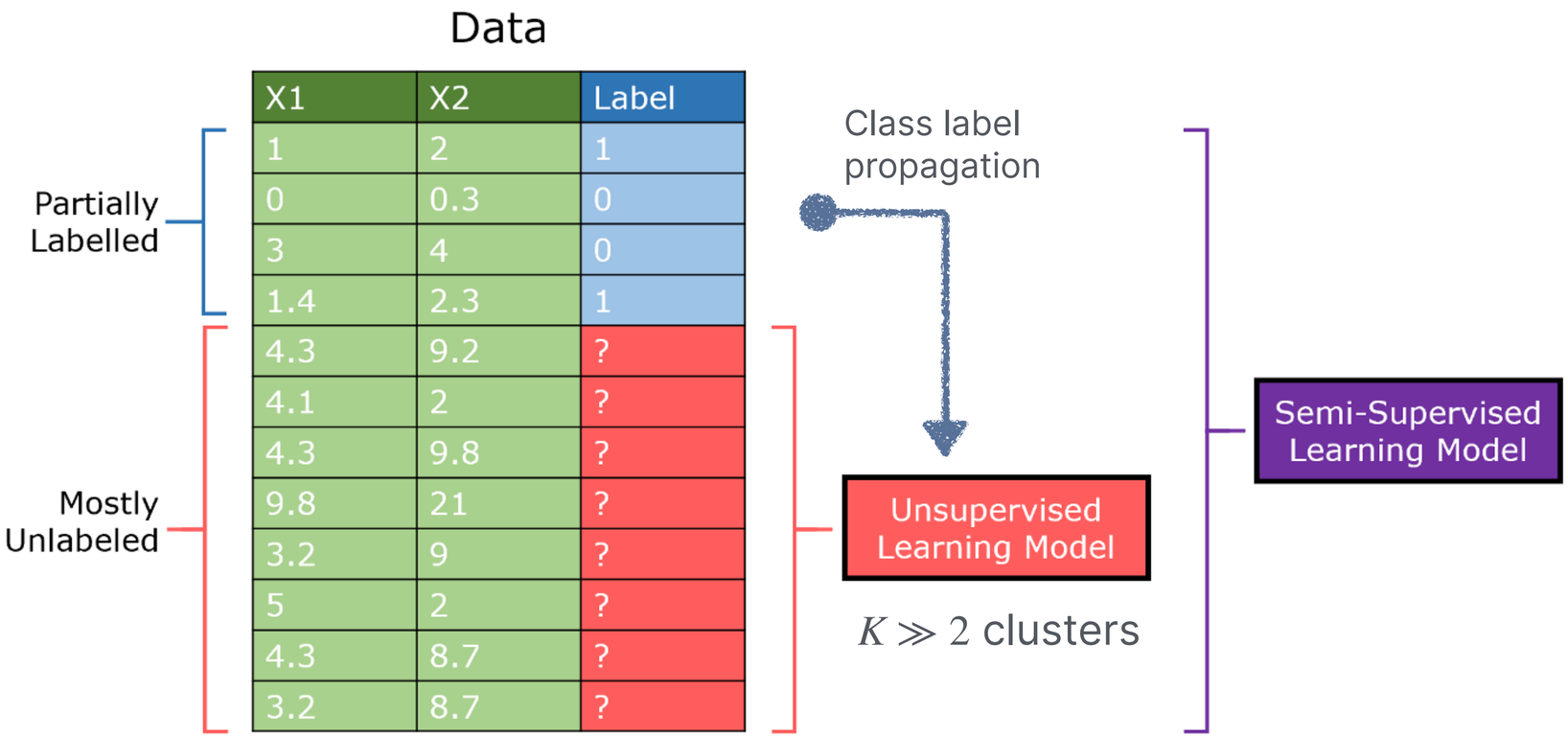
Constructed features

C_1 C_2 $\cdots$ C_K

0	x		0
0	0		0
0	0		x
x	0		0
0	0		0

One-hot encoding
Inverse-distance encoding

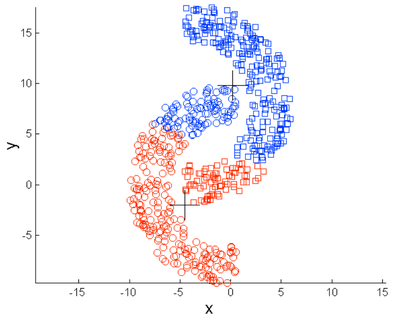
Clustering App: Semi-Supervised Learning



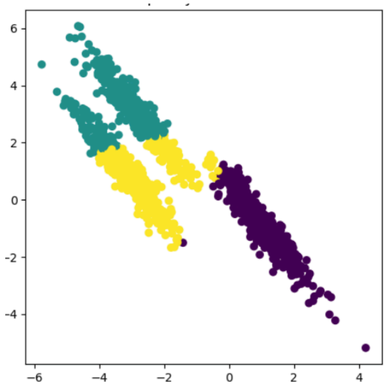
https://towardsdatascience.com/supervised-learning-but-a-lot-better-semi-supervised-learning-a42dff534781

Limitations of K-Means

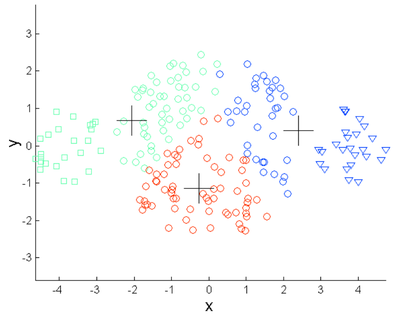
Non-convex shapes



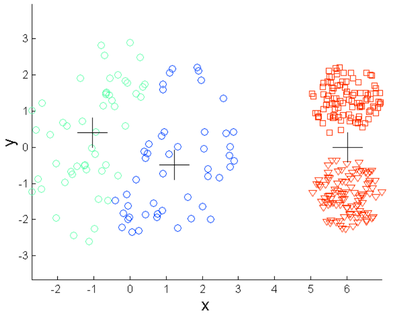
Non-spherical shapes



Different sizes



Different densities
(Unequal variances)



Summary

Clustering to find natural groupings of data points

Require the calculation of dissimilarity among objects.

K-means clustering

- Approximate clusterings that minimize SSE between objects and their centroids.
- Elbow method or Silhouette coefficients to determine the number of clusters

Need to interpret clustering results

Clustering Lab

(Group of 4 to 6)

Part I

- From sheets 'Dataset1', 'Dataset2', 'Dataset3', 'Dataset4' in clustering.xlsx, apply K-mean clustering to determine natural groupings of the data and visualize the clustering results. You may use either Elbow method or Silhouette coefficients to determine appropriate number of clusters.

Part II

The wine dataset contains chemical analysis of wines grown in the same region in Italy. The analysis determined the quantities of 13 constituents including

Load data from 'WINE' sheet in clustering.xlsx

- Alcohol
- Malic acid
- Ash (inorganic matter remaining after evaporation and incineration).
- Alkalinity of ash (Ability to neutralize acids or resist changes that cause acidity)
- Magnesium
- Total phenols (Similar to alcohols but form stronger hydrogen bonds)
- Flavanoids (Compound with antioxidant properties)
- Nonflavanoid phenols
- Proanthocyanins (Compound contributing to wine color and mouthfeel properties)
- Color intensity
- Hue
- OD280/OD315 of diluted wines (Ratio of wavelength absorbance indicating taste, color, and aging characteristics)
- Proline (Amino acid to boost viscosity, sweetness and flavor)

Apply PCA to reduce the dataset dimension that captures at least 80% of original variance. Then, cluster the reduced-dimension data.