# Plan of Attack

21 November 2023 09:37

Unsupervised learning

↓

Clustering

↓

Kmean

- intuition
- python
- assumption
- limitations
- Kmeans variations

Assignment

# Types of ML Learning

21 November 2023 09:38

90% → 5-10

DSMP 1

Types

→ dog

→ cat

Supervised

Unsupervised

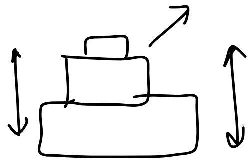
Reinforcement →

Regress Class

input labelx

unlabelled

99%



input/output

labels

Yann Le Cunn

Fb → CNN

5000

1000

100 milk

80 egg

unsupervised

deep learning

Language model

Anamoly detection

Clustering

Association

Dimentionality reduction

credit card fraud

Isolation forest

unlabelled

Kmeans

DBSCAN

GMM

Aprioi

Eclat

PCA

tsne

1000 cols → dim

input → cols → feature

Visualis → 10 cols → 10 dim

2dim

3 dim

Unsupervised learnin → clustering → Kmeans

LLM → Large language models

LLM → Large language models

↓

language models

↓

nlp model → text data

↓

next word prediction

wikipedia

Chap

unsupervised

↓

labels

# Applications of Clustering

21 November 2023 09:39

1. Customer Segmentation
2. Data Analysis
3. Semi Supervised Learning
4. Image Segmentation

google photos

Kmeans
Heir
~~DBSCAN~~
GMM

# K Means Geometric Intuition

21 November 2023 09:38

100dim

k=2

stop → 2d | 3d | nd

(euclidean)

cgpa

Kmean++

## Problem Statement

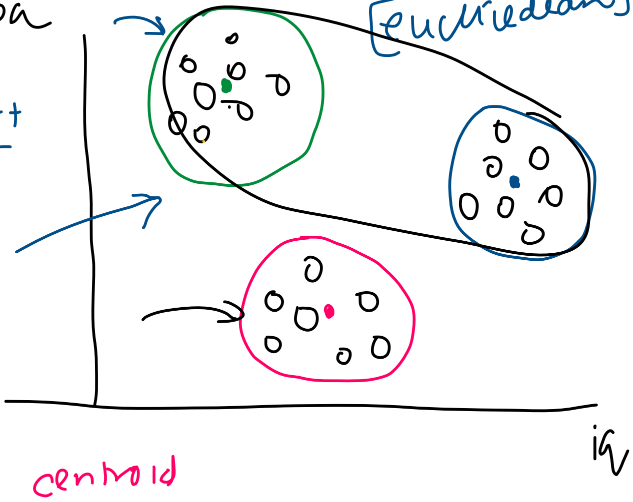
1. Decide k clusters → K=3
2. Initialize centroids
3. Start Iterating loop →

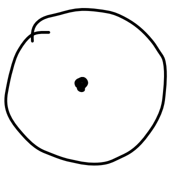
[Assign clusters] ⇄

→ Move centroid

[Check and stop] → previous position centroid = current position centroid → stop

continue



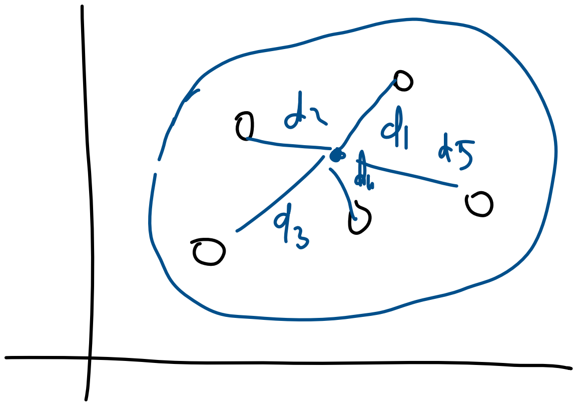



## [Elbow Method] → inertia → WCSS

21 November 2023 13:34

↳ within cluster sum of squared distance

How do we decide the correct value of K



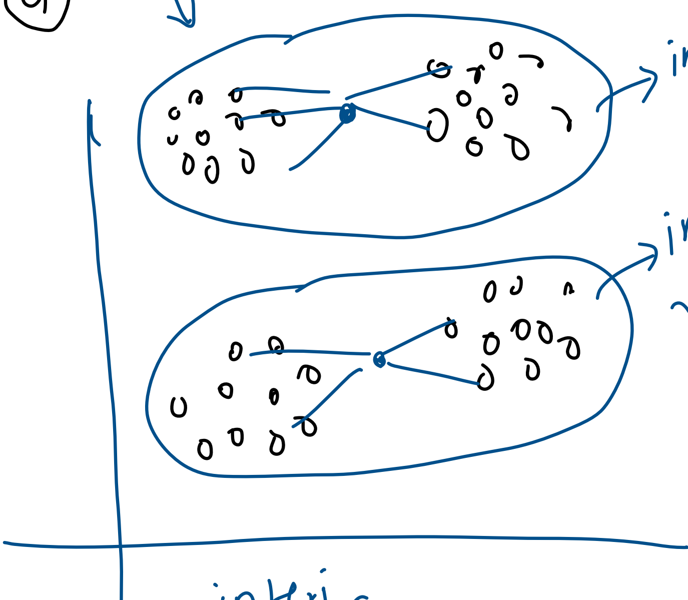

100

$$WCSS = \underbrace{d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2}_{n}$$

4

100 points

elbow method




1

inertia(1) = X

K=2

$$inertia(2) = \underbrace{int(a)}_{X_1} + \underbrace{int(b)}_{Y_2} \rightarrow Y$$

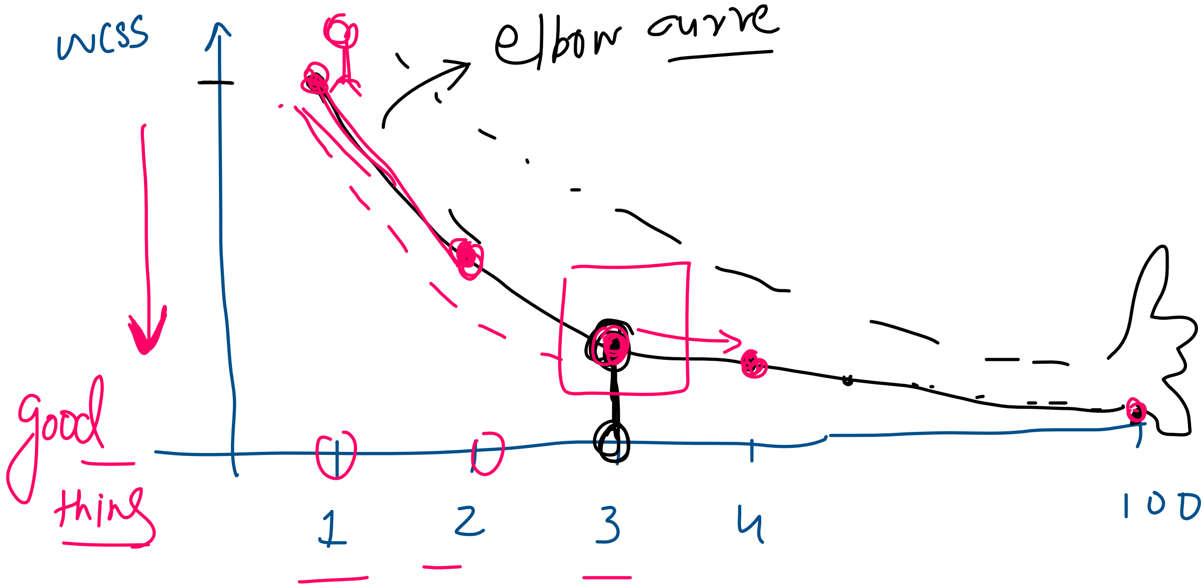
1) X > Y

2) Y > X

intertra(100)

1 > 2 > 3 > 4 > 5 . . . > 100

1 > 2 > 3 > 4 > 5 . . . > 100




[[WCSS] → variance]

1. Why square of distance
2. Why the word Inertia?

# Limitations of Elbow Method

21 November 2023 17:35

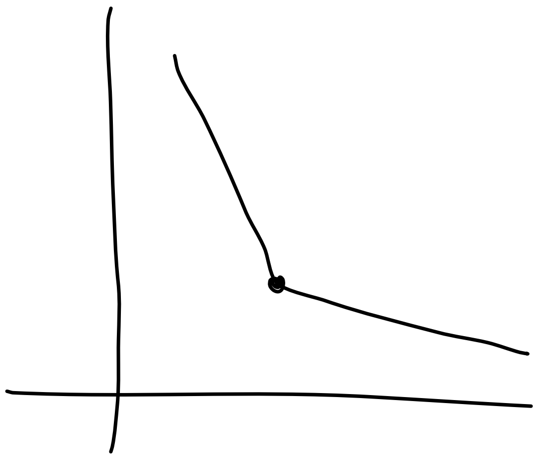
Subjectivity in Identifying the Elbow: The biggest challenge with the elbow method is the subjective nature of identifying the "elbow" point. The point where the inertia starts decreasing at a slower rate can be open to interpretation and may not be clear-cut, especially in datasets where the decrease in inertia is gradual.

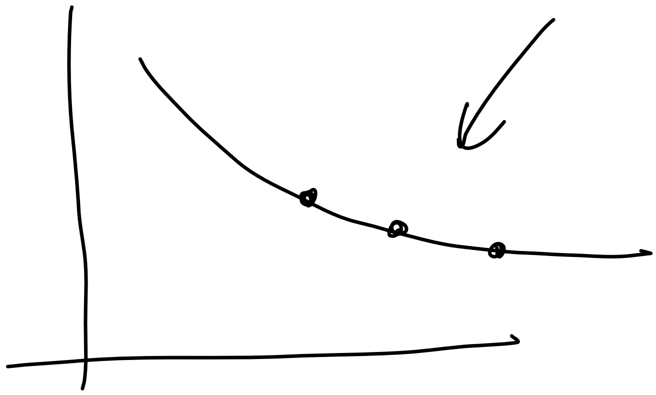
Not Suitable for All Datasets: The method does not work well if the data is not very clustered or if the clusters have an irregular shape. In such cases, the elbow might not be distinct, leading to ambiguity in choosing the right number of clusters.

Performance with Large Number of Features: The elbow method can become less effective as the number of features in the dataset increases. High-dimensional data can make the identification of a clear elbow more difficult.

Doesn't Consider Cluster Quality: The elbow method focuses solely on the variance within the clusters and does not take into account the quality of the clusters formed. It's possible to choose a k where clusters are not meaningful or well-separated.

Sensitivity to Scaling: Like K-means clustering itself, the results of the elbow method can be sensitive to the scale of the data. Features with larger scales can dominate the result, potentially leading to suboptimal choices of k.





$f_1$ (1000) $\longleftrightarrow$ age

$\rightarrow$ salary

7500

# Code Example

21 November 2023 09:38

# Assumptions of KMeans

21 November 2023 18:01

Spherical Cluster Shape: K-means assumes that the clusters are spherical and isotropic, meaning they are uniform in all directions. Consequently, the algorithm works best when the actual clusters in the data are circular (in 2D) or spherical (in higher dimensions).

Similar Cluster Size: The algorithm tends to perform better when all clusters are of approximately the same size. If one cluster is much larger than others, K-means might struggle to correctly assign the points to the appropriate cluster.

Equal Variance of Clusters: K-means assumes that all clusters have similar variance. The algorithm uses the Euclidean distance metric, which can bias the clustering towards clusters with lower variance.

Clusters are Well Separated: The algorithm works best when the clusters are well separated from each other. If clusters are overlapping or intertwined, K-means might not be able to distinguish them effectively.

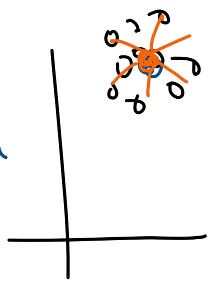
Number of Clusters (k) is Predefined: K-means requires the number of clusters (k) to be specified in advance. Choosing the right value of k is crucial, but it is not always straightforward and typically requires domain knowledge or additional methods like the Elbow method or Silhouette analysis.

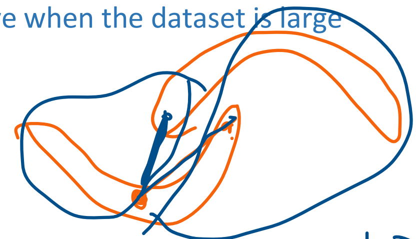
Large n, Small k: K-means is generally more efficient and effective when the dataset is large (large n) and the number of clusters is small (small k).

#rows / cluster

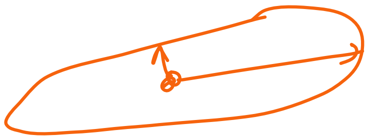
10K / less

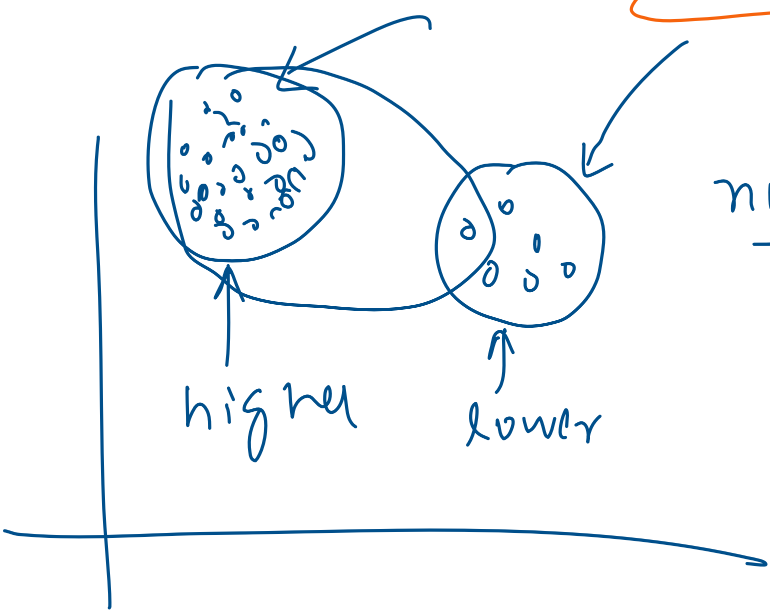
3 cluster



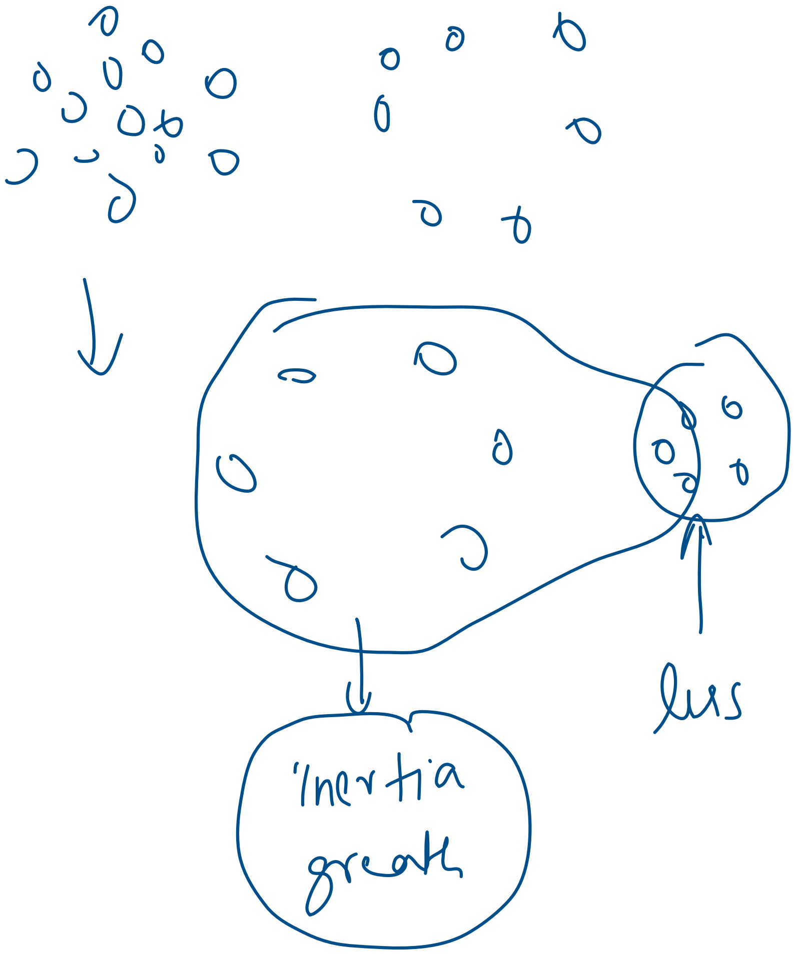


100 data / 50 clus






not ideal

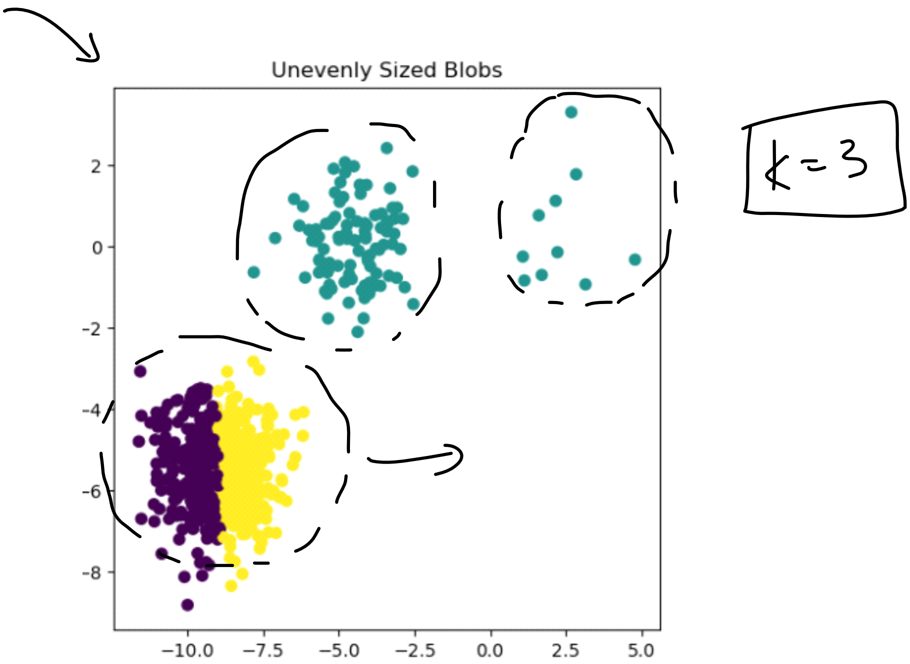

less
'Inertia greater

# Limitations of K Means

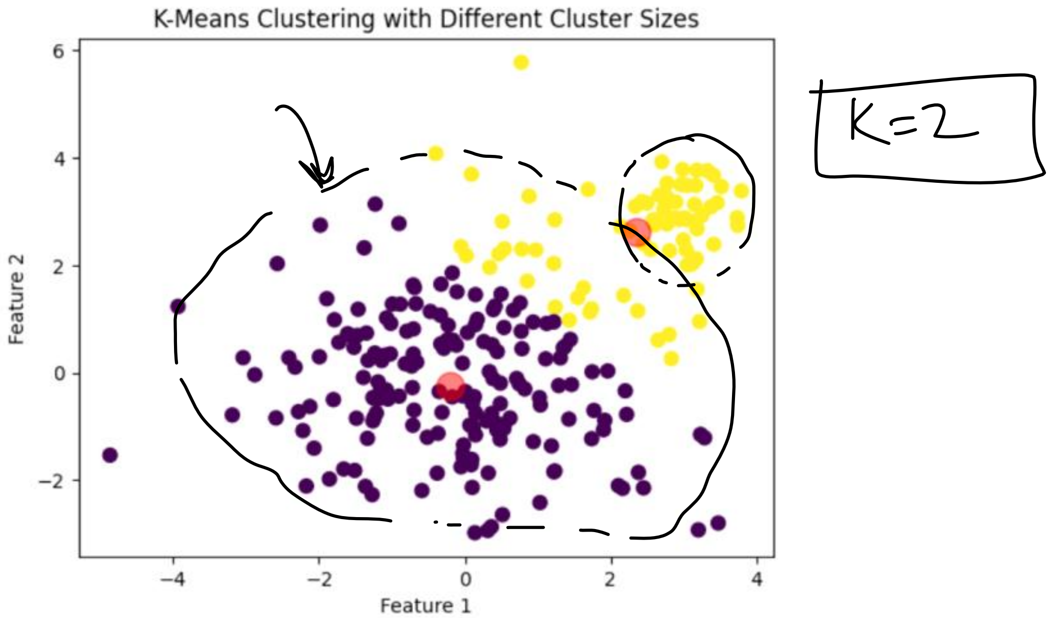
21 November 2023 09:39

- ✓ Number of Clusters: Determining the optimal number of clusters (k) is not straightforward and often requires domain knowledge or methods like the elbow method.

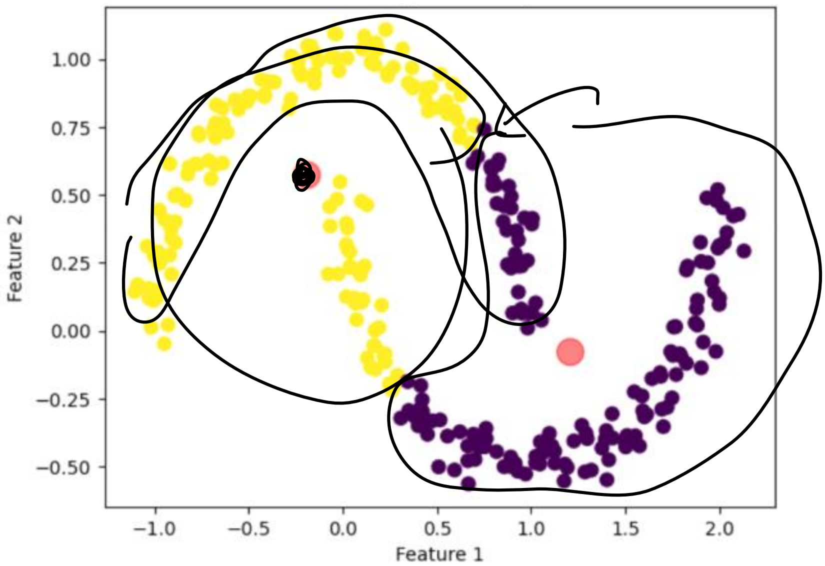
- ✓ Requires clusters of similar sizes: Kmeans requires the clusters to be of similar sizes.




Similar variance between clusters: Kmeans requires the clusters to be of similar variance.



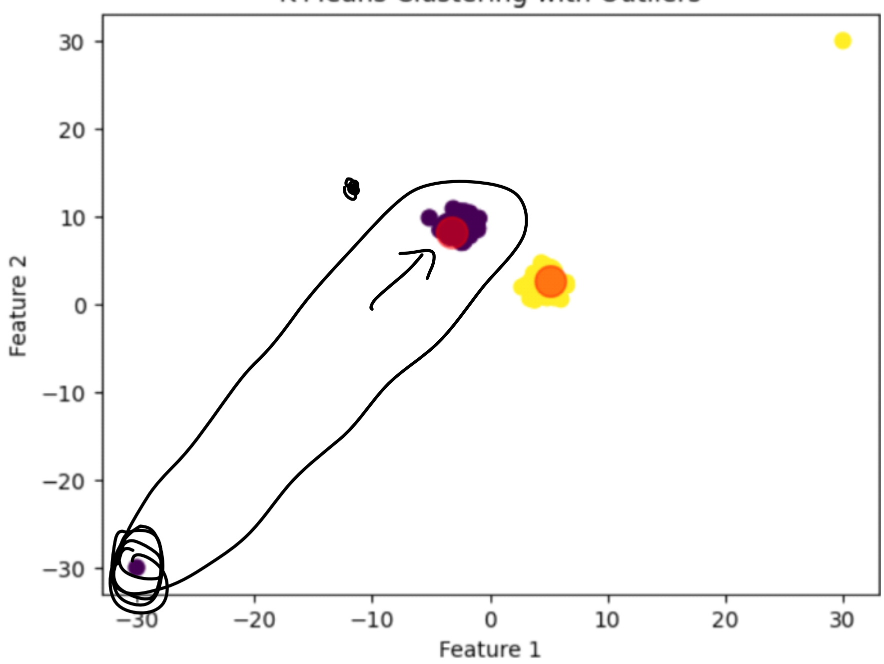

- ✓ Assumption of Spherical Clusters: KMeans assumes that clusters are spherical and of similar size, which might not be the case in real-world data.




K=2

- ✓ Vulnerability to Outliers: Outliers can significantly distort the mean value of a cluster, leading to misleading results.

K-Means Clustering with Outliers




K=2

soft clustering

→ [0.78, 0.62]

- ✓ Hard Clustering: Each data point is forced into exactly one cluster, which may not be suitable for all applications, especially where data can belong to multiple clusters.
- ✓ High-Dimensional Challenges: In very high-dimensional spaces, the distance between data points can become less meaningful, affecting the performance of KMeans.
- ✓ Sensitive to Scale: The measure is sensitive to the scale of the features. Hence, feature scaling (like standardization) is often recommended before applying K-means.

↳ KMeans

↳ euclidean

→ euclidean

→ high dim

# Variations of KMeans

21 November 2023 13:57

KMeans++: Improves the initialization phase of KMeans by spreading out the initial centroids, which can lead to better and more consistent results.

Mini-Batch KMeans: Uses small random batches of data for each iteration rather than the full dataset. This approach significantly speeds up computation, especially for large datasets, while achieving similar results to the standard KMeans.

K-Medoids: Instead of using the mean of a cluster's points, K-Medoids uses the most centrally located data point (medoid) of a cluster. This makes it more robust to outliers compared to KMeans.

Fuzzy C-Means : Allows each data point to belong to multiple clusters with varying degrees of membership, rather than assigning each point to only one cluster. This approach is useful in scenarios where data points naturally belong to more than one group.

Incremental KMeans - also known as online KMeans or streaming KMeans, is a variant of the KMeans algorithm designed to handle streaming data

# Homework

21 November 2023 09:40