Mikhail Hushchyn

Clustering #1

Clustering. K-Means. Quality Metrics.

2021

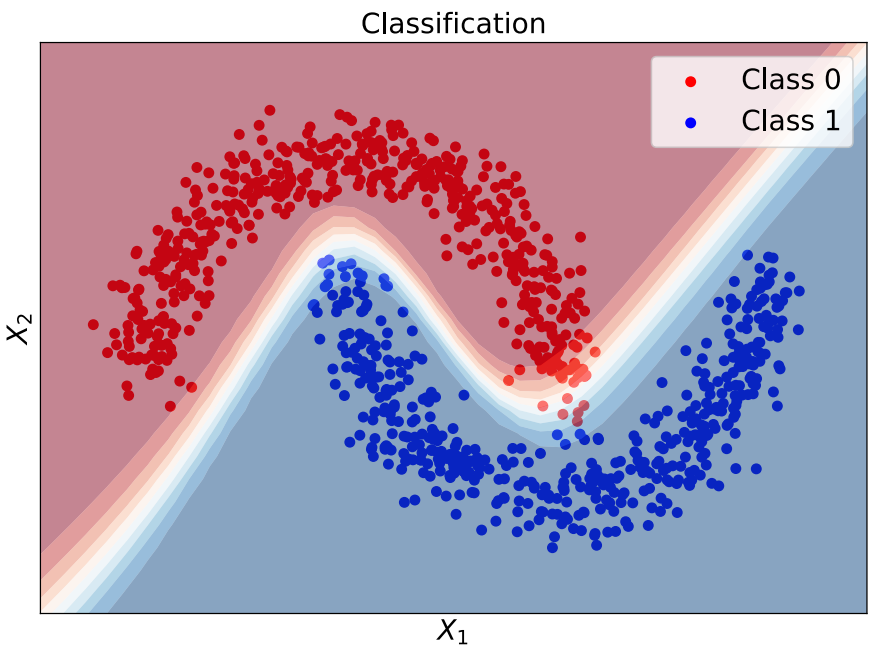
Outline

- Clustering
- K-Means algorithm
- Quality metrics

Clustering

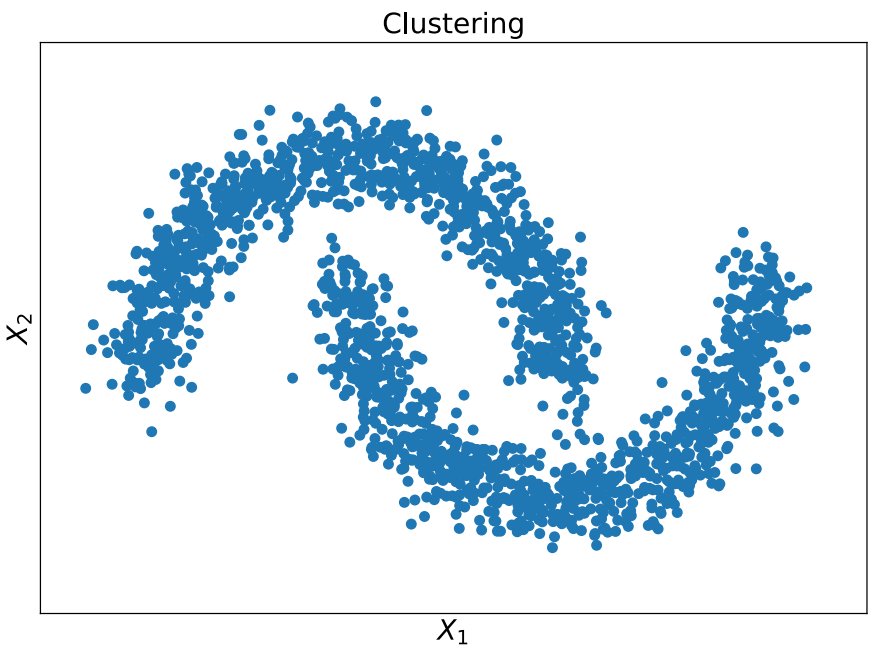
Clustering vs classification

- In classification, we have object features X and class labels $y \in \{0, 1\}$
- A classifier learns decision rule f, so that $f(X) \approx y$
- The trained classifier predicts class labels for new objects

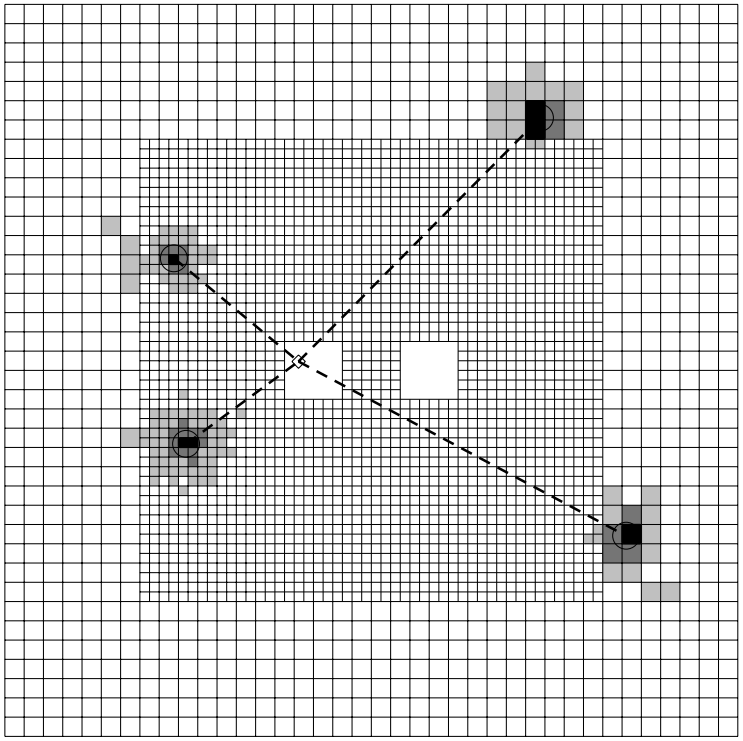


Clustering vs classification

- In clustering, we don't have class labels y
- The goal is to divide all objects into separate groups using only object features X
- Objects inside groups are similar
- Objects from different groups are dissimilar



Example of clustering



Clusters in EM calorimeter of KTEV experiment for $K \to \pi^0\pi^0$ decay.

Clustering assumptions

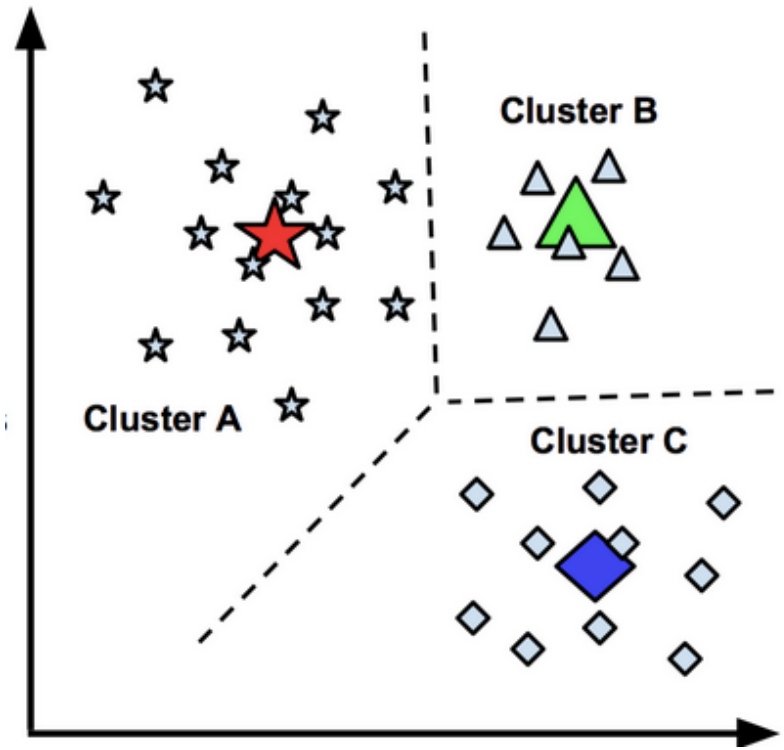
Most of clustering algorithms are based on the following assumptions:

- Objects form dense clusters
- Objects from one cluster are similar
- Objects from different clusters are dissimilar
- Objects similarity is often based on distance between them
- Distances between neighbors within one cluster are smaller than between objects from different clusters

K-Means

Clustering intuition

- Each cluster is represented by its center
- All objects are assigned to the closest center
- The goal is to find such centers that form the most compact clusters



Link: https://medium.com/@msdasila90/basics-k-means-clustering-algorithm-a77c539c9e00

Notations

- Consider a sample with N objects $\{x_n\}_{n=1}^{N}$.
- We will search for K clusters with centers $\{\mu_1, \mu_2, \dots, \mu_K\}$.
- Criterion to find the best centers is minimum of **within-cluster distance**:

$$Q = \sum_{n=1}^{N} \min_k \rho(x_n, \mu_k) \to \min_{\mu_1, \dots, \mu_K}$$

- Each object x_n is assigned to a cluster $z_n \in \{1, 2, \dots, K\}$ as:

$$z_n = \arg\min_k \rho(x_n, \mu_k)$$

General algorithm

```
initialize μ_1, ...μ_K from
random training objects

WHILE not converged:
    FOR n = 1, 2, ...N:
        z_n = arg min_k ρ(x_n, μ_k)

    FOR k = 1, 2, ...K:
        μ_k = arg min_μ Σ_{n:z_n=k} ρ(x_n, μ)

RETURN z_1, ...z_N
```

Assign each object to the nearest center

Update the centers

Algorithm variations

- Distance $\rho(x_n, \mu_k)$ can be defined in different ways.
- If $\rho(x_n, \mu_k) = \|x_n - \mu_k\|_2^2$, we get **K-Means algorithm**
- If $\rho(x_n, \mu_k) = \|x_n - \mu_k\|_1$, we get **K-Medians algorithm**

K-Means algorithm

Initialize μ_j, $j = 1, 2, \ldots K$.

WHILE not converged:

FOR $i = 1, 2, \ldots N$:

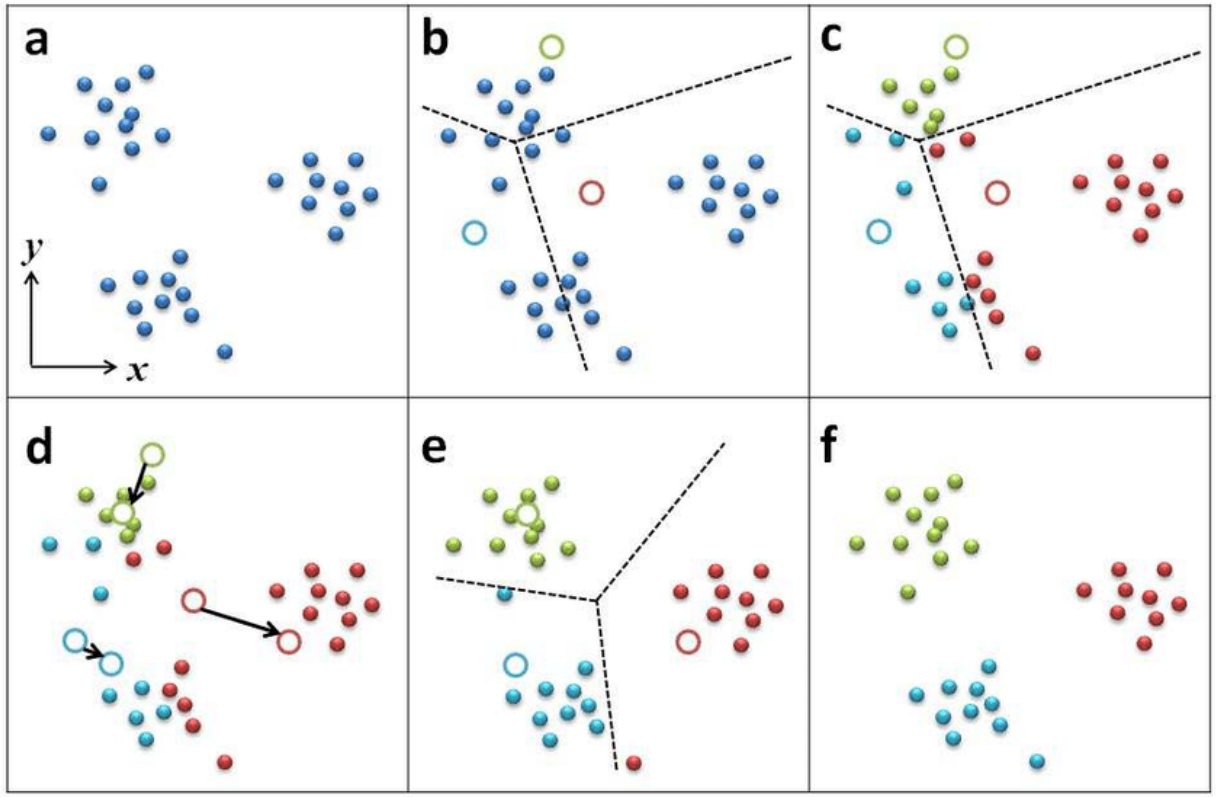
find cluster number of x_i:

$z_i = \arg\min_{j \in \{1,2,\ldots K\}} ||x_i - \mu_j||_2^2$

FOR $j = 1, 2, \ldots K$:

$\mu_j = \frac{1}{\sum_{n=1}^{N} \mathbb{I}[z_n = j]} \sum_{n=1}^{N} \mathbb{I}[z_n = j] x_i$

K-Means demonstration



Properties #1

- Initializaion:
  - Centers $\{\mu_k\}_{k=1}^{K}$ are usually initialized randomly from training objects
  - Number of clusters (and centers) K is fixed
- Convergence criteria:
  - Iterations limit is reached
  - Centers stop changing significantly
  - Cluster assignments $\{z_n\}_{n=1}^{N}$ stop changing

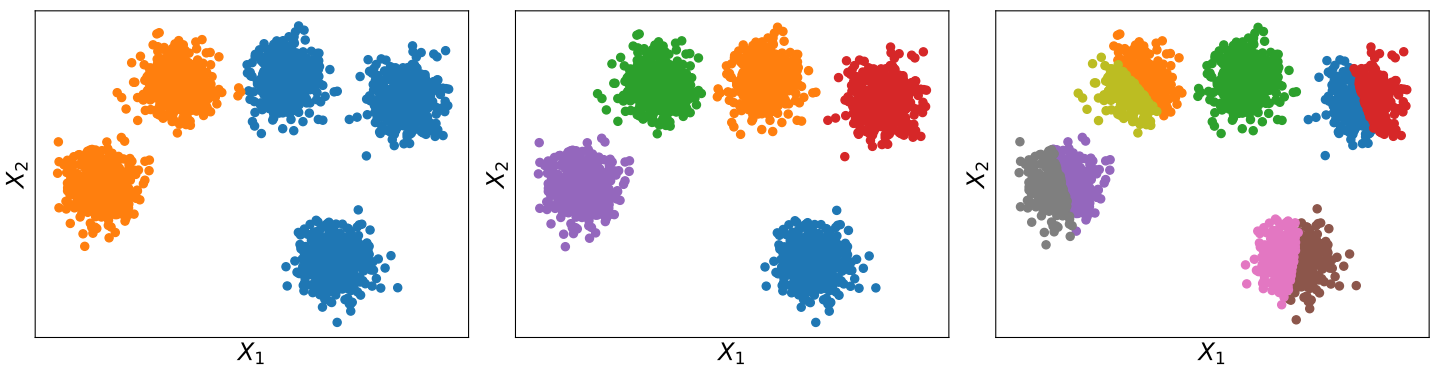
Properties #2

- Solution
  - Depends on starting positions of centers
  - Sensitive to outliers, may create single-object clusters
  - It is recommended to run the algorithm with several different initializations and select solution with the minimal within-cluster distance Q

Elbow method

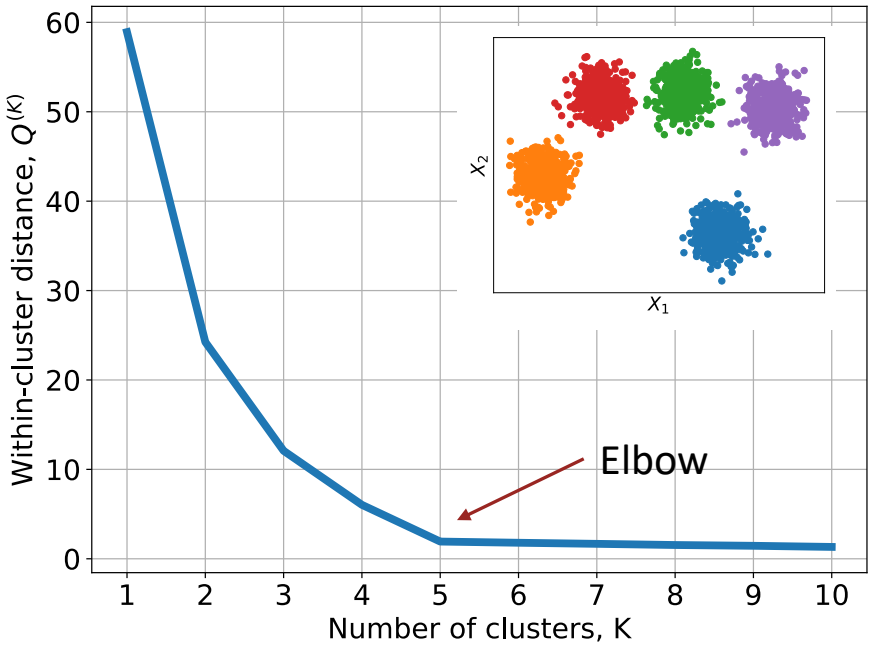
- How to estimate optimal number of clusters K?
- Consider within-cluster distances $Q^{(K)}$ for all possible K:

$$Q^{(K)} = \sum_{n=1}^{N} \left\| x_n - \mu_{z_n} \right\|_2^2 \to \min_{z_1,\ldots,z_N,\mu_1,\ldots,\mu_K}$$



Elbow method

- $Q^{(K)}$ decreases with increasing K
- The dependence has elbow at the optimal number of clusters ($K = 5$)
- Let's try to formalize it

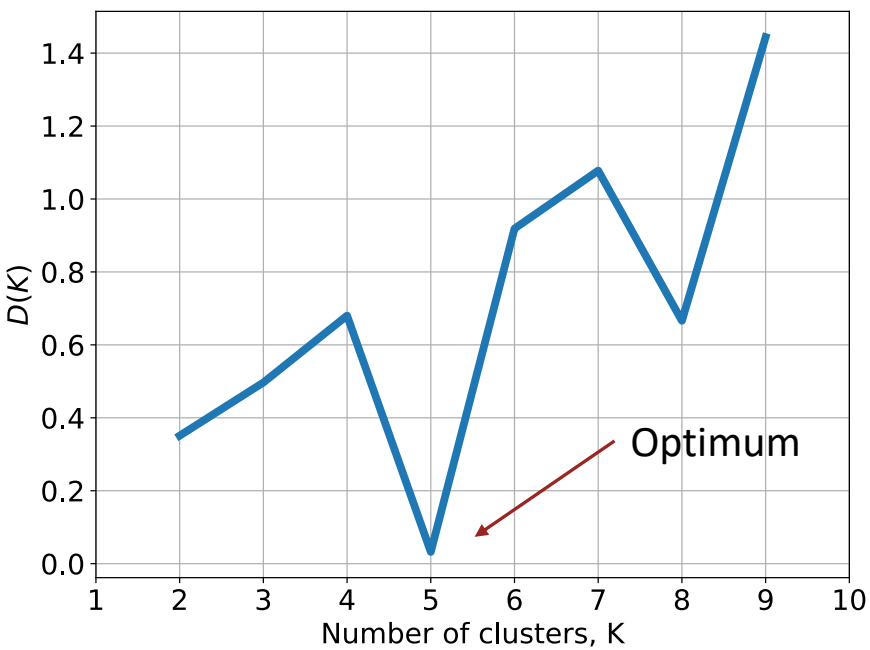


Elbow method

- Let's define $D(K)$:

$$D(K) = \frac{|Q^{(K+1)} - Q^{(K)}|}{|Q^{(K)} - Q^{(K-1)}|}$$

- This function takes small value for the optimal number of clusters



Quality Metrics

Quality metrics

There are two kinds of quality metrics for clustering:

- Supervised
  - Based on ground truth of object labels
  - Invariant to cluster naming
- Unsupervised
  - Based on intuition about “good” clusters:
    - Objects from the same cluster are similar / close to each other
    - Objects from different clusters are dissimilar / distant from each other

Rand Index

Rand Index (RI) is supervised quality metric defined as:

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

TP – number of pairs in the same cluster in predictions and the ground truth,

TN – number of pairs from different clusters in predictions and the ground truth,

FP – number of pairs in the same cluster in predictions, but from different clusters in the ground truth,

FN – number of pairs in the same cluster in the ground truth, but from the different clusters in predictions.

Adjusted Rand Index

Adjusted Rand Index (ARI) is modification of RI:

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{RI}_{\mathrm{Expected}}}{\mathrm{RI}_{\mathrm{Max}} - \mathrm{RI}_{\mathrm{Expected}}}$$

ARI has a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clustering is ideal

Metrics for classification

- $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
- $\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
- $\text{F1} - \text{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$
- Fowlkes-Mallows Index (FMI) = $\frac{\text{TP}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})}}$
- others

Silhouette

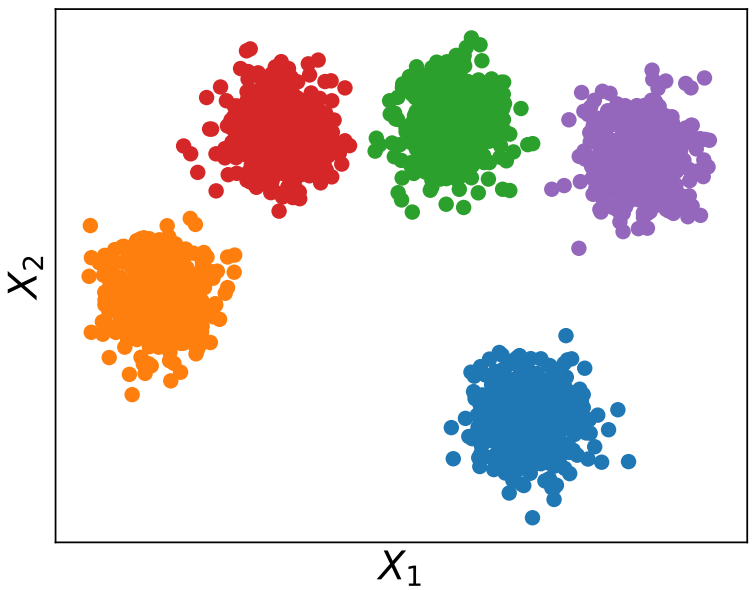
Silhouette is unsupervised quality metric defined as:

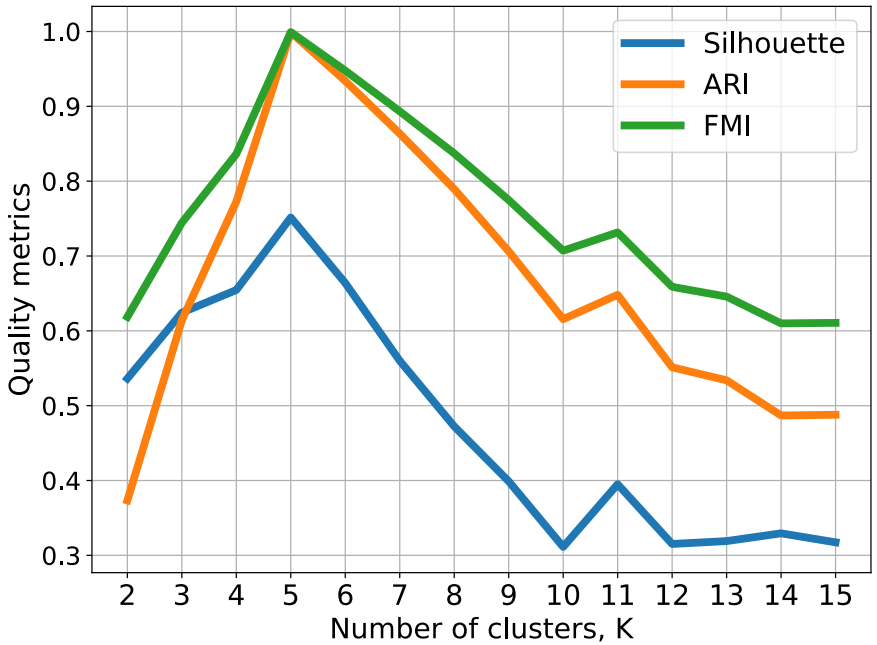
$$\text{Silhouette} = \frac{1}{N}\sum_{i=1}^{N} \frac{d_i - s_i}{\max\{d_i, s_i\}}$$

s_i - mean distance between the i-th object and all objects in the same cluster,

d_i - mean distance between the i-th object and all objects in the nearest cluster.

Example


X_2
X_1


Silhouette
ARI
FMI
Quality metrics
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
2 3 4 5 6 7 8 9 10 11 12 13 14 15
Number of clusters, K

Summary

Summary

- Clustering
  - Clustering is a filed of unsupervised machine learning
  - Its goal is to divide objects into groups based on their similarities
- K-Means algorithm
  - Clusters are represented by their centers
  - The centers are optimized to minimize within-cluster distance
- Quality metrics
  - Supervised metrics uses ground truth (ARI, FMI)
  - Unsupervised metrics are based on intuition of ”good” clusters (Silhouette)