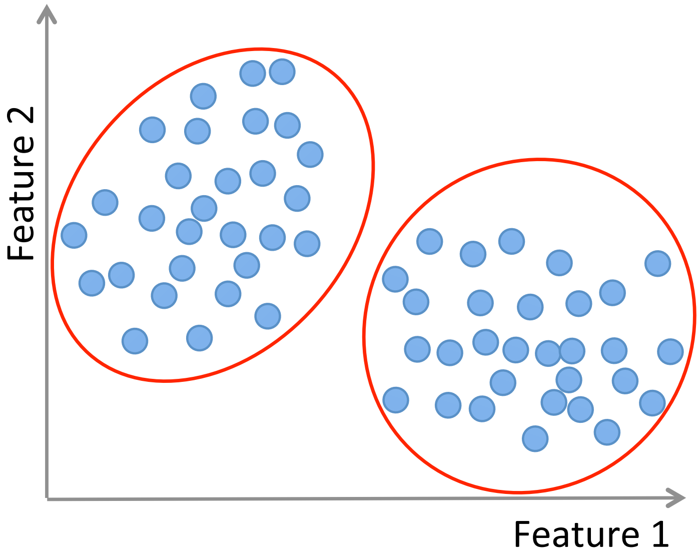
Artificial Intelligence

Machine Learning
Unsupervised learning



Unsupervised Learning

Training data: "examples" x.

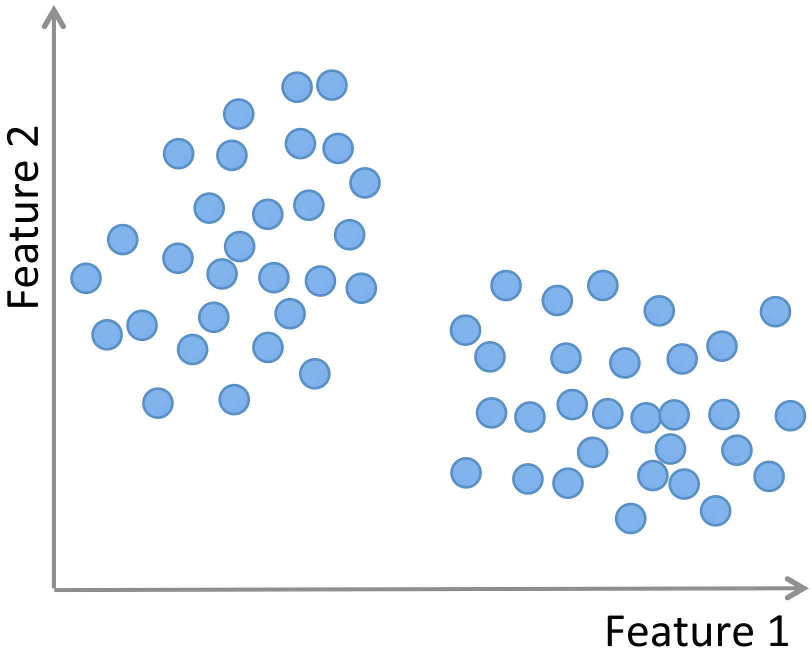
$$x_1, \ldots, x_n, \; x_i \in X \subset \mathbb{R}^n$$

- **Clustering/segmentation:**

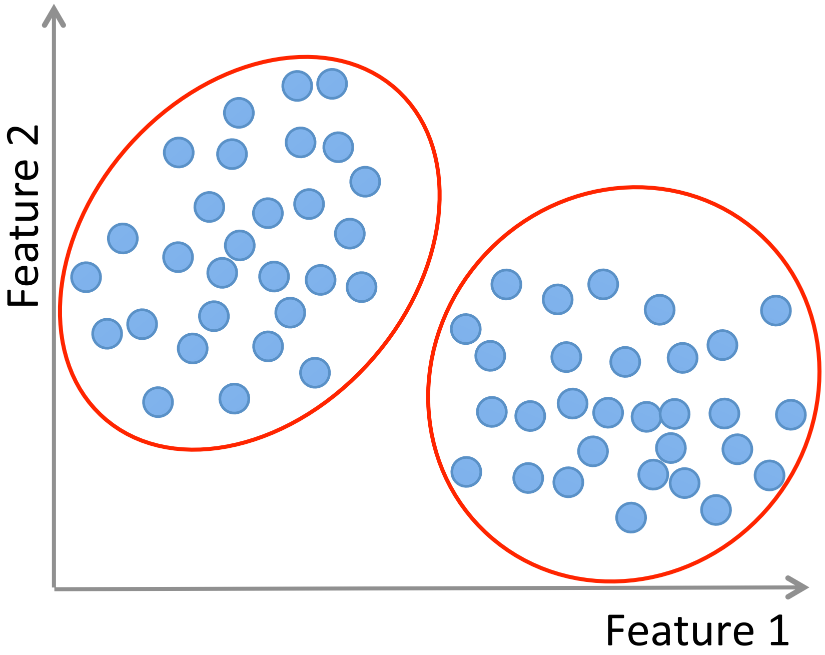
  $f : \mathbb{R}^d \longrightarrow \{C_1, \ldots C_k\}$ (set of clusters).

  Example: Find clusters in the population, fruits, species.

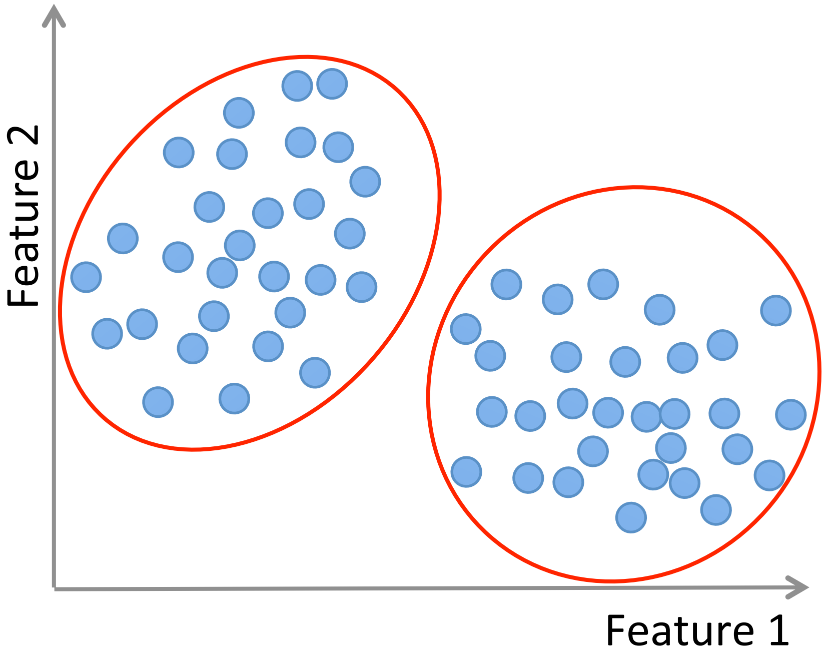
Unsupervised learning



Unsupervised learning



Unsupervised learning

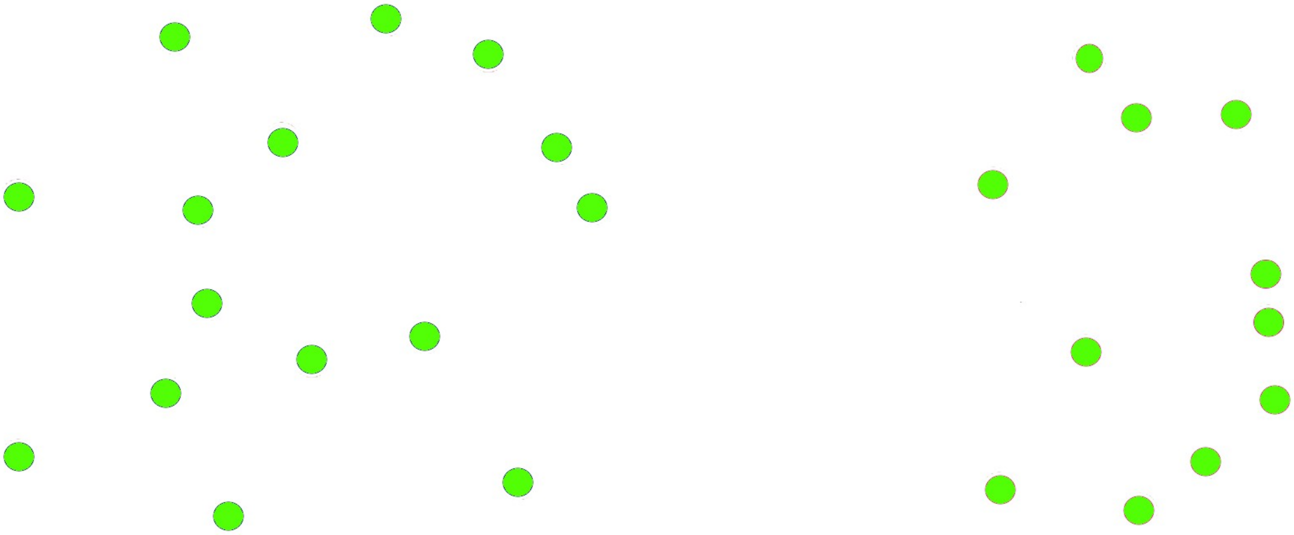


Methods: K-means, gaussian mixtures, hierarchical agglomerative clustering, spectral clustering, DBScan, etc.

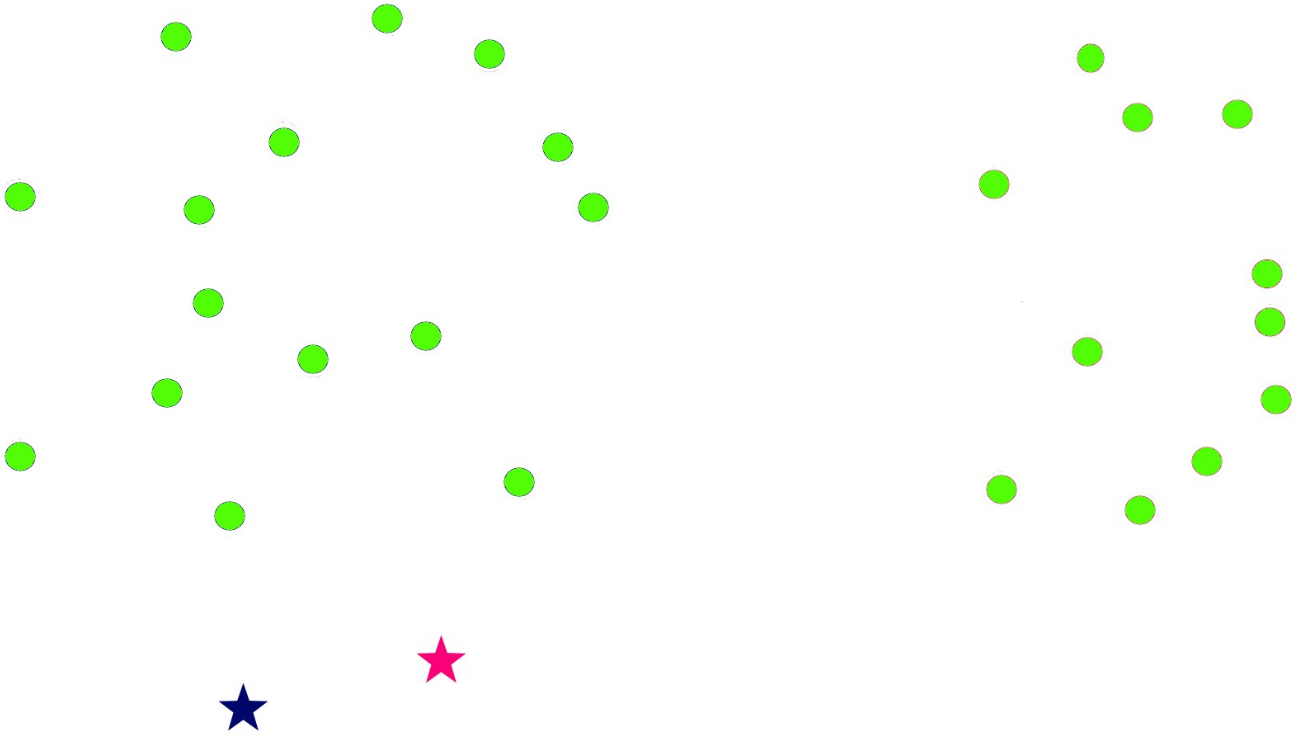
Clustering examples

- Clustering of the population by their demographics.
- Clustering of geographic objects (mineral deposits, houses, etc.)
- Clustering of stars
- Audio signal separation. Example?
- Image segmentation. Example?

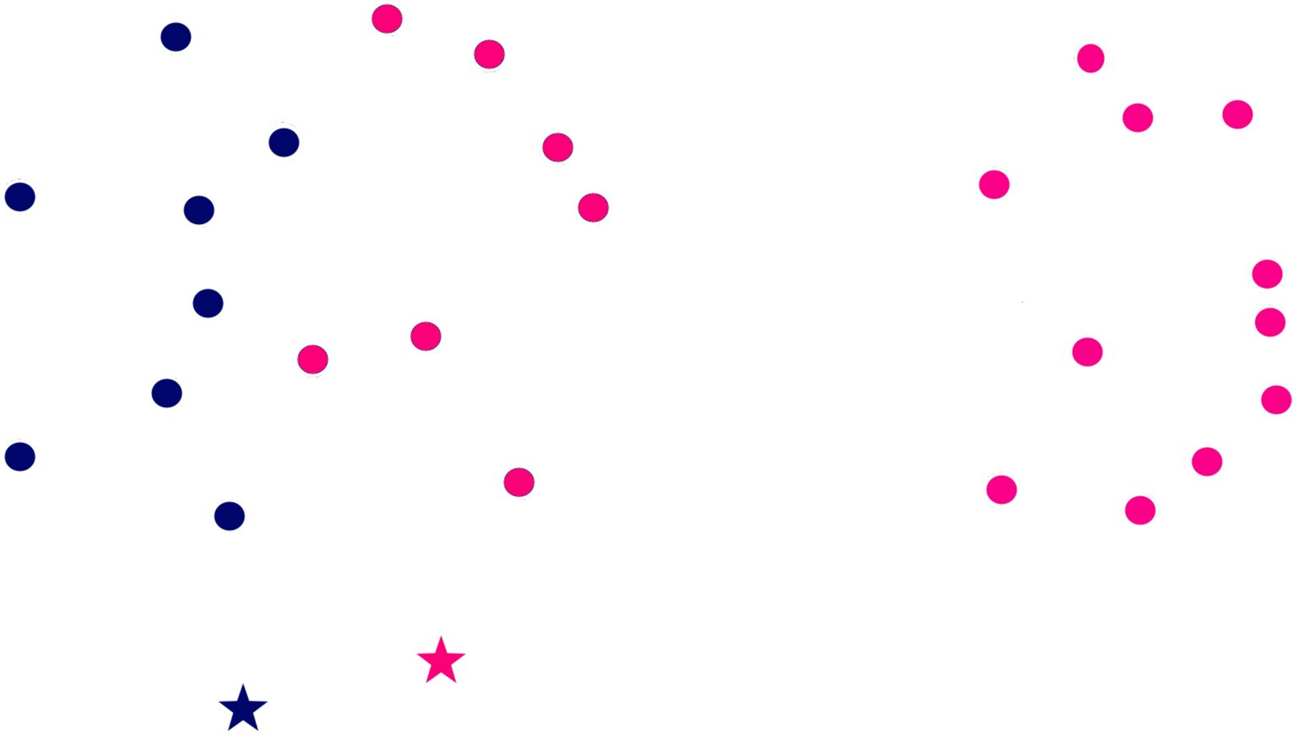
K-Means: example



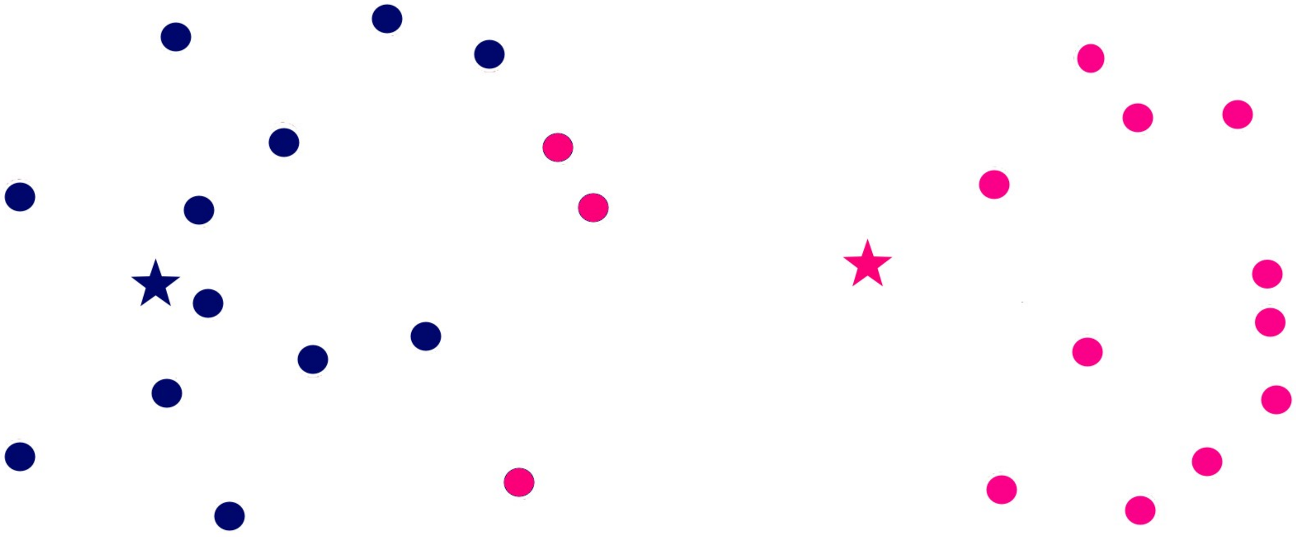
K-Means: example



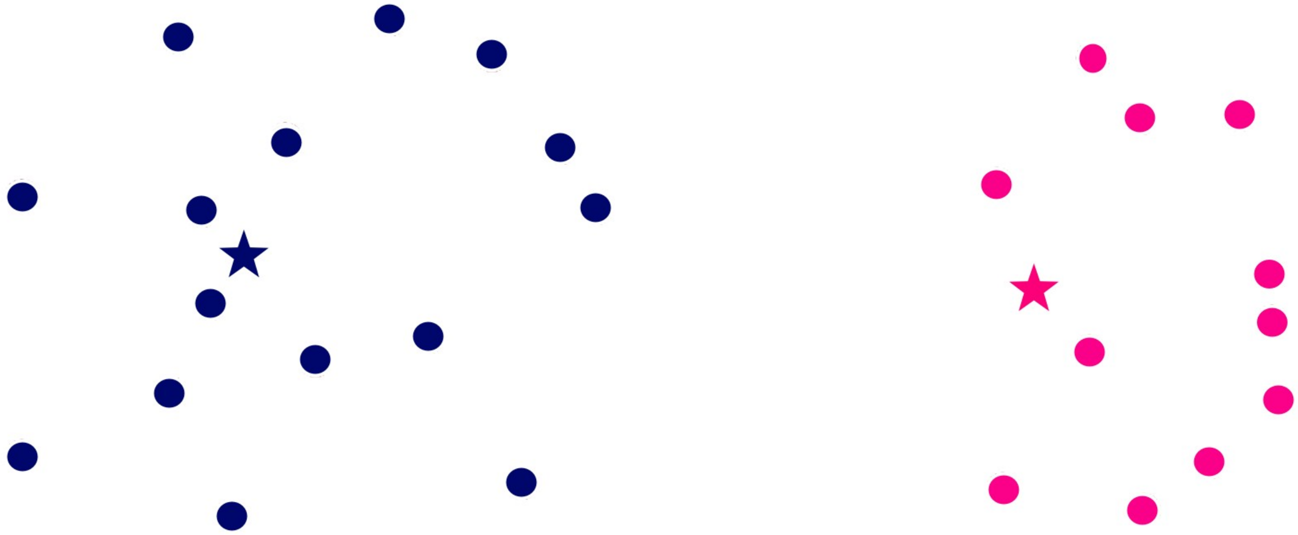
K-Means: example



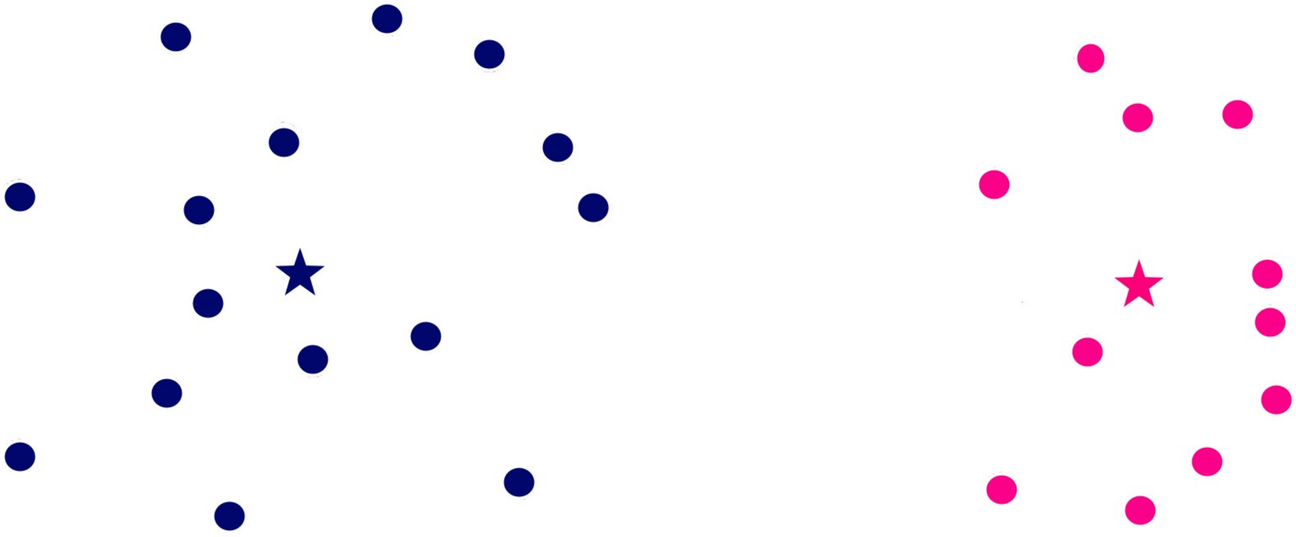
K-Means: example



K-Means: example



K-Means: example



Clustering: K-Means

- **Goal**: Assign each example $(x_1, \ldots, x_n)$ to one of the k clusters $\{\mathcal{C}_1, \ldots \mathcal{C}_k\}$.

Clustering: K-Means

- **Goal**: Assign each example $(x_1, \ldots, x_n)$ to one of the k clusters $\{\mathcal{C}_1, \ldots \mathcal{C}_k\}$.
- μ_j is the mean of all examples in the j^{th} cluster.

Clustering: K-Means

- **Goal**: Assign each example $(x_1, \ldots, x_n)$ to one of the k clusters $\{\mathcal{C}_1, \ldots \mathcal{C}_k\}$.
- μ_j is the mean of all examples in the j^{th} cluster.
- **Minimize**:

$$J = \sum_{j=1}^{k} \sum_{x_i \in \mathcal{C}_j} ||x_i - \mu_j||^2$$

Clustering: K-Means

Algorithm K-Means:

```
  Initialize randomly $\mu_1, \cdots \mu_k$.
```

Clustering: K-Means

Algorithm K-Means:

```
Initialize randomly $\mu_1, \cdots \mu_k$.
Repeat
  Assign each point $x_i$ to the cluster with the closest $\mu_j$.
```

Clustering: K-Means

Algorithm K-Means:

```
Initialize randomly $\mu_1, \cdots \mu_k$.
Repeat
  Assign each point $x_i$ to the cluster with the closest $\mu_j$.
  Calculate the new mean for each cluster as follows:
      $\mu_j = \frac{1}{|\mathcal{C}_j|} \sum_{x_i \in \mathcal{C}_j} x_i$
Until convergence$^*$.
```

Clustering: K-Means

Algorithm K-Means:

```
Initialize randomly $\mu_1, \cdots \mu_k$.
Repeat
 Assign each point $x_i$ to the cluster with the closest $\mu_j$.
 Calculate the new mean for each cluster as follows:
```

$$\mu_j = \frac{1}{|\mathcal{C}_j|} \sum_{x_i \in \mathcal{C}_j} x_i$$

```
 Until convergence*.
```

*Convergence: Means no change in the clusters OR maximum number of iterations reached.

K-Means: pros and cons

+ Easy to implement

BUT...

- Need to know K
- Suffer from the curse of dimensionality
- No theoretical foundation

K-Means: questions

1. How to set k to optimally cluster the data?

2. How to evaluate your model?

3. How to cluster non circular shapes?

K-Means: question 1

How to set k to optimally cluster the data?

G-means algorithm (Hamerly and Elkan, NIPS 2003)

1. Initialize k to be a small number
2. Run k-means with those cluster centers, and store the resulting centers as C
3. Assign each point to its nearest cluster
4. Determine if the points in each cluster fit a Gaussian distribution (Anderson-Darling test).
5. For each cluster, if the points seem to be normally distributed, keep the cluster center. Otherwise, replace it with two cluster centers.
6. Repeat this algorithm from step 2. until no more cluster centers are created.

K-Means: question 2

How to evaluate your model?

- Not trivial (as compared to counting the number of errors in classification).
- Internal evaluation: using same data. high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity. E.g., Davies-Bouldin index that takes into account both the distance inside the clusters and the distance between clusters. The lower the value of the index, the wider is the separation between different clusters, and the more tightly the points within each cluster are located together.
- External evaluation: use of ground truth of external data. E.g., mutual information, entropy, adjusted random index, etc.

K-Means: question 3

How to cluster non circular shapes?

There are other methods: spectral clustering, DBSCAN, BIRCH, etc. that handle other shapes.