K-Means Clustering

Introduction to Artificial Intelligence with Mathematics
Lecture Notes

Ganguk Hwang

Department of Mathematical Sciences
KAIST

Unsupervised learning algorithm (no labels)

Clustering

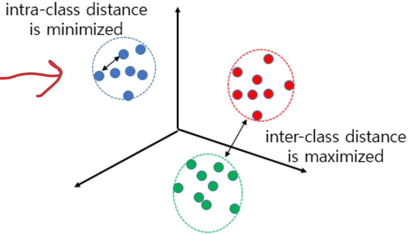
Multiple Clusters

data pts

Clustering is the process of grouping a set of observations into classes of similar observations.

clusters

- high intra-class similarity
- low inter-class similarity different clusters
- It is the most common form of unsupervised learning.
- Note that clustering is subjective.



k-means clustering

data pts

k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

- k-means clustering is one of unsupervised learning algorithms.
- It creats a labeling of observations with cluster labels.
- The labels are derived exclusively from the observations.

k-means clustering is described as follows:

For a given set of observations $\{\mathbf{x}_i | \mathbf{x}_i = (x_{i1}, ... x_{ip}), 1 \leq i \leq n\}$,

- if centroids of k clusters are denoted by $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_k\}$, and partitions are denoted by $\mathcal{C} = \{C_1, C_2, ... C_k\}$, then k-means clustering aims to partition the n observations into $k (\leq n)$ sets $C_1, C_2, ... C_k$ so as to minimize the within-cluster sum of squares (WCSS) (i.e., variance).

prob. function for k-means

Formally, the objective is to find:

data point

centroid

$$J(\mathcal{C}, \boldsymbol{\mu}) := \arg\min_{\mathcal{C}, \boldsymbol{\mu}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} ||\mathbf{x}_j - \boldsymbol{\mu}_i||^2$$

cluster

It seems to be hard to find a solution, but we have a nice heuristic algorithm.

k-means clustering algorithm

1. Let $t = 0$ and start with initial guesses $\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \cdots, \boldsymbol{\mu}_k^{(0)}$ for cluster centers (centroids).
2. For each observation, find the closest cluster centeroid.

$$C_i^{(t)} = \{\mathbf{x}_l : ||\mathbf{x}_l - \boldsymbol{\mu}_i^{(t)}||^2 \leq ||\mathbf{x}_l - \boldsymbol{\mu}_j^{(t)}||^2,\ \forall j, 1 \leq j \leq k\}$$

Centroids

(Find $\mathcal{C}$ to minimize $J(\mathcal{C}, \boldsymbol{\mu})$ while fixing $\boldsymbol{\mu}$.)

3. Replace each centroid by the average of observations in its partition.

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{x_j \in C_i^{(t)}} \mathbf{x}_j$$

(Find $\boldsymbol{\mu}$ to minimize $J(\mathcal{C}, \boldsymbol{\mu})$ while fixing $\mathcal{C}$.)

4. Iterate steps 2 and 3 until convergence (No guarantee but good)

k-means clustering as Expectation Maximization

Let $r_{ij} = 1$ if $\mathbf{x}_j \in C_i$ and $r_{ij} = 0$ if $\mathbf{x}_j \notin C_i$. We then have

$$\begin{aligned} J(\mathcal{C}, \boldsymbol{\mu}) &:= \arg\min_{\mathcal{C},\boldsymbol{\mu}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} ||\mathbf{x}_j - \boldsymbol{\mu}_i||^2 \\ &= \arg\min_{\mathcal{C},\boldsymbol{\mu}} \sum_{i=1}^{k} \sum_{j=1}^{n} r_{ij} ||\mathbf{x}_j - \boldsymbol{\mu}_i||^2. \end{aligned}$$

We want to find $\mathcal{C}$, equivalently r_{ij} and $\boldsymbol{\mu}$ to minimize J.

Best values

Step 1: Find r_{ij} to minimize $J(\mathcal{C}, \boldsymbol{\mu})$ while fixing $\boldsymbol{\mu}$. (Expectation)

$$r_{ij} = \begin{cases} 1 & \text{if } i = \arg\min_l ||\mathbf{x}_j - \boldsymbol{\mu}_l||^2 \\ 0 & \text{otherwise.} \end{cases}$$

$x_j \in C_i$

Step 2: Find $\boldsymbol{\mu}$ to minimize $J(\mathcal{C}, \boldsymbol{\mu})$ while fixing r_{ij}. (Maximization)

$$2\sum_{j=1}^{n} r_{ij}(\mathbf{x}_j - \boldsymbol{\mu}_i) = \mathbf{0}$$

Gradient of objective function J wrt μ_i's

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} r_{ij}\mathbf{x}_j}{\sum_{j=1}^{n} r_{ij}}$$

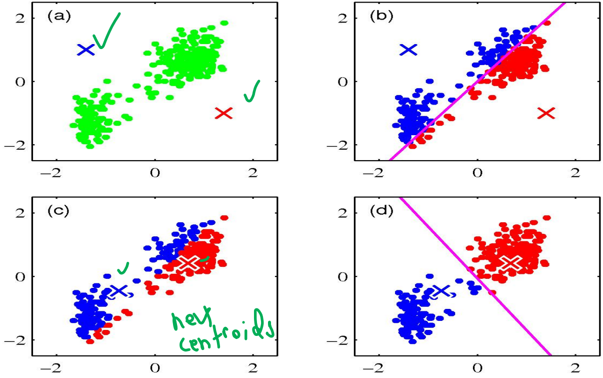
Centroid

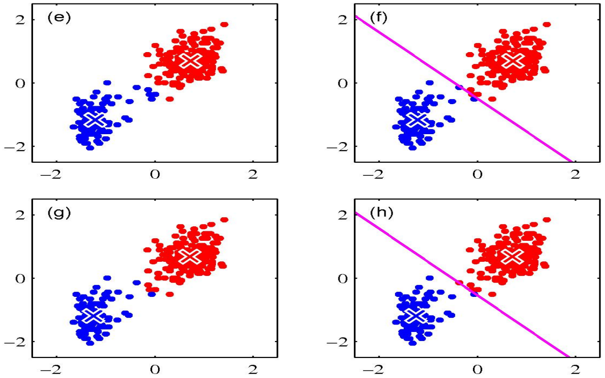
Example:

Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer 2006.

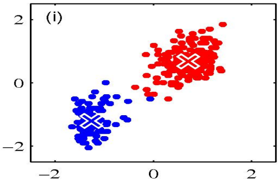
K=2 clusters

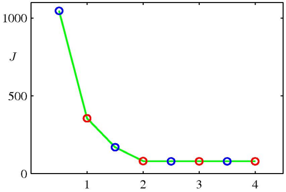
initial centroids





converged





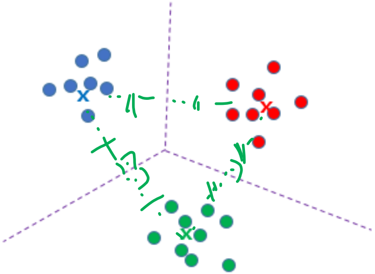
The value of J decreases as we iterate the algorithm as shown above.

cf. blue points (E steps) and red points (M steps)

Expectation

Maximization

When the Euclidean distance is used as a metric, it results in Voronoi cells.



3 boundaries

equilateral triangle

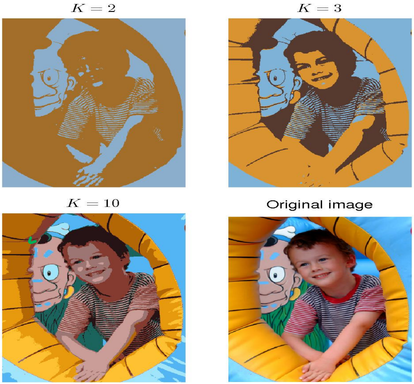
In practice:

- Try many random starting centroids (observations) and choose a solution with the smallest sum of squares
- How to choose the number k of clusters?

Select suitable K

Example: k-means clustering is applied on pixel colour values

- Pixels in each cluster are coloured by cluster mean
- Each pixel (e.g., 24-bit colour value) is represented by the cluster number (e.g., 4 bits for $k = 10$), which is a compressed version.
- This is a good example of vector quantization



3 colors

10 color pixels

Drawbacks of k-means clustering

- The value of k should be given as an input parameter.
- It might converge to a local minumum (not a global minumum).
- Its performance is sensitive to outliers.
- Euclidean distance is used as a metric and variance is used as a measure of cluster scatter, which limits the applicability of the algorithm.