CS 6316 Machine Learning

Clustering

Yangfeng Ji

Department of Computer Science
University of Virginia

Clustering

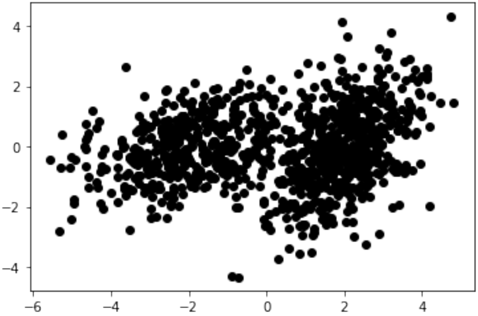
Clustering

Clustering is the task of grouping a set of objects such that similar objects end up in the same group and dissimilar objects are separated into different groups

[Shalev-Shwartz and Ben-David, 2014, Page 307]

Motivation

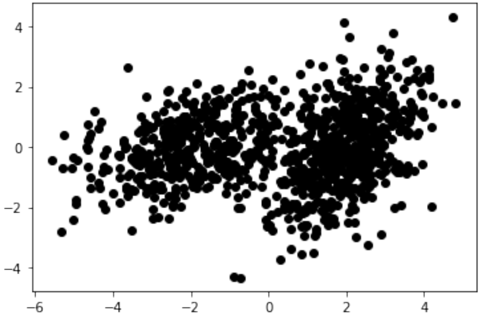
A good clustering can help us understand the data



[MacKay, 2003, Chap 20]

Movitation(II)

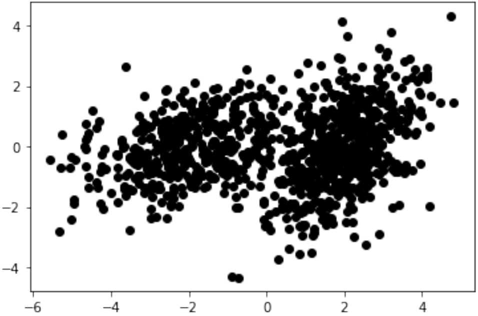
A good clustering has predictive power and can be useful to build better classifiers



[MacKay, 2003, Chap 20]

Motivation (III)

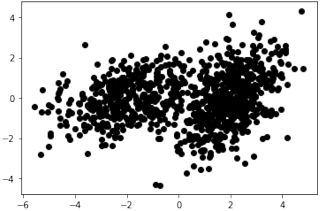
Failures of a cluster model may highlight interesting properties of data or a single data point



[MacKay, 2003, Chap 20]

Challenges

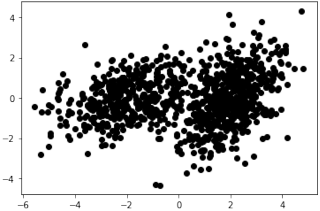
- Lack of *ground truth* — like any other unsupervised learning tasks



[Shalev-Shwartz and Ben-David, 2014, Page 307]

Challenges

- Lack of *ground truth* — like any other unsupervised learning tasks
- Definition of *similarity* measurement
  - Two images are similar
  - Two documents are similar

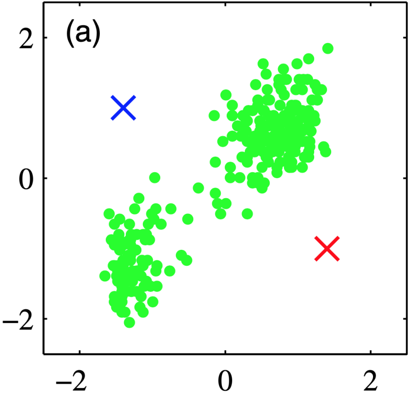


[Shalev-Shwartz and Ben-David, 2014, Page 307]

K-Means Clustering

K-Means Clustering

- A data set $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$
- Partition the data set into some number K of clusters
- K is a hyper-parameter given before learning
- Another example task of unsupervised learning



Objective Function

- Introduce $r_i \in [K]$ for each data point $\boldsymbol{x}_i$, which is a determinstric variable
- The objective function of k-means clustering

$$J(\boldsymbol{r}, \boldsymbol{\mu}) = \sum_{i=1}^{m} \sum_{k=1}^{K} \delta(r_i = k) \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2 \qquad (1)$$

where $\{\boldsymbol{\mu}_k\}_{k=1}^{K} \in \mathbb{R}^d$. Each $\boldsymbol{\mu}_k$ is called a *prototype* associated with the k-th cluster.

Objective Function

- Introduce $r_i \in [K]$ for each data point $\boldsymbol{x}_i$, which is a determinstric variable
- The objective function of k-means clustering

$$J(\boldsymbol{r}, \boldsymbol{\mu}) = \sum_{i=1}^{m} \sum_{k=1}^{K} \delta(r_i = k) \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2 \tag{1}$$

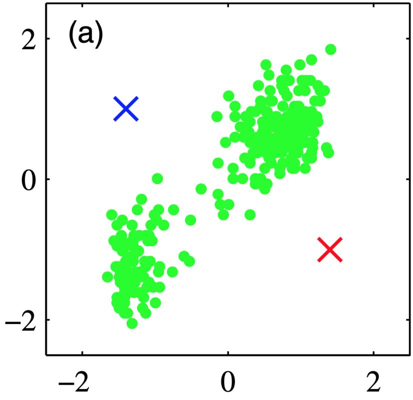
where $\{\boldsymbol{\mu}_k\}_{k=1}^{K} \in \mathbb{R}^d$. Each $\boldsymbol{\mu}_k$ is called a *prototype* associated with the k-th cluster.

- Learning: minimize equation 1

$$\operatorname*{argmin}_{\boldsymbol{r}, \boldsymbol{\mu}} J(\boldsymbol{r}, \boldsymbol{\mu}) \tag{2}$$

Learning: Initialization

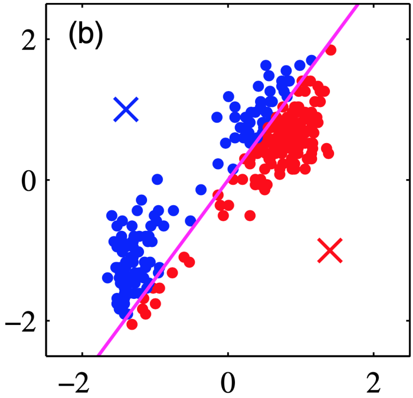
Randomly initialize $\{\mu_k\}_{k=1}^{K}$



Learning: Assignment Step

Given $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$, for each $\boldsymbol{x}_i$, find the value of r_i is equivalent to assign the data point to a cluster

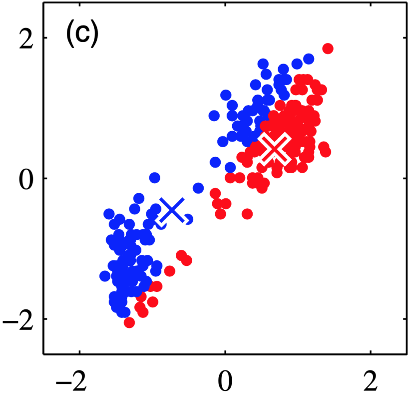
$$r_i \leftarrow \operatorname*{argmin}_{k'} \|\boldsymbol{x}_i - \boldsymbol{\mu}_{k'}\|_2^2 \tag{3}$$



Learning: Update Step

Given $\{r_i\}_{i=1}^m$, the algorithm updates $\boldsymbol{\mu}_k$ as

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^m \delta(r_i = k)\boldsymbol{x}_i}{\sum_{i=1}^m \delta(r_i = k)} \tag{4}$$



Algorithm

With some randomly initialized $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$, iterate the following two steps until converge

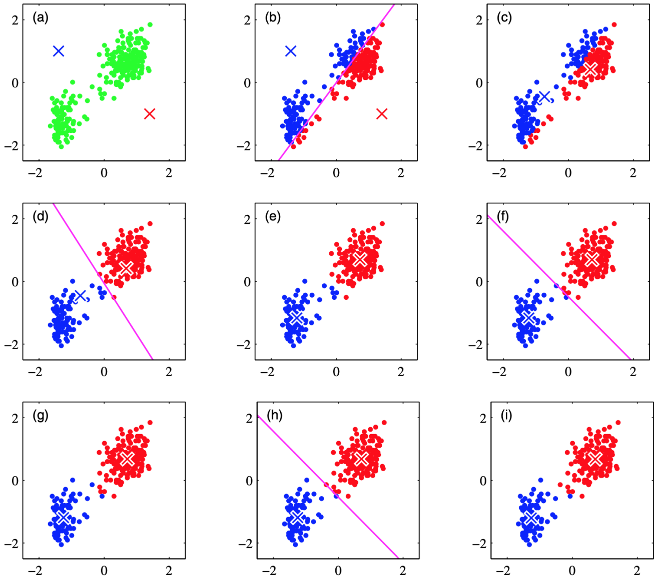
Assignment Step Assign r_i for each $\boldsymbol{x}_i$

$$r_i \leftarrow \operatorname*{argmin}_{k'} \|\boldsymbol{x}_i - \boldsymbol{\mu}_{k'}\|_2^2 \tag{5}$$

Update Step Updates $\boldsymbol{\mu}_k$ with $\{r_i\}_{i=1}^{m}$

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{m} \delta(r_i = k)\boldsymbol{x}_i}{\sum_{i=1}^{m} \delta(r_i = k)} \tag{6}$$

Example (Cont.)



From GMMs to K-means

Gaussian Mixture Models

Consider a GMM with two components

$$
\begin{aligned}
q(\boldsymbol{x}, z) &= q(z)q(\boldsymbol{x} \mid z) \\
&= \alpha^{\delta(z=1)} \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)^{\delta(z=1)} \\
&\quad \cdot (1-\alpha)^{\delta(z=2)} \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)^{\delta(z=2)} \qquad (7)
\end{aligned}
$$

Gaussian Mixture Models

Consider a GMM with two components

$$
\begin{aligned}
q(\boldsymbol{x}, z) &= q(z)q(\boldsymbol{x} \mid z) \\
&= \alpha^{\delta(z=1)} \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)^{\delta(z=1)} \\
&\quad \cdot (1-\alpha)^{\delta(z=2)} \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)^{\delta(z=2)}
\end{aligned} \tag{7}
$$

And the marginal probability $p(\boldsymbol{x})$ is

$$
\begin{aligned}
q(\boldsymbol{x}) &= q(z=1)q(\boldsymbol{x} \mid z=1) + q(z=2)q(\boldsymbol{x} \mid z=2) \\
&= \alpha \cdot \mathcal{N}(x; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\alpha) \cdot \mathcal{N}(x; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)
\end{aligned} \tag{8}
$$

A Special Case

Consider the first component in this GMM with parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$

- Assume $\boldsymbol{\Sigma}_1 = \epsilon \boldsymbol{I}$, then

$$
\begin{aligned}
|\boldsymbol{\Sigma}_1| &= \epsilon^d && (9)\\
(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) &= \frac{1}{\epsilon} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 && (10)
\end{aligned}
$$

A Special Case

Consider the first component in this GMM with parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$

- Assume $\boldsymbol{\Sigma}_1 = \epsilon \boldsymbol{I}$, then

$$
\begin{aligned}
|\boldsymbol{\Sigma}_1| &= \epsilon^d && (9)\\
(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) &= \frac{1}{\epsilon} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 && (10)
\end{aligned}
$$

- A Gaussian component can be simplified as

$$
\begin{aligned}
q(\boldsymbol{x}_i \mid z_i = 1) &= \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right)\\
&= \frac{1}{(2\pi\epsilon)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\epsilon} \|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2^2 \right) && (11)
\end{aligned}
$$

A Special Case

Consider the first component in this GMM with parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$

- Assume $\boldsymbol{\Sigma}_1 = \epsilon \boldsymbol{I}$, then

$$\begin{aligned} |\boldsymbol{\Sigma}_1| &= \epsilon^d && (9) \\ (\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) &= \frac{1}{\epsilon} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 && (10) \end{aligned}$$

- A Gaussian component can be simplified as

$$\begin{aligned} q(\boldsymbol{x}_i \mid z_i = 1) &= \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \Big) \\ &= \frac{1}{(2\pi\epsilon)^{\frac{d}{2}}} \exp\big(-\frac{1}{2\epsilon} \|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2^2 \big) && (11) \end{aligned}$$

- Similar results with the second component with $\boldsymbol{\Sigma}_2 = \epsilon \boldsymbol{I}$

A Special Case (II)

From the previous discussion, we know that, given $\boldsymbol{\theta}$, $q(z_i \mid \boldsymbol{x}_i)$ is computed as

$$
\begin{aligned}
q(z_i = 1 \mid \boldsymbol{x}_i) &= \frac{\alpha \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}{\alpha \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\alpha) \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)} \\
&= \frac{\alpha \exp(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2^2)}{\alpha \exp(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2^2) + (1-\alpha)\exp(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_2\|_2^2}
\end{aligned}
$$

A Special Case (II)

From the previous discussion, we know that, given $\boldsymbol{\theta}$, $q(z_i \mid \boldsymbol{x}_i)$ is computed as

$$
\begin{aligned}
q(z_i = 1 \mid \boldsymbol{x}_i) &= \frac{\alpha \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}{\alpha \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\alpha) \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)} \\
&= \frac{\alpha \exp(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2^2)}{\alpha \exp(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2^2) + (1-\alpha)\exp(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_2\|_2^2}
\end{aligned}
$$

- When $\epsilon \rightarrow 0$

$$
q(z_i = 1 \mid \boldsymbol{x}_i) \rightarrow \begin{cases} 1 & \|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2 < \|\boldsymbol{x}_i - \boldsymbol{\mu}_2\|_2 \\ 0 & \|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2 > \|\boldsymbol{x}_i - \boldsymbol{\mu}_2\|_2 \end{cases} \tag{12}
$$

A Special Case (II)

From the previous discussion, we know that, given $\boldsymbol{\theta}$, $q(z_i \mid \boldsymbol{x}_i)$ is computed as

$$\begin{aligned} q(z_i = 1 \mid \boldsymbol{x}_i) &= \frac{\alpha \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}{\alpha \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-\alpha) \cdot \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)} \\ &= \frac{\alpha \exp(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2^2)}{\alpha \exp(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2^2) + (1-\alpha) \exp(-\frac{1}{2\epsilon}\|\boldsymbol{x}_i - \boldsymbol{\mu}_2\|_2^2} \end{aligned}$$

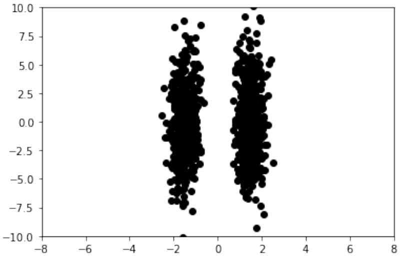
- When $\epsilon \rightarrow 0$

$$q(z_i = 1 \mid \boldsymbol{x}_i) \rightarrow \begin{cases} 1 & \|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2 < \|\boldsymbol{x}_i - \boldsymbol{\mu}_2\|_2 \\ 0 & \|\boldsymbol{x}_i - \boldsymbol{\mu}_1\|_2 > \|\boldsymbol{x}_i - \boldsymbol{\mu}_2\|_2 \end{cases} \tag{12}$$

- r_i in K-means is a very special case of z_i in GMM

When K-means Will Fail?

Recall that K-means is an extreme case of GMM with $\boldsymbol{\Sigma} = \epsilon \boldsymbol{I}$ and $\epsilon \to 0$

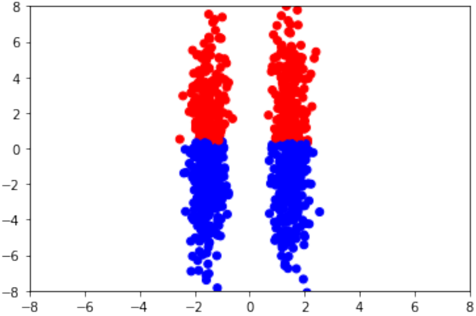


Parameters

$$\begin{aligned} \boldsymbol{\mu}_1 = [1.5, 0]^\top \qquad & \boldsymbol{\mu}_2 = [-1.5, 0]^\top \\ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \quad = & \quad \mathrm{diag}(0.1, 10.0) \end{aligned} \tag{13}$$

When K-means Will Fail? (II)

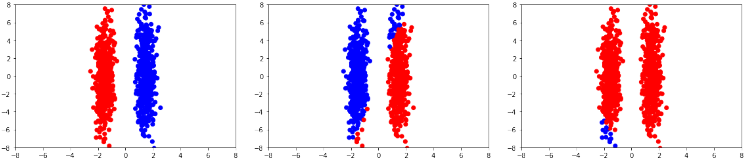
Recall that K-means is an extreme case of GMM with $\boldsymbol{\Sigma} = \epsilon \boldsymbol{I}$ and $\epsilon \longrightarrow 0$



How About GMM?

With the following setup[1]

- Randomly initialize GMM parameters (instead of using K-means to initalize)
- Set `covariance_type` to be `tied`



[1] Please refer to the demo code for more detail

Reference

Bishop, C. M. (2006).
Pattern recognition and machine learning.
springer.

MacKay, D. (2003).
Information theory, inference and learning algorithms.
Cambridge university press.

Shalev-Shwartz, S. and Ben-David, S. (2014).
Understanding machine learning: From theory to algorithms.
Cambridge university press.