

# CS 6316 Machine Learning

## Clustering

---

Yangfeng Ji

Information and Language Processing Lab  
Department of Computer Science  
University of Virginia



# Clustering

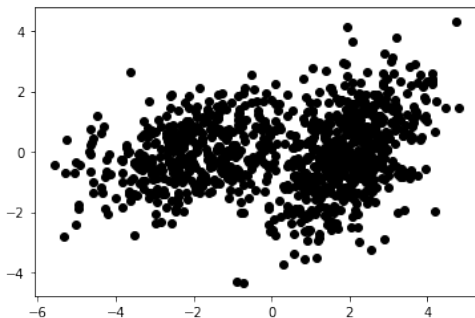
---

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

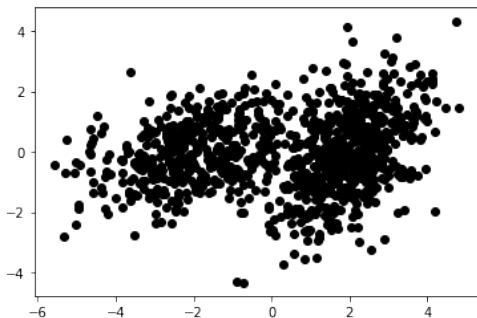
A good clustering can help us understand the data



[MacKay, 2003, Chap 20]

## Movitation(II)

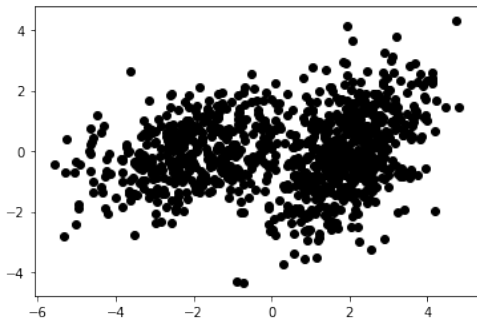
A good clustering has predictive power and can be useful to build better classifiers



[MacKay, 2003, Chap 20]

## Motivation (III)

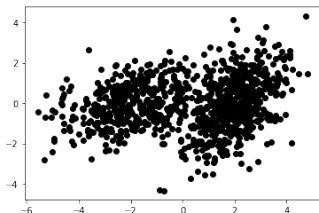
Failures of a cluster model may highlight interesting properties of data or a single data point



[MacKay, 2003, Chap 20]

# Challenges

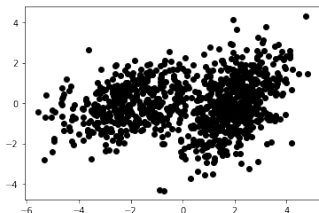
- ▶ Lack of *ground truth* — like any other unsupervised learning tasks



[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]

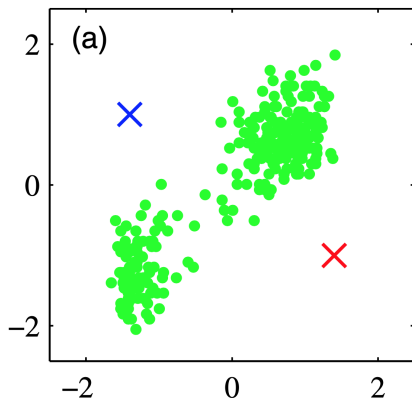


# *K*-Means Clustering

---

# K-Means Clustering

- ▶ A data set  $S = \{x_1, \dots, x_m\}$  with  $x_i \in \mathbb{R}^d$
- ▶ 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  $x_i$ , which is a deterministic variable
- ▶ The objective function of  $k$ -means clustering

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

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  $x_i$ , which is a deterministic variable
- ▶ The objective function of  $k$ -means clustering

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

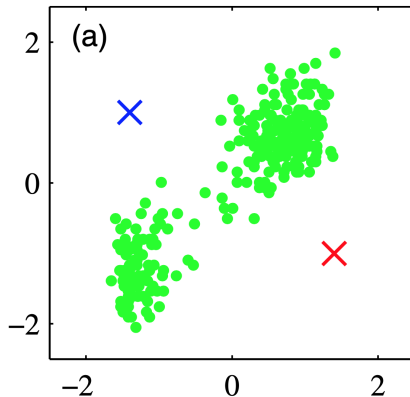
where  $\{\boldsymbol{\mu}_k\}_{k=1}^K \in \mathbb{R}^d$ . Each  $\boldsymbol{\mu}_k$  is called a *prototype* associated with the  $k$ -th cluster.

- ▶ Learning: minimize equation 1

$$\underset{\mathbf{r}, \boldsymbol{\mu}}{\operatorname{argmin}} J(\mathbf{r}, \boldsymbol{\mu}) \quad (2)$$

# Learning: Initialization

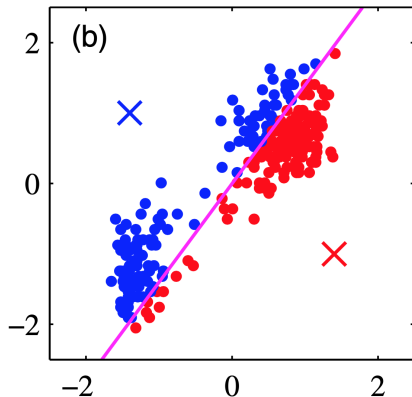
Randomly initialize  $\{\mu_k\}_{k=1}^K$



# Learning: Assignment Step

Given  $\{\mu_k\}_{k=1}^K$ , for each  $x_i$ , find the value of  $r_i$  is equivalent to assign the data point to a cluster

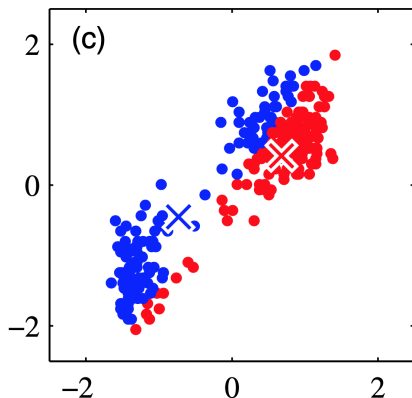
$$r_i \leftarrow \underset{k'}{\operatorname{argmin}} \|x_i - \mu_{k'}\|_2^2 \quad (3)$$



# Learning: Update Step

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

$$\mu_k = \frac{\sum_{i=1}^m \delta(r_i = k) x_i}{\sum_{i=1}^m \delta(r_i = k)} \quad (4)$$



► The updated  $\mu_k$  equals to the **mean** of all data points in cluster  $k$

► The **mean** of this cluster is called  $\mu_k$  (mean of cluster  $k$ )

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

**Assignment Step** Assign  $r_i$  for each  $x_i$

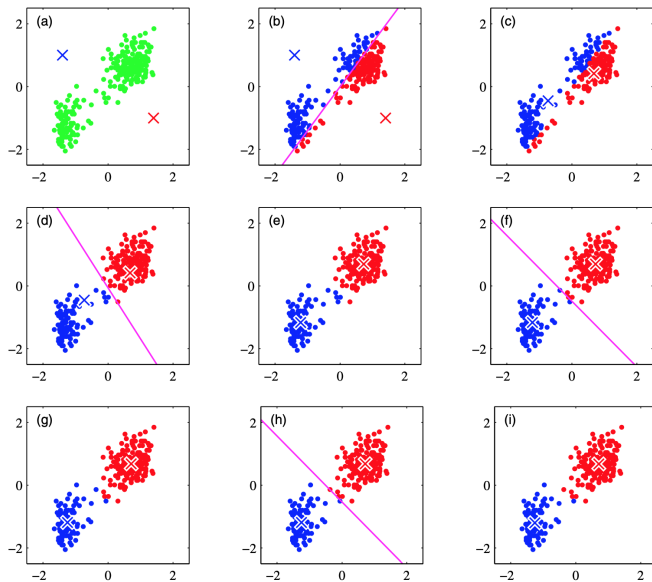
$$r_i \leftarrow \underset{k'}{\operatorname{argmin}} \|x_i - \mu_{k'}\|_2^2 \quad (5)$$

**Update Step** Updates  $\mu_k$  with  $\{r_i\}_{i=1}^m$

$$\mu_k = \frac{\sum_{i=1}^m \delta(r_i = k) x_i}{\sum_{i=1}^m \delta(r_i = k)} \quad (6)$$



# Example (Cont.)



## From GMMs to $K$ -means

---

Consider a GMM with two components

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

Consider a GMM with two components

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

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

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

# A Special Case

Consider the first component in this GMM with parameters  $\mu_1$  and  $\Sigma_1$

- ▶ Assume  $\Sigma_1 = \epsilon I$ , then

$$|\Sigma_1| = \epsilon^d \quad (9)$$

$$(x - \mu_1)^T \Sigma_1^{-1} (x - \mu) = \frac{1}{\epsilon} \|x - \mu\|_2^2 \quad (10)$$

# A Special Case

Consider the first component in this GMM with parameters  $\mu_1$  and  $\Sigma_1$

- ▶ Assume  $\Sigma_1 = \epsilon I$ , then

$$|\Sigma_1| = \epsilon^d \quad (9)$$

$$(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) = \frac{1}{\epsilon} \|x - \mu\|_2^2 \quad (10)$$

- ▶ A Gaussian component can be simplified as

$$\begin{aligned} q(x_i | z_i = 1) &= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_i - \mu_1)^T \Sigma_1^{-1} (x_i - \mu_1)\right) \\ &= \frac{1}{(2\pi\epsilon)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\epsilon} \|x_i - \mu_1\|_2^2\right) \end{aligned} \quad (11)$$

# A Special Case

Consider the first component in this GMM with parameters  $\mu_1$  and  $\Sigma_1$

- ▶ Assume  $\Sigma_1 = \epsilon I$ , then

$$|\Sigma_1| = \epsilon^d \quad (9)$$

$$(x - \mu_1)^\top \Sigma_1^{-1} (x - \mu_1) = \frac{1}{\epsilon} \|x - \mu\|_2^2 \quad (10)$$

- ▶ A Gaussian component can be simplified as

$$\begin{aligned} q(x_i | z_i = 1) &= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (x_i - \mu_1)^\top \Sigma_1^{-1} (x_i - \mu_1)\right) \\ &= \frac{1}{(2\pi\epsilon)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\epsilon} \|x_i - \mu_1\|_2^2\right) \end{aligned} \quad (11)$$

- ▶ Similar results with the second component with  $\Sigma_2 = \epsilon I$

## A Special Case (II)

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

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



## A Special Case (II)

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

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

► When  $\epsilon \rightarrow 0$

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

## A Special Case (II)

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

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

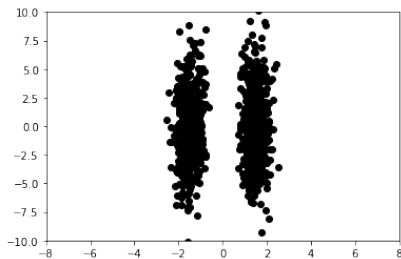
- ▶ When  $\epsilon \rightarrow 0$

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

- ▶  $r_i$  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  $\Sigma = \epsilon I$  and  $\epsilon \rightarrow 0$

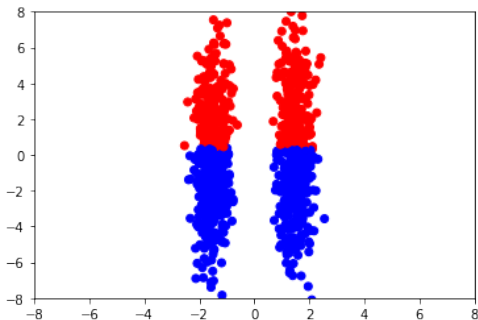


Parameters

$$\begin{aligned}\mu_1 &= [1.5, 0]^T & \mu_2 &= [-1.5, 0]^T \\ \Sigma_1 = \Sigma_2 &= \text{diag}(0.1, 10.0)\end{aligned}\tag{13}$$

## When $K$ -means Will Fail? (II)

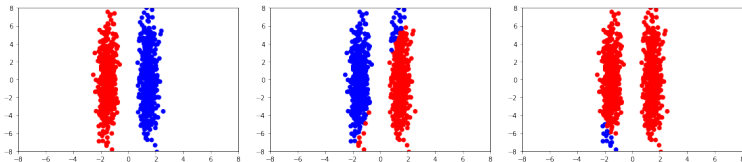
Recall that  $K$ -means is an extreme case of GMM with  $\Sigma = \epsilon I$  and  $\epsilon \rightarrow 0$



# How About GMM?

With the following setup<sup>1</sup>

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

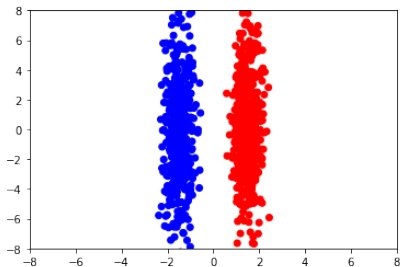


---

<sup>1</sup>Please refer to the demo code for more detail

# Spectral Clustering

Instead of computing the distance between data points to some prototypes, spectral clustering is purely based on the similarity between data points, which can address the problem like this



[Shalev-Shwartz and Ben-David, 2014, Section 22.3]



Bishop, C. M. and Nasrabadi, N. M. (2006).  
*Pattern recognition and machine learning*, volume 4.  
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.