EM Algorithm

Chapter 9, Statistical learning methods

Yu Hao

January 16, 2019

NaMI, Tongji University

Table of contents

1. K-means
2. Gaussian Mixture Model
3. EM Algorithm

K-means

K-means recap

- Randomly initialize K centers
  - ▷ $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}$
- Classify: Assign each point $j \in \{1, \ldots, N\}$ to nearest center:
  - ▷ $C^{(t)}(j) \leftarrow \arg\min_i \|\boldsymbol{\mu}_i - \mathbf{x}_j\|^2$
- Recenter: $\boldsymbol{\mu}_i$ becomes centroid of its points
  - ▷ $\boldsymbol{\mu}_i^{(t+1)} \leftarrow \arg\min_{\boldsymbol{\mu}} \sum_{j:C(j)=i} \|\boldsymbol{\mu} - \mathbf{x}_j\|^2$
  - ▷ Equivalent to $\boldsymbol{\mu}_i \leftarrow$ average of its points!

What is K-means optimizing?

- Potential function $F(\boldsymbol{\mu}, C)$ of centers $\boldsymbol{\mu}$ and point allocations C:

$$F(\boldsymbol{\mu}, C) = \sum_{j=1}^{N} \left\| \boldsymbol{\mu}_{C(j)} - \boldsymbol{x}_j \right\|^2 \tag{1}$$

- Optimal K-means:
  - ▷ $\min_{\boldsymbol{\mu}} \min_{C} F(\boldsymbol{\mu}, C)$

K-means algorithm

- Optimize potential function:

$$\min_{\boldsymbol{\mu}} \min_{C} F(\boldsymbol{\mu}, C) = \min_{\boldsymbol{\mu}} \min_{C} \sum_{i=1}^{K} \sum_{j:C(j)=i} \left\|\boldsymbol{\mu}_i - \boldsymbol{x}_i\right\|^2 \quad (2)$$

- K-means algorithm:
  - ▷ (1) Fix $\boldsymbol{\mu}$, optimize C

$$\min_{C} \sum_{j=1}^{N} \left\|\boldsymbol{\mu}_{C(j)} - x_j\right\|^2 = \sum_{j=1}^{N} \min_{C(j)} \left\|\boldsymbol{\mu}_{C(j)} - \boldsymbol{x}_j\right\|^2$$

Exactly first step: assign each point to the nearest cluster center

K-means algorithm

- Optimize potential function:

$$\min_{\boldsymbol{\mu}} \min_{C} F(\boldsymbol{\mu}, C) = \min_{\boldsymbol{\mu}} \min_{C} \sum_{i=1}^{K} \sum_{j:C(j)=i} \left\|\boldsymbol{\mu}_i - \boldsymbol{x}_i\right\|^2 \quad (2)$$

- K-means algorithm:
  - ▷ (2) Fix C, optimize $\boldsymbol{\mu}$

$$\min_{\boldsymbol{\mu}} \sum_{i=1}^{K} \sum_{j:C(j)=i} \left\|\boldsymbol{\mu}_i - \boldsymbol{x}_j\right\|^2 = \sum_{i}^{K} \min_{\boldsymbol{\mu}_i} \sum_{j:C(j)=i} \left\|\boldsymbol{\mu}_i - \boldsymbol{x}_j\right\|^2$$

Exactly second step: average of points in cluster i

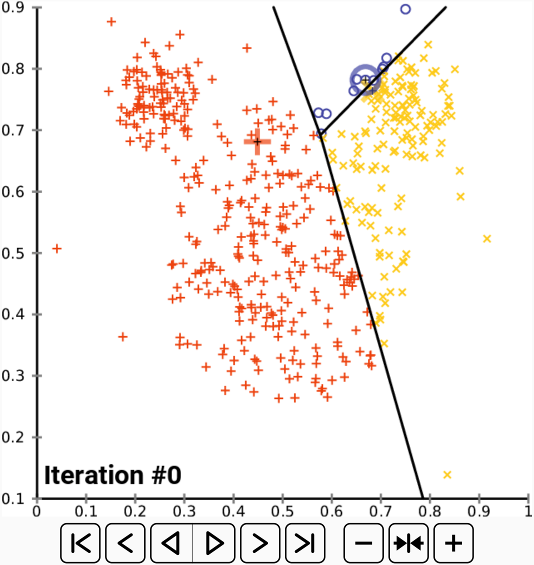
K-means algorithm

- Optimize potential function:

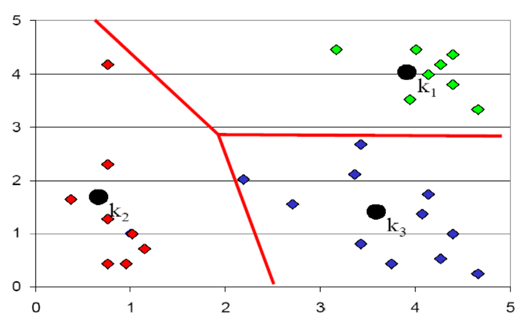
$$\min_{\boldsymbol{\mu}} \min_{C} F(\boldsymbol{\mu}, C) = \min_{\boldsymbol{\mu}} \min_{C} \sum_{i=1}^{K} \sum_{j:C(j)=i} \left\| \boldsymbol{\mu}_i - \boldsymbol{x}_i \right\|^2 \quad (2)$$

- K-means algorithm:
  - ▷ (1) Fix $\boldsymbol{\mu}$, optimize C **Expectation step**
  - ▷ (2) Fix C, optimize $\boldsymbol{\mu}$ **Maximization step**
  - ▷ Today, we will see a generalization of this approach: **EM algorithm**

Iterations of K-means



K-means decision boundaries



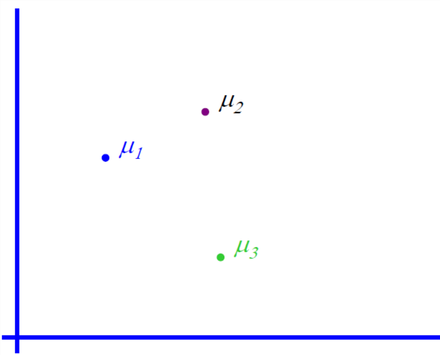
"Linear"
Decision
Boundaries

- **Generative Model:**
  Assume data comes from a mixture of K Gaussians distributions with same variance.

K-means: Generative model

Mixture of K Gaussians distributions: (Multi-model distribution)

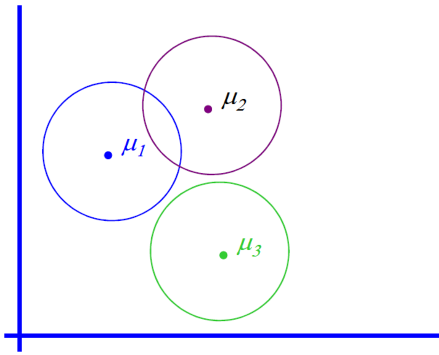
- There are K components
- Component i has an associated mean vector $\boldsymbol{\mu}_i$



K-means: Generative model

Mixture of K Gaussians distributions: (Multi-model distribution)

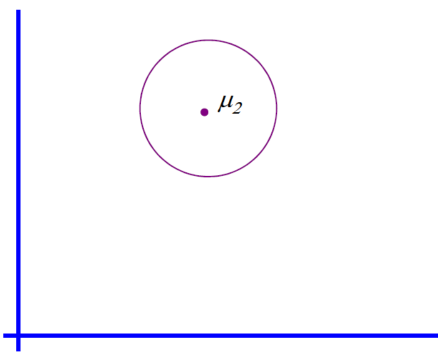
- There are K components
- Component i has an associated mean vector $\boldsymbol{\mu}_i$
- Each component generates data from a Gaussian with mean $\boldsymbol{\mu}_i$ and covariance matrix $\sigma^2\mathbf{I}$



K-means: Generative model

Mixture of K Gaussians distributions: (Multi-model distribution)

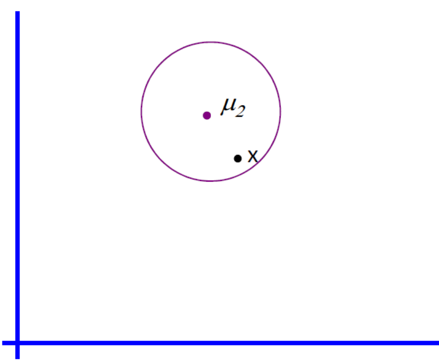
- Each data point is generated according to the following recipe:
- (1) Pick a component at random: choose component i with probability $P(y = i)$



K-means: Generative model

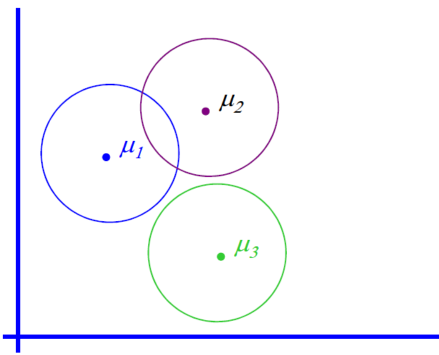
Mixture of K Gaussians distributions: (Multi-model distribution)

- Each data point is generated according to the following recipe:
- (1) Pick a component at random: choose component i with probability $P(y = i)$
- (2) Data point $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 \mathbf{I})$



K-means: Generative model

Mixture of K Gaussians distributions: (Multi-model distribution)



- $p(\boldsymbol{x}|y=i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 \mathbf{I})$
- $p(\boldsymbol{x}) = \sum_i p(\boldsymbol{x}|y=i) P(y=i)$
  - Mixture component
  - Mixture proportion

K-means: Generative model

Mixture of K Gaussians distributions: (Multi-model distribution)

- $p(\mathbf{x}|y=i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2\mathbf{I})$

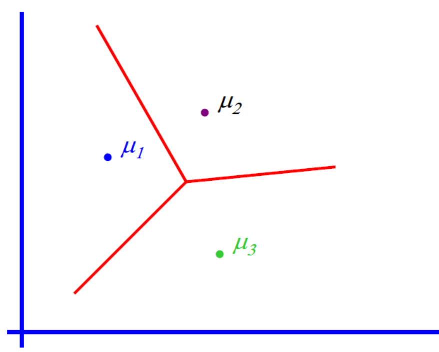
- Gaussian Bayes Classifier:

$$\log \frac{P(y=i|\mathbf{x})}{P(y=j|\mathbf{x})}$$

$$= \log \frac{p(\mathbf{x}|y=i)P(y=i)}{p(\mathbf{x}|y=j)P(y=j)}$$

$$= \mathbf{w}^T\mathbf{x}$$

- $\mathbf{w}$ depends on $\boldsymbol{\mu}, \sigma^2, P(y)$



"Linear Decision boundary"
second-order terms cancel out

Gaussian Mixture Model

EM Algorithm