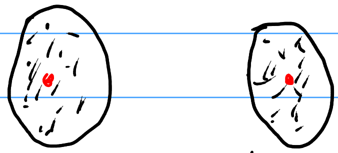
K-Means Clustering + Mixture Models

Unsupervised Learning

Clustering



$n=5,\ k=3$

$C_1 = \{X_1, X_5\}$

$C_2 = \{X_2\}$

$C_3 = \{X_3, X_4\}$

Given $X = \{X_1, \dots, X_n\}$, partition in $C_1, \dots, C_k$. ← hyperparam

Goals (informally):

- Points in a cluster are similar
- points in diff clusters are dissimilar

$$\underset{\text{partition } C_1,\dots,C_k}{\text{minimize}} \sum_{j=1}^{k} \underbrace{W(C_j)}_{\text{Cost fn for points in cluster } j}$$

K-means: objective function

$$W(C_j) = \frac{1}{|C_j|} \sum_{\substack{x_i, x_{i'} \\ \in C_j}} \|X_i - X_{i'}\|_2^2 = 2 \sum_{x_i \in C_j} \|X_i - \mu_j\|_2^2$$

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} X_i$$

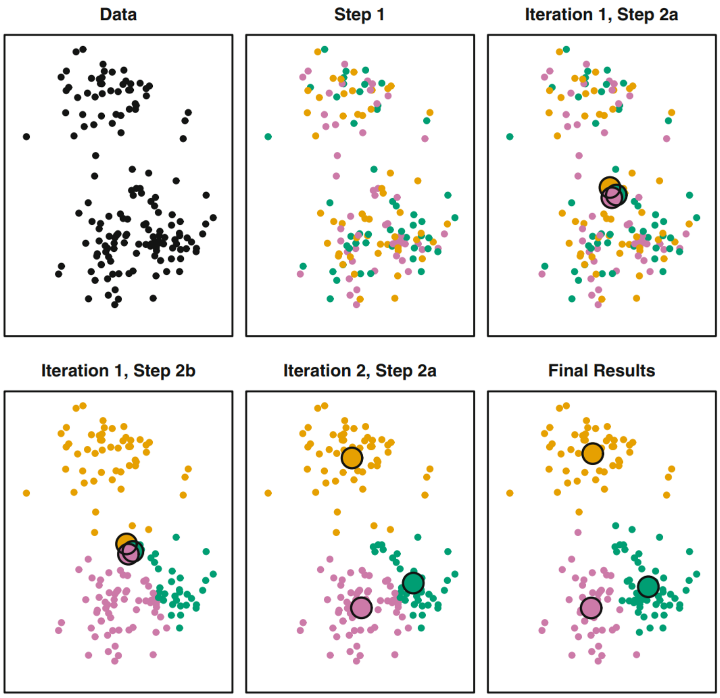
Slow Alg: Try all partitions: k^n partitions

Alg:

- Initialize partition $C_1, \dots, C_k$

Repeat until converged {

- For each cluster C_j, compute centroid $\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} X_i$
- For each point X_i, assign it to cluster w/ nearest centroid: $\underset{j}{\operatorname{argmin}} \|X_i - \mu_j\|_2$.



Drawbacks

- Slow converge
- Local optimum

Solutions:

- Repeat many times
- Better initialization
  (K-means++)

converge faster

Generative Models

distribution

$X_1, \dots, X_n$, drawn i.i.d. from some dist p_θ

Goal: Output $\hat{p} \approx p_\theta$

Simple: $X_1, \dots, X_n \sim N(\mu, 1)$ $p_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$

Maximum likelihood est (MLE):

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_\mu(X_i) = \underset{\mu}{\operatorname{argmax}} \sum_{i=1}^{n} -(X_i - \mu)^2$$

$$= \frac{1}{n} \sum X_i$$

$\boxed{Xi = u}$

Output $\hat{p} = N(\hat{\mu}, 1)$

Mixture Model

$$p_\theta(x) = \sum_{j=1}^{k} \pi_j \, p^{(j)}_{\theta^{(j)}}(x)$$

$\{\pi_j\text{'s}$, $\theta^{(j)}\text{'s}$

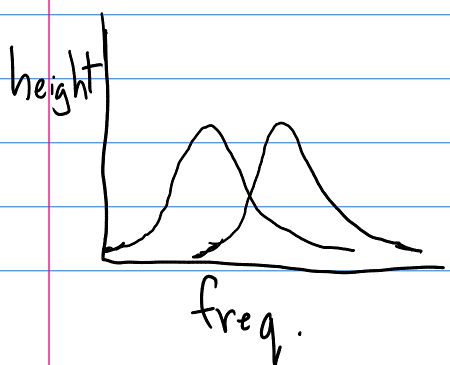
Prob dist for component j

mixing weight

$\pi_j \geq 0$, $\sum \pi_j = 1$

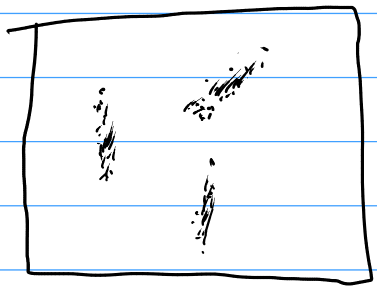
Sampling procedure:

1. Draw sample from $\{1, \dots, k\}$, from π
2. Output sample from $p^{(j)}_{\theta^{(j)}}(x)$



Gaussian Mixture Model (GMM)

$$P_\theta(x) = \sum_{j=1}^{k} \pi_j N(\mu_j, \Sigma_j, x)$$



Z_i = component that X_i was sampled from

$Z_i \in \{1, \dots, K\}$

$$\hat{\theta} = \underset{\theta}{\arg\max} \sum_{i=1}^{n} \log P_\theta(X_i)$$

$$= \underset{\theta=(\mu_j, \Sigma_j, \pi_j)}{\arg\max} \sum \log\left(\sum_{j=1}^{K} 1\{Z_i=j\} \pi_j N(\mu_j, \Sigma_j, X_i)\right)$$

- Use $\hat{\pi}_j = \frac{1}{n} \sum_{Z_i=j} 1$

$$\hat{\mu}_j = \frac{1}{\sum_{Z_i=j} 1} \sum_{Z_i=j} X_i, \quad \hat{\Sigma}_j = \frac{1}{\sum_{Z_i=j} 1} \sum_{Z_i=j} (X_i - \hat{\mu}_j)(X_i - \hat{\mu}_j)^T$$

Expectation Maximization (EM):

- "Soft" version of k-means

eg: point 50% chance belong to cluster 1, 20% belong to cluster 2...

1. Given θ, fractionally assign points X_i to clusters
2. Given fractional assignment of X_i to clusters, compute best θ.

$$\arg\max_{\theta} \sum_{i=1}^{n} \log P_\theta(x_i)$$

$$\log P_\theta(x_i)$$

label of the point being j

$$= \log \sum_{j=1}^{k} P_\theta(x_i, z_i = j)$$

$$= \log \sum_{j=1}^{k} \frac{q_i(z_i=j)}{q_i(z_i=j)} P_\theta(x_i, z_i = j)$$

$$= \log \sum_{j=1}^{k} q_i(z_i=j) \left(\frac{P_\theta(x_i, z_i=j)}{q_i(z_i=j)} \right)$$

$$= \log \mathop{E}_{z_i \sim q_i}\left[\frac{P_\theta(x_i, z_i)}{q_i(z_i)} \right]$$

$$\geq \mathop{E}_{z_i \sim q_i}\left[\log\left(\frac{P_\theta(x_i, z_i)}{q_i(z_i)} \right) \right]$$

$$= \sum_{j=1}^{k} q_i(z_i=j) \log P_\theta(x_i, z_i=j) - \sum_{j=1}^{k} q_i(z_i=j) \log q_i(z_i=j)$$

$$\arg\max_{\theta} \sum_{i=1}^{n} \log P_\theta(x_i) \geq$$

$$\arg\max_{\theta, q_i} \sum_{i=1}^{n} \sum_{j=1}^{k} q_i(z_i=j) \log P_\theta(x_i, z_i=j) - \sum_{j=1}^{k} q_i(z_i=j) \log q_i(z_i=j)$$

$$\underset{\theta, q_i}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{K} q_i(z_i=j) \log P_\theta(x_i, z_i=j) - \sum_{j=1}^{K} q_i(z_i=j) \log q_i(z_i=j)$$

E step: Fix θ, opt q_i's

$$\underset{\text{dist } q_i}{\operatorname{argmax}} \sum_{j=1}^{K} q_i(z_i=j) \log P_\theta(x_i, z_i=j) - \sum_{j=1}^{K} q_i(z_i=j) \log q_i(z_i=j)$$

$$\underset{\text{dist } q_i}{\operatorname{argmax}} \sum_{j=1}^{K} q_i(z_i=j) \log P_\theta(z_i=j \mid x_i) - \sum_{j=1}^{} q_i(z_i=j) \log q_i(z_i=j) + \overbrace{\cancel{\sum_{j=1}^{K} q_i(z_i=j)}}^{1} \log P_\theta(x_i)$$

$$= \underset{\text{dist } q_i}{\operatorname{argmin}} \underbrace{\sum_{j=1}^{K} q_i(z_i=j) \log\left(\frac{q_i(z_i=j)}{P_\theta(z_i=j \mid x_i)}\right)}_{KL(q_i(z_i) \,\|\, P_\theta(z_i \mid x_i))}$$

Measure of dist

≥ 0

$= 0$ when $q_i(z_i) = P_\theta(z_i \mid x_i)$

$= P_\theta(z_i \mid x_i)$

M Step: Fix q_i's, opt θ

$$\underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{K} q_i(z_i=j) \log P_\theta(x_i, z_i=j)$$

Often solvable in closed form

Algo:

- Initialize params θ
- Run E step
- Run M step

} Repeat

GMMs

E step: $q_i(z_i=j) = P_\theta(z_i=j \mid X_i) = \dfrac{P_\theta(z_i=j, X_i)}{P_\theta(X_i)}$

$$= \frac{\pi_j N(\mu_j, \Sigma_j, X_i)}{\sum_{l=1}^{k} \pi_l N(\mu_l, \Sigma_l, X_i)}$$

(Compute for all X_i, $j \in \{1, \ldots, k\}$)

M step: (Simplify, 1D, Var=1. $P_\theta(x) = \sum \pi_j N(\mu_j, 1, x)$)

$$\operatorname*{argmax}_\theta \sum_{i=1}^{n} \sum_{j=1}^{K} q_i(z_i=j) \log P_\theta(x_i, z_i=j)$$

$$= \operatorname*{argmax}_\theta \sum_{i=1}^{n} \sum_{j=1}^{K} q_i(z_i=j) \log\left(\pi_j \exp\left(-\frac{(x_i-\mu_j)^2}{2}\right)\right)$$

$$= \operatorname*{argmax}_\theta \sum \sum q_i(z_i=j)\left(\log \pi_j - \frac{(x_i-\mu_j)^2}{2}\right)$$

Focus on μ_j's

$$\operatorname*{argmin}_{\mu_j} \sum \sum q_i(z_i=j) \frac{1}{2}(x_i-\mu_j)^2 \Rightarrow \sum_{i=1}^{n} -q_i(z_i=j)(x_i-\mu_j) = 0$$

$$\mu_j = \frac{\sum_{i=1}^{n} q_i(z_i=j) X_i}{\sum_{i=1}^{n} q_i(z_i=j)}$$

$$\operatorname*{argmax}_{\text{dist } \pi} \sum_{i=1}^{n} \sum_{j=1}^{K} q_i(z_i=j) \log \pi_j \iff \operatorname*{argmax}_{\pi} \sum_{i=1}^{n} \sum_{j=1}^{K} q_i(z_i=j) \log \pi_j + \lambda\left(\sum \pi_j - 1\right)$$

s.t. $\sum_{j=1}^{K} \pi_j = 1$

KKT condition

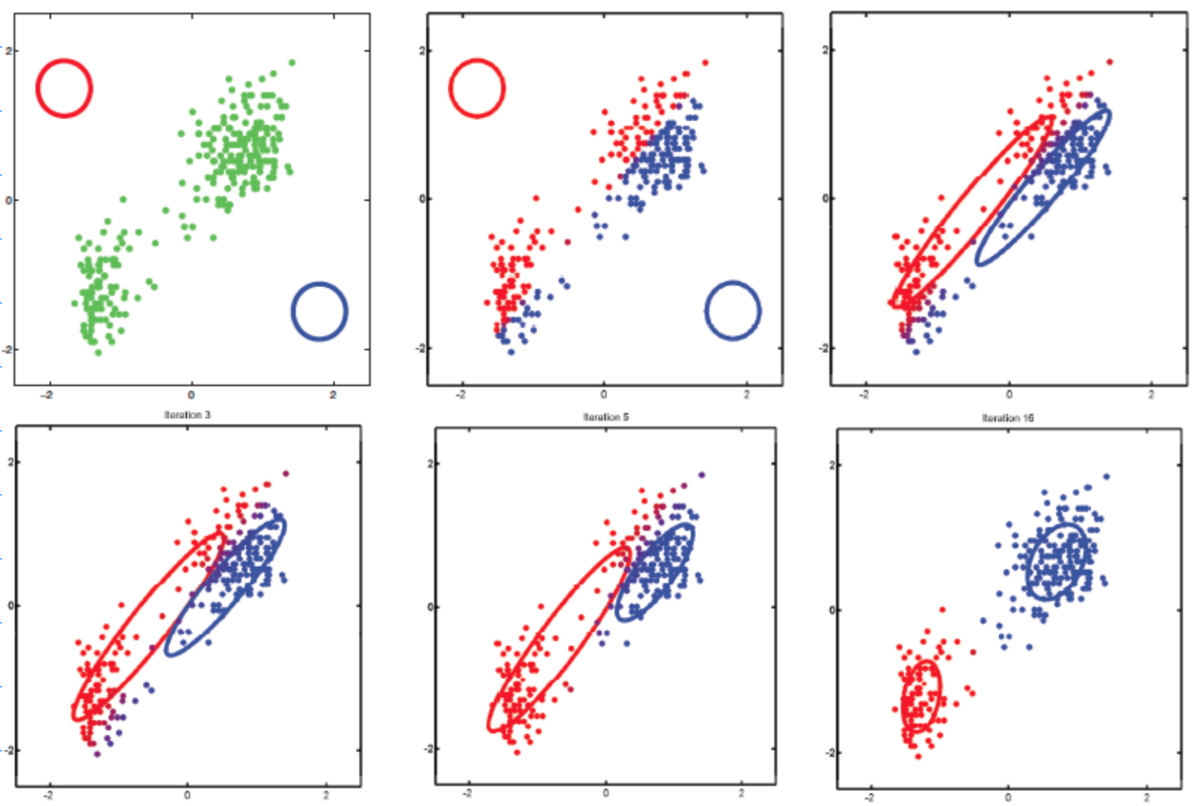
$$\arg\max_{\pi} \sum_{i=1}^{n}\sum_{j=1}^{k} q_i(z_i=j) \log \pi_j + \lambda\left(\sum \pi_j - 1\right)$$

$$\frac{d}{d\pi_j} \hookrightarrow \sum_{i=1}^{n} \frac{q_i(z_i=j)}{\pi_j} + \lambda = 0$$

$$\pi_j = -\frac{1}{\lambda}\sum_{i=1}^{n} q_i(z_i=j)$$

$$\hookrightarrow \sum_{j=1}^{k} \pi_j = -\frac{1}{\lambda} \underbrace{\sum_{i=1}^{n} \underbrace{\sum_{j=1}^{k} q_i(z_i=j)}_{1}}_{n} = -\frac{n}{\lambda} = 1 \Rightarrow \lambda = -n$$

$$\pi_j = \frac{1}{n}\sum_{i=1}^{n} q_i(z_i=j)$$



Distribution