# Machine Learning for Data Science (CS4786)
# Lecture 7

Gaussian Mixture Models

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2017fa/

# K-means Clustering

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

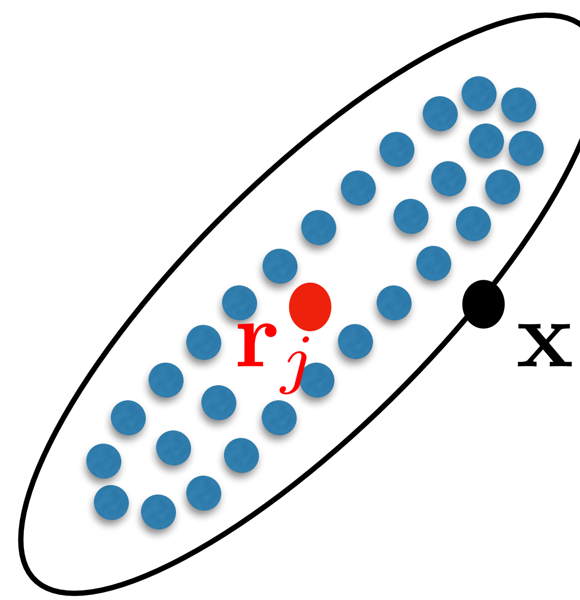
$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \|\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1}\|$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t$$

  3. $m \leftarrow m + 1$

# General Ellipsoid




$$d(\mathbf{x}, C_j) = (\mathbf{x} - \mathbf{r}_j)^\top \Sigma_j^{-1} (\mathbf{x} - \mathbf{r}_j)$$

$$\Sigma_j = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x} - \mathbf{r}_j)(\mathbf{x}_t - \mathbf{r}_j)^\top$$

# Ellipsoidal Clustering

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$ and ellipsoids $\hat{\Sigma}_j^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \ \underbrace{(\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1})^\top \left(\hat{\Sigma}^{m-1}\right)^{-1} (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1})}_{d(\mathbf{x}_t, C_j)}$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t \qquad \hat{\Sigma}^m = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top$$

  3. $m \leftarrow m + 1$

# K-means: pitfalls

- Looks for spherical clusters ✓
- Of same radius ✓
- And with roughly equal number of points ✗

# Mixture Distribution

$$\forall j \leq K,\ \pi(j) \geq 0$$

$$\sum_{j=1}^{K} \pi(j) = 1$$

- Pi models proportion of points belonging to each cluster
- We update Pi as we go
- Finally we expect Pi to contain proportion of points we expect in each cluster

# Towards Hard Gaussian Mixture Model

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$, ellipsoids $\hat{\Sigma}_j^0$ and initial proportions $\pi^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\operatorname{argmin}} \ \underbrace{(\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1})^\top \left(\hat{\Sigma}^{m-1}\right)^{-1} (\mathbf{x}_t - \hat{\mathbf{r}}_j^{m-1}) - \log(\pi_j^{m-1})}_{d(\mathbf{x}_t, C_j)}$$

  2. For each $j \in [K]$, set new representative as

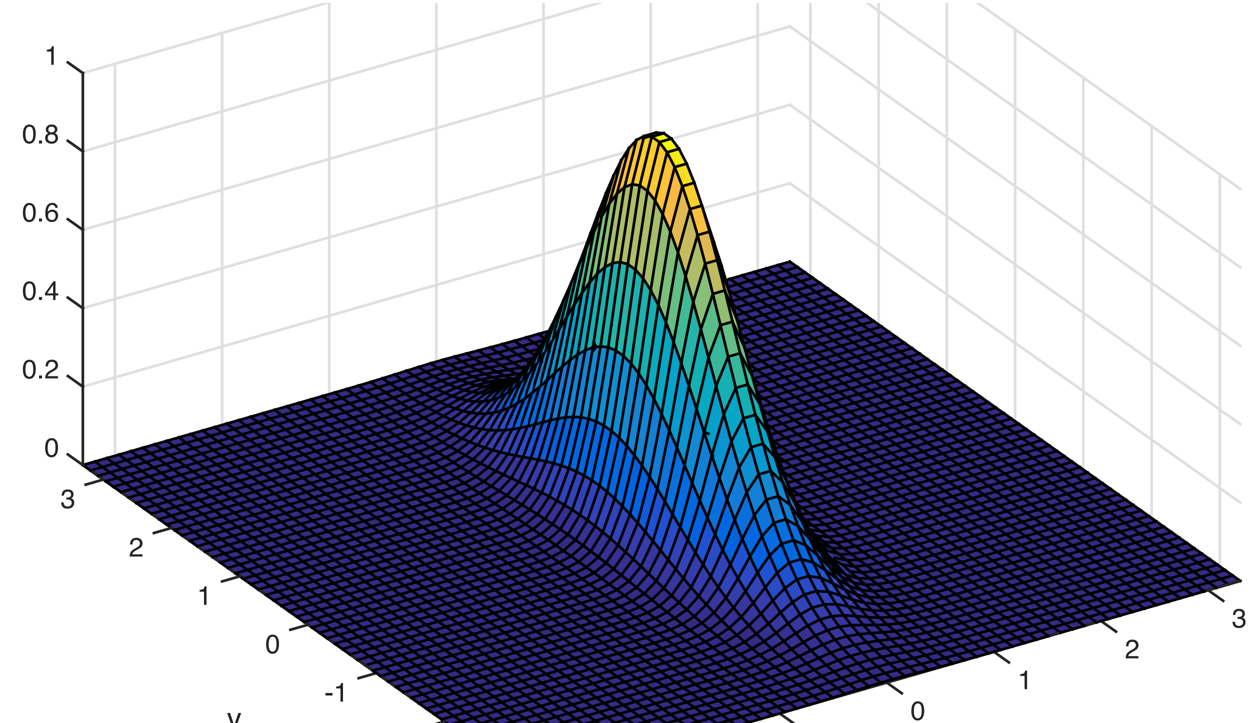
$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t \qquad \hat{\Sigma}^m = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top \qquad \pi_j^m = \frac{|C_j^m|}{n}$$

  3. $m \leftarrow m + 1$

# Multivariate Gaussian

- Two parameters:
  - Mean $\mu \in \mathbb{R}^d$
  - Covariance matrix $\Sigma$ of size dxd

$$p(x;\mu,\Sigma) = (2\pi)^{d/2}\det(\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)$$



# Hard Gaussian Mixture Model

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$, ellipsoids $\hat{\Sigma}_j^0$ and initial proportions $\pi^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$\hat{c}^m(\mathbf{x}_t) = \underset{j \in [K]}{\arg\max}\ p(\mathbf{x}_t, \hat{\mathbf{r}}_j^{m-1}, \hat{\Sigma}_j^{m-1}) \times \pi^m(j)$$

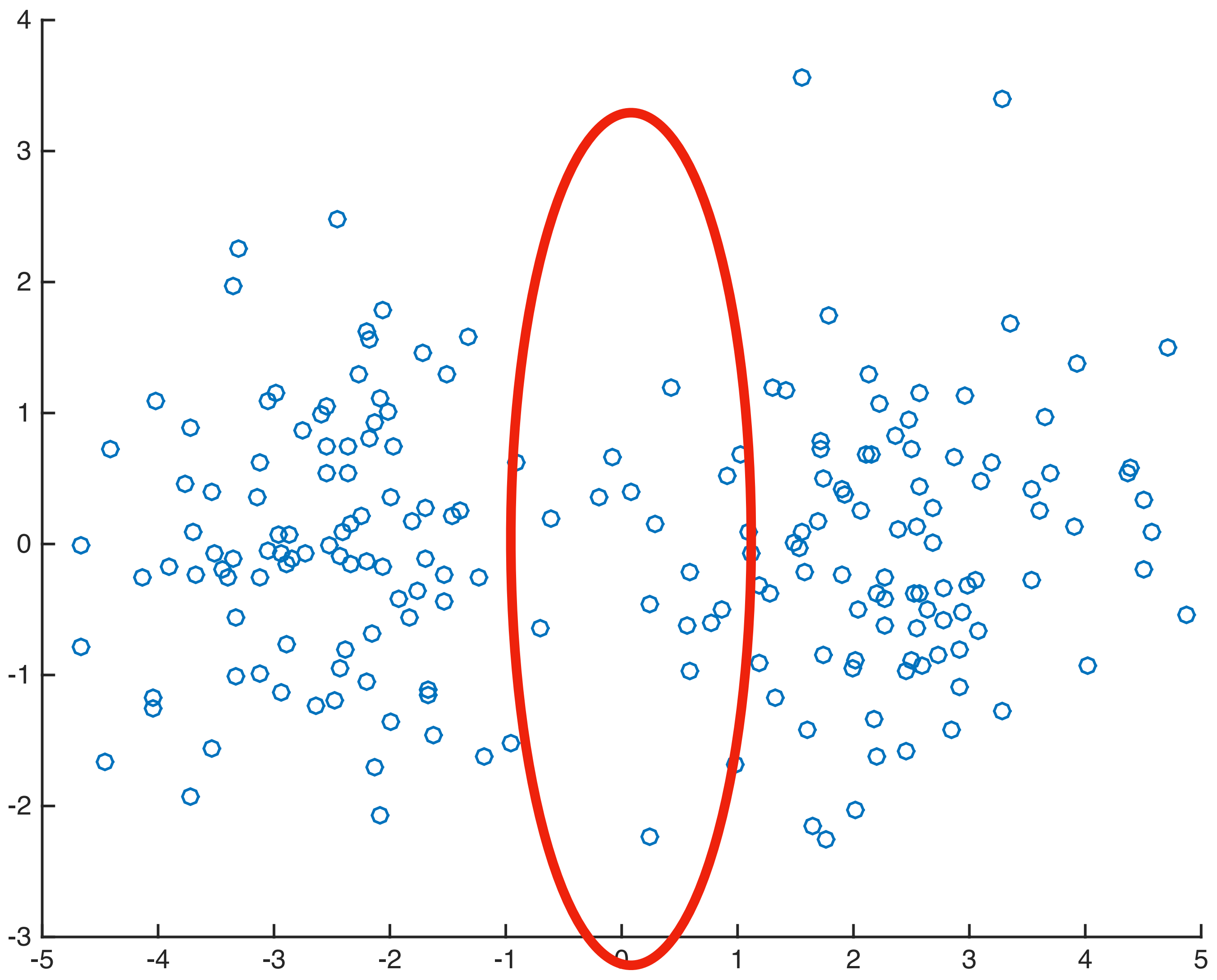
$$d(\mathbf{x}_t, C_j)$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{1}{|\hat{C}_j^m|} \sum_{\mathbf{x}_t \in \hat{C}_j^m} \mathbf{x}_t \qquad \hat{\Sigma}_j^m = \frac{1}{|C_j|} \sum_{t \in C_j} (\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top \qquad \pi_j^m = \frac{|C_j^m|}{n}$$

  3. $m \leftarrow m + 1$

# Pitfall of Hard Assignment



# (Soft) Gaussian Mixture Model

- For all $j \in [K]$, initialize cluster centroids $\hat{\mathbf{r}}_j^0$ and ellipsoids $\hat{\Sigma}_j^0$ randomly and set $m = 1$
- Repeat until convergence (or until patience runs out)
  1. For each $t \in \{1, \ldots, n\}$, set cluster identity of the point

$$Q_t^m(j) = p(\mathbf{x}_t, \hat{\mathbf{r}}_j^{m-1}, \hat{\Sigma}_j^{m-1}) \times \pi^m(j)$$

  2. For each $j \in [K]$, set new representative as

$$\hat{\mathbf{r}}_j^m = \frac{\sum_{t=1}^n Q_t(j)\mathbf{x}_t}{\sum_{t=1}^n Q_t(j)} \qquad \hat{\Sigma}_j^m = \frac{\sum_{t=1}^n Q_t(j)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)(\mathbf{x}_t - \hat{\mathbf{r}}_j^m)^\top}{\sum_{t=1}^n Q_t(j)}$$

$$\pi_j^m = \frac{\sum_{t=1}^n Q_t(j)}{n}$$

  3. $m \leftarrow m + 1$

# Demo

# How to choose K

- Elbow method:
  - plot Objective versus K, typically it monotonically decreases.
  - Pick point where there is a kink (explanation in variance is not as much)
  - Intuition: look at rate of change
- Add to objective penalty + p(K) and minimize, where p increases with K
  - intuition we prefer smaller clusters
  - Use prior knowledge to pick p
  - (AIC, BIC etc can been seen to be specific cases)