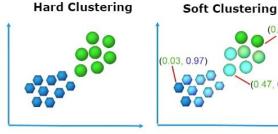


Hard and Soft Clusters

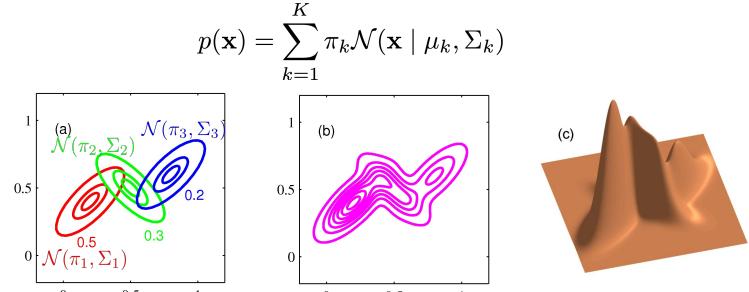
- K-Means uses **hard clustering assignment**.
 - A point belongs to exactly one cluster.
- Mixture of Gaussians uses **soft clustering**.
 - A point could be explained by more than one cluster.
 - Different clusters take different levels of “responsibility” (posterior probability) for that point.
 - (It was actually generated by only one cluster, but we don’t know which one; we assign a probability)



28

Mixtures of Gaussians

- Mixtures of Gaussians make it possible to describe much richer distributions.

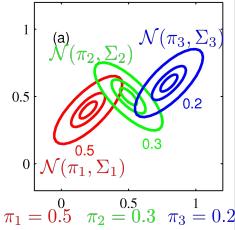


29

Mixtures of Gaussians

- Note the mixing coefficients in

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k) \quad \sum_{k=1}^K \pi_k = 1$$



- Let \mathbf{z} in $\{0,1\}^K$ be a 1-of- K random variable;

$$p(z_k = 1) = \pi_k \quad p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

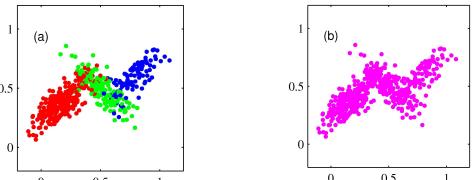
$$p(\mathbf{x} | z_k = 1) = \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x} | \mathbf{z}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)$$

30

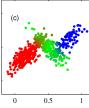
Mixtures of Gaussians

- To generate samples from a Gaussian mixture distribution $p(\mathbf{x})$, use $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) p(\mathbf{x} | \mathbf{z})$:
 - Select a value \mathbf{z} from the marginal $p(\mathbf{z})$;
 - Then select a value \mathbf{x} from $p(\mathbf{x} | \mathbf{z})$ for that \mathbf{z} .



31

EM for Gaussian Mixtures (1-page summary)



- Initialize parameters randomly $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$
- Repeat until convergence (alternating optimization)
 - E Step: Given fixed parameters θ , optimize $q(\mathbf{Z})$: find posterior probabilities of \mathbf{Z} given \mathbf{X} , $q(\mathbf{Z}) = p(\mathbf{Z} | \mathbf{X}, \theta)$
 - M Step: Given fixed $q(\mathbf{Z})$ (or $\gamma(z_{nk})$), optimize θ (estimate mean, covariance, etc.): equivalent to MLE for Gaussian Discriminant Analysis using $q(\mathbf{Z})$ as pseudo-counts!

52

Relation to K-means

- In Gaussian mixture, we fix the covariance matrix for each cluster as $\sigma^2 I$
- We take $\sigma^2 \rightarrow 0$ then
- the update equations converge to doing K-means clustering

EM for Gaussian Mixtures: E-step

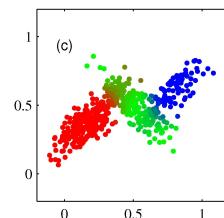
Derivation: in the next lecture!

- Initialize means, covariances, and mixing coefficients for the K Gaussians.

- E Step: Given the parameters, evaluate the responsibilities.

$$q(\mathbf{Z}) = p(\mathbf{Z} | \mathbf{X}, \theta)$$

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)} | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}^{(n)} | \mu_j, \Sigma_j)} = P(z_k = 1 | \mathbf{x}^{(n)})$$

(Recap) E-step: compute posterior → optimal $q(\mathbf{Z})$ 

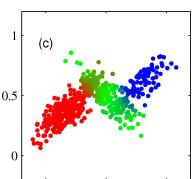
50

Mixtures of Gaussians: E-step

- Need to calculate $p(\mathbf{Z} | \mathbf{X}, \theta)$, i.e., *soft assignments*
- Responsibility is the degree (posterior prob.) to which each Gaussian explains an observation \mathbf{X} .

Q. Verify this! (Hint: Use Bayes Rule)

$$q^{(n)}(\mathbf{z}_k) = p(\mathbf{z}_k | \mathbf{x}^{(n)}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)} | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}^{(n)} | \mu_j, \Sigma_j)} =: \gamma(z_{nk})$$



26

(Hint: Use Bayes Rule)

Mixtures of Gaussians: M-step

General formula for M-step:

$$\arg \max_{\theta} \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}^{(n)} | \theta)$$

Plug in for GMM: $\theta = \{\pi_k, \mu_k, \Sigma_k \mid k \in \{1 \dots K\}\}$

$$\begin{aligned} \operatorname{argmax}_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} & \sum_{n=1}^N \sum_{k=1}^K q^{(n)}(\mathbf{z}_k) \log p(\mathbf{z}_k, \mathbf{x}^{(n)} | \pi_k, \mu_k, \Sigma_k) \\ \text{s.t. } & \sum_{k=1}^K \pi_k = 1 \end{aligned}$$

27

Mixtures of Gaussians: M-step

Let's first simplify the expression $J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\begin{aligned} J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{n=1}^N \sum_{k=1}^K q^{(n)}(\mathbf{z}_k) \log p(\mathbf{z}_k, \mathbf{x}^{(n)} | \pi_k, \mu_k, \Sigma_k) \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \left(\log \pi_k + \log \frac{1}{(2\pi)^{m/2} (\det \Sigma_k)^{1/2}} - \frac{1}{2} (\mathbf{x}^{(n)} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}^{(n)} - \mu_k) \right) \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k - \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \left((2\pi)^{m/2} (\det \Sigma_k)^{1/2} \right) \\ &\quad - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}^{(n)} - \mu_k) \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \det \Sigma_k \\ &\quad - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}^{(n)} - \mu_k) + \text{const} \end{aligned}$$

28

Mixtures of Gaussians: M-step

$$\begin{aligned} J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \det \Sigma_k \\ &\quad - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}^{(n)} - \mu_k) + \text{const} \end{aligned}$$

- Maximize J w.r.t. μ_k by differentiating w.r.t. μ_k and setting the gradient to 0:

$$\begin{aligned} \frac{\partial J}{\partial \mu_k} &= \sum_{n=1}^N \gamma(\mathbf{z}_{nk}) \Sigma_k^{-1} (\mathbf{x}^{(n)} - \mu_k) = 0 \\ \mu_k &= \frac{\sum_{n=1}^N \gamma(\mathbf{z}_{nk}) \mathbf{x}^{(n)}}{\sum_{n=1}^N \gamma(\mathbf{z}_{nk})} \end{aligned}$$

29

Mixtures of Gaussians: M-step

$$\begin{aligned} J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \det \Sigma_k \\ &\quad - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}^{(n)} - \mu_k) + \text{const} \end{aligned}$$

- To find Σ_k , we use change of variables: $\mathbf{M}_k = \Sigma_k^{-1}$

$$\begin{aligned} J(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{M}) &= \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k + \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \det \mathbf{M}_k \\ &\quad - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k)^\top \mathbf{M}_k (\mathbf{x}^{(n)} - \mu_k) + \text{const} \end{aligned}$$

30

Mixtures of Gaussians: M-step

$$\begin{aligned} J(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{M}) &= \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k + \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \det \mathbf{M}_k \\ &\quad - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k)^\top \mathbf{M}_k (\mathbf{x}^{(n)} - \mu_k) + \text{const} \end{aligned}$$

- Maximize J w.r.t. \mathbf{M}_k by differentiating w.r.t. \mathbf{M}_k and setting the gradient to 0: *Note: $\frac{\partial \log |\det \mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^\top \frac{\partial \mathbf{a}^\top \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{a} \mathbf{b}^\top$

$$\frac{\partial J}{\partial \mathbf{M}_k} = \frac{1}{2} \sum_{n=1}^N \gamma(\mathbf{z}_{nk}) \mathbf{M}_k^{-1} - \frac{1}{2} \sum_{n=1}^N \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k) (\mathbf{x}^{(n)} - \mu_k)^\top = 0$$

$$\Sigma_k = \mathbf{M}_k^{-1} = \frac{\sum_{n=1}^N \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k) (\mathbf{x}^{(n)} - \mu_k)^\top}{\sum_{n=1}^N \gamma(\mathbf{z}_{nk})}$$

31

Mixtures of Gaussians: M-step

$$\begin{aligned} J(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \det \Sigma_k \\ &\quad - \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}^{(n)} - \mu_k) + \text{const} \end{aligned}$$

- Finally we need: $\max_{\boldsymbol{\pi}} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k$ s.t. $\sum_{k=1}^K \pi_k = 1$

- Use Lagrange multipliers

$$L(\boldsymbol{\pi}, \alpha) = \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k - \alpha \left(\sum_{k=1}^K \pi_k - 1 \right)$$

32

Mixtures of Gaussians: M-step

- Finally we need: $\max_{\boldsymbol{\pi}} \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k$ s.t. $\sum_{k=1}^K \pi_k = 1$

- Use Lagrange multipliers

$$L(\boldsymbol{\pi}, \alpha) = \sum_{n=1}^N \sum_{k=1}^K \gamma(\mathbf{z}_{nk}) \log \pi_k - \alpha \left(\sum_{k=1}^K \pi_k - 1 \right)$$

- Setting $\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^N \gamma(\mathbf{z}_{nk}) \frac{1}{\pi_k} - \alpha = 0$ gives $\pi_k = \frac{\sum_{n=1}^N \gamma(\mathbf{z}_{nk})}{\alpha}$

- Using the constraint $\sum_{k=1}^K \pi_k = 1$, we get:

$$\pi_k = \frac{\sum_{n=1}^N \gamma(\mathbf{z}_{nk})}{\sum_{k=1}^K \sum_{n=1}^N \gamma(\mathbf{z}_{nk})} = \frac{\sum_{n=1}^N \gamma(\mathbf{z}_{nk})}{N}$$

Mixtures of Gaussians: M-step (putting together)

- The mean of a cluster is the weighted mean, weighted by the responsibilities.

$$\mu_k = \frac{\sum_{n=1}^N \gamma(\mathbf{z}_{nk}) \mathbf{x}^{(n)}}{N_k}$$

[N_k = sum of pseudo-counts γ_{nk} over n.]

- N_k is the effective number of points in cluster k

$$N_k = \sum_{n=1}^N \gamma(\mathbf{z}_{nk})$$

$$\pi_k = \frac{N_k}{N}$$

- Likewise for covariance:

$$\Sigma_k = \frac{\sum_{n=1}^N \gamma(\mathbf{z}_{nk}) (\mathbf{x}^{(n)} - \mu_k) (\mathbf{x}^{(n)} - \mu_k)^\top}{N_k}$$

33

EM for Gaussian Mixtures: Summary

- Initialize means, covariances, and mixing coefficients for the K Gaussians.
- E Step: Given the parameters, evaluate the responsibilities or the posterior

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)} | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}^{(n)} | \mu_j, \Sigma_j)} = P(z_k = 1 | \mathbf{x}^{(n)})$$

35

EM for Gaussian Mixtures: Summary

- M Step: Given the responsibilities, re-evaluate the coefficients.
- $$\pi_k^{\text{new}} = \frac{N_k}{N} = \frac{\sum_n \gamma(z_{nk})}{N}$$
- $$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}^{(n)}$$
- $$\sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}^{(n)} - \mu_k^{\text{new}}) (\mathbf{x}^{(n)} - \mu_k^{\text{new}})^{\top}$$
- Stop when either coefficients or log likelihood converges.

36

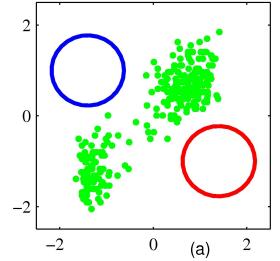
Summary: EM for Gaussian Mixtures

- Initialize parameters randomly $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$
 - Repeat until convergence (alternating optimization)
 - E Step: Given fixed parameters θ , set $q^{(n)}(\mathbf{z}) = p(\mathbf{z} | \mathbf{x}^{(n)}, \theta)$
 - $$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)} | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}^{(n)} | \mu_j, \Sigma_j)} = P(z_k = 1 | \mathbf{x}^{(n)})$$
 - M Step: Given fixed $q(\mathbf{z}^{(n)})$'s for $\mathbf{x}^{(n)}$'s (or $\gamma(z_{nk})$), update θ :
- $$\pi_k^{\text{new}} = \frac{N_k}{N} = \frac{\sum_n \gamma(z_{nk})}{N} \quad \arg \max_{\theta} \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}^{(n)} | \theta)$$
- $$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}^{(n)}$$
- $$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}^{(n)} - \mu_k^{\text{new}}) (\mathbf{x}^{(n)} - \mu_k^{\text{new}})^{\top}$$

37

EM Example

- Initialize parameters: means, covariances, and mixing coefficients.

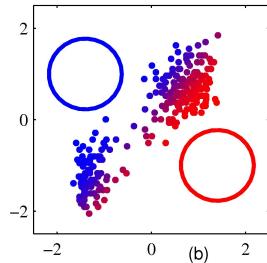


(a)

38

EM Example

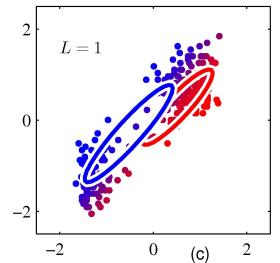
- First E Step



39

EM Example

- First M Step

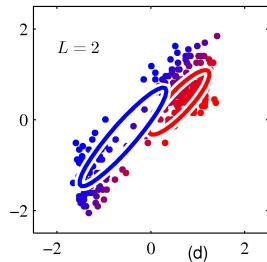


(c)

40

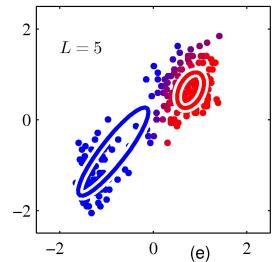
EM Example

- Second E and M Steps



41

- Three more E-M cycles

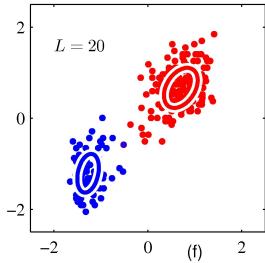


(e)

42

EM Example

- Fifteen E-M cycles later



Summary: Expectation Maximization

Core Idea:

- Iterative method for parameter estimation with hidden data (latent variables).
- Finds maximum likelihood estimates (local optimum) via EM.

Steps: (alternating optimization until convergence)

- **E-Step:** Estimate the posterior probability (expected value) of latent variables.
- **M-Step:** Maximize the “expected” likelihood to update parameters.

Key Points:

- Converges to a local optimum (initialization matters).
- Commonly used in clustering (e.g., Gaussian Mixture Models) and other models with missing data.

43

44

Summary: K-Means vs Gaussian Mixtures

K-means:

- Clusters data by minimizing within-cluster variance.
- Hard assignments; best for spherical, similar-sized clusters.
- Fast, but sensitive to initialization.

Gaussian Mixture Models (GMM):

- Models data as a mix of Gaussian distributions via EM.
- Soft assignments; handles elliptical, overlapping clusters.
- More flexible yet computationally intensive.
- K-means can be shown to be a special case of GMM

Key Considerations:

- Both need pre-specified # clusters and careful initialization.
- Use K-means for speed; choose GMM for probabilistic clustering or probabilistic modeling of data.

45