Clustering and Mixture Models
Outline

- Clustering intuition
- Mixture models
- Mixture of Gaussians
- Expectation maximization
- Variational expectation maximization
Clustering

unsupervised

Lots of variants:
- Hard cluster assignment
- Distribution-based
- Hierarchical, etc.
Mixture Models

\[ X = \{x_1, \ldots, x_n\} \]

\[ P(X) = \prod_{i=1}^{n} \sum_{c_i=1}^{K} p(c_i) p(x_i|c_i) \]

- probability that example \( i \) is in cluster \( c_i \)
- probability of \( x_i \) if \( i \) is in cluster \( c_i \)

**generative** process:
1. Sample cluster
2. Sample data example from cluster distribution
Gaussian Mixture Model

\[ P(x) = \sum_{c=1}^{K} p(c) \frac{1}{\sqrt{2\pi} |\Sigma_c|} \exp \left( -\frac{1}{2} (x - \mu_c)^\top \Sigma_c^{-1} (x - \mu_c) \right) \]

- multinomial cluster membership
- multivariate Gaussian data \( \mathcal{N}(x|\mu_c, \Sigma_c) \)
“clouds” can overlap
no identity for clusters
Expectation Maximization Recipe

Input: $x_i \quad i \in \{1, \ldots, n\}$

GMM parameters:
\[
p(c) \quad \mu_c \quad \Sigma_c \quad c \in \{1, \ldots, K\}
\]

Latent variables:
\[
z_i \in \{1, \ldots, K\}
\]

Latent variable probabilities: $p(z_i)$
\[
\sum_{c=1}^{K} p(c) = \sum_{c=1}^{K} p(z_i = c) = 1
\]

E-step: fit latent variable probabilities
\[
p(z_i = c) \leftarrow \frac{p(c) \mathcal{N}(x_i | \mu_c, \Sigma_c)}{\sum_{c'=1}^{K} p(c') \mathcal{N}(x_i | \mu_{c'}, \Sigma_{c'})}
\]

M-step: fit GMM parameters using expected likelihood
\[
p(c) \leftarrow \frac{1}{n} \sum_{i=1}^{n} p(z_i = c)
\]
\[
\mu_c \leftarrow \frac{\sum_{i=1}^{n} p(z_i = c) x_i}{\sum_{i=1}^{n} p(z_i = c)}
\]
\[
\Sigma_c \leftarrow \frac{\sum_{i=1}^{n} p(z_i = c) (x_i - \mu_c) (x_i - \mu_c)^\top}{\sum_{i=1}^{n} p(z_i = c)}
\]
Non-Convexity of GMM NLL

no worse solution?

better solution?
EM as Maximizing Lower Bound

\[
\mathbb{E}_z [\log p(X|\mu, \Sigma, \theta, Z)]
\]

\[
\log p(X|\mu, \Sigma, \theta)
\]
EM as Maximizing Lower Bound
Initialization

• Some heuristics:
  • Completely random
  • Fit a single Gaussian to all data; randomly perturb K copies
  • Randomly initialize cluster memberships. Start with M-step
EM Likelihood Landscape

equally bad local maxima

equally good global maxima

bad local minima

equally bad local maxima

Wrong
Global maximum?
\[ \sum_{c=1}^{K} p(c) \frac{1}{\sqrt{2\pi|\Sigma_c|}} \exp \left( -\frac{1}{2} (x_i - \mu_c)^\top \Sigma_c^{-1} (x_i - \mu_c) \right) \]

\[ \Sigma_c \to [0] \quad L \to \infty \]
\[ \int_{-\infty}^{\infty} p(x) \, dx = 1 \]
\[ \int_{-\infty}^{\infty} p(x) \, dx = 1 \]
\[ \int_{-\infty}^{\infty} p(x) \, dx = 1 \]
Fixes

- Good initialization
- Constrain covariance to have some bandwidth on each dimension
Summary of EM for GMMs

- Gaussian mixture models: fit data with weighted combination of Gaussians
- Non-convex likelihood
  - Estimate probability of each point being in each Gaussian
  - Use probabilities to maximize expected likelihood
  - Iterate until local minimum
Variational Derivation of EM
Marginal Likelihood

\[ p(X|\theta) = \int_Z p(X, Z|\theta) dZ \]

\[ \sum_Z p(X, Z|\theta) \]

e.g.,

\[ X = \{x_1, \ldots, x_n\} \]

\[ \theta = \{\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \ldots p(c)\} \]

\[ Z = \{z_1, \ldots, z_n\} \quad \text{(cluster memberships)} \]

\[ p(X, Z|\theta) = \prod_{i=1}^{n} p(z_i) \mathcal{N}(x_i|\mu_{z_i}, \Sigma_{z_i}) \]

\[ \log p(X|\theta) = \log \sum_Z p(X, Z|\theta) \]

log marginal likelihood

\[ \argmax_\theta \log \sum_Z p(X, Z|\theta) \]

learning objective
Jensen’s Inequality

For any convex function $\varphi$, 

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$$

For any concave function $\phi$, 

$$\phi(\mathbb{E}[X]) \geq \mathbb{E}[\phi(X)]$$

$$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log X]$$
Variational Bound

\[
\log(\mathbb{E}[X]) \geq \mathbb{E}[\log X]
\]

\[
\log \sum_Z p(X, Z|\theta) = \log \sum_Z \frac{q(Z)}{q(Z)} p(X, Z|\theta) \quad \sum_Z q(Z) = 1
\]

\[
= \log \sum_Z q(Z) \frac{p(X, Z|\theta)}{q(Z)}
\]

\[
\geq \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)}
\]

\[
= \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)
\]
**Variational Bound**

\[
\log \sum_Z p(X, Z|\theta) \geq \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)
\]

We can pick any \( q \) distribution and the bound holds

\[
\arg\max_{\theta, q \in Q} \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)
\]

\[
q(Z) = \prod_{i=1}^n q(z_i) \quad \sum q(z_i) = 1
\]
Fully Factorized Variational Family

\[
\argmax_{\theta, q \in Q} \sum_Z q(Z) \log p(X, Z|\theta) - \sum_Z q(Z) \log q(Z)
\]

\[
q(Z) = \prod_{i=1}^n q(z_i) \quad \sum_{z_i} q(z_i) = 1
\]

\[
\argmax_{\theta, q \in Q} \sum_{i=1}^n \sum_{z_i} q(z_i) \log p(x_i, z_i|\theta) - q(z_i) \log q(z_i)
\]
Point Distributions

\[
\arg\max_{\theta, q \in Q} \sum_Z q(Z) \log p(X, Z | \theta) - \sum_Z q(Z) \log q(Z)
\]

\[
q(Z) = \prod_{i=1}^n q(z_i)
\]

\[
q(z_i) = \begin{cases} 
1 & \text{if } z_i = \hat{z}_i \\
0 & \text{otherwise}
\end{cases}
\]

\[
\arg\max_{\theta, q \in Q} \sum_{i=1}^n \sum_{z_i} q(z_i) \log p(x_i, z_i | \theta) - q(z_i) \log q(z_i)
\]

\[
\arg\max_{\theta, q \in Q, \hat{Z}} \sum_{i=1}^n \log p(x_i, \hat{z}_i | \theta)
\]

Point distributions are often easier to compute, but less robust.
Point Distributions for GMMs

\[
\argmax_{\theta,q\in Q,\hat{Z}} \sum_{i=1}^{n} \log p(x_i, \hat{Z}_i|\theta) \\
\sum_{i=1}^{n} \log \mathcal{N}(x_i|\mu_{\hat{Z}_i}, \Sigma_{\hat{Z}_i})
\]

\[
\hat{Z}_i \leftarrow \argmax_{z} \log \mathcal{N}(x_i|\mu_z, \Sigma_z)
\]

\[
\mu_z \leftarrow \frac{\sum_{i;\hat{Z}_i=z} x_i}{\sum_{i;\hat{Z}_i=z} 1} \\
\Sigma_z \leftarrow \frac{\sum_{i;\hat{Z}_i=z} (x_i - \mu_i)(x_i - \mu_i)^\top}{\sum_{i;\hat{Z}_i=z} 1}
\]

K-means

\[
\hat{Z}_i \leftarrow \argmin_{z} \|x_i - \mu_z\|
\]

assign points to closest mean

\[
\mu_z \leftarrow \frac{\sum_{i;\hat{Z}_i=z} x_i}{\sum_{i;\hat{Z}_i=z} 1}
\]

set means to average of points in cluster
Example

input data
Example

assign points to initialized means
Example

average each cluster
Example

repeat for all clusters
Example

repeat for all clusters
Example

end of iteration 1
Example

end of iteration 2
Example

end of iteration 3
Example

end of iteration 4
Example

end of iteration 5
Example

end of iteration 6
Example

end of iteration 7
Example

converged at iteration 8
Summary of Variational EM

• Used Jensen’s inequality to derive lower bound on log marginal likelihood

• Bound uses variational distribution $q$. We get to choose what family of $q$ distributions to consider

• Using fully-factorized multinomial distributions for $q$ gets EM

• Fully-factorized point distributions gets “hard”-EM, and using fixed, spherical covariance gets K-means