Model-Based Clustering

- Same setting and goal as center-based clustering
- Statistical model of clustering

Given a sequence $f_1(x) \ldots f_K(x)$ of probability densities / distribution and probabilities $p_1 \ldots p_K \geq 0$ $p_1 + \ldots + p_K = 1$, suppose $\vec{x}_1 \ldots \vec{x}_N$ were generated in the following

- Assign cluster label, Z_i

  Set $Z_i = z_i$ w/ probability p_i
  $Z_i \in \{1, \ldots k\}$

- Draw $\vec{x}_i$ conditional on $Z_i = z_i$

$$X_i \sim f_{Z_i}(x)$$

- Repeat independently for $i = 1, \ldots N$

Equivalently, we could draw X_i from the mixture distribution

$$f_{mix}(x) \quad \sum_{j=1}^{K} p_j f_j(x)$$

Ex: Gaussian mixture model

$$f_{mix}(x \mid p_1 \ldots p_K, \vec{\mu}_1 \ldots \vec{\mu}_K, \Sigma_1 \ldots \Sigma_K)$$
$$= \sum_{j=1}^{K} \frac{p_j}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}-\vec{\mu}_j)\right)$$

Ex: Can use mixture of any type of distribution or combinations!

t, t & Gaussian, ... Cauchy

MAIN IDEA

FOR FIXED K AND DENSITIES $f_1 \ldots f_K$ DEFINED IN TERMS OF PARAMETERS ϕ

1) FIND THE MLE, $\hat{\phi}$, OF ϕ TO IDENTIFY THE BEST MODEL FOR MIXTURE DENSITY (EXPECTATION-MAXIMIZATION)

2) USE $f_{MIX}(x \mid \hat{\phi})$ TO ASSIGN CLUSTER LABELS

CAN REPEAT FOR DIFFERENT CHOICES OF K (IF SUITABLE) AND COMPARE RESULTS (AIC, BIC)

__NOTE__: THE CHOICE OF DENSITIES $f_1 \ldots f_K$ AND SIMPLIFYING ASSUMPTIONS ON PARAMETERS CAN GREATLY INFLUENCE THE PERFORMANCE OF MODEL-BASED CLUSTERING. FOR THIS LESSON, WE WILL FOCUS ON MIXTURES OF GAUSSIANS (ALTHOUGH THIS CHOICE STRUGGLES W/ NON-CONVEX CLUSTERS).

- $f_{MIX}(x \mid \phi)$ $\quad \phi = (p_1 \ldots p_K, \vec{\mu}_1 \ldots \vec{\mu}_K, \Sigma_1 \ldots \Sigma_K)$

$$= \sum_{j=1}^{K} p_j \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2} (\vec{x}-\vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}-\mu_j)\right)$$

- LIKELIHOOD OF DATA

$$L(\phi \mid \vec{x}_1 \ldots \vec{x}_N) = \prod_{i=1}^{N} f_{MIX}(x_i \mid \phi)$$

WE DON'T OBSERVE Z_i's

$$= \prod_{i=1}^{N} \left(\sum_{j=1}^{K} p_j \frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2} (\vec{x}-\mu_j)^T \Sigma_j^{-1} (\vec{x}-\mu_j)\right) \right)$$

- LOG-LIKELIHOOD OF DATA

$$\log L(\phi \mid \vec{x}_1 \ldots \vec{x}_N) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{K} \ldots \right)$$

CHOOSING K BASED ON AIC / BIC

FOR NOW, ASSUME WE CAN FIND $\hat{\phi} = (\hat{p}_1 \dots \hat{p}_K, \hat{\vec{\mu}}_1 \dots \hat{\vec{\mu}}_K, \hat{\Sigma}_1 \dots \hat{\Sigma}_K)$

MAXIMIZING, i.e.

$$\hat{L} = \max_{\phi} L(\phi \mid \vec{x}_1 \dots \vec{x}_n) = L(\hat{\phi} \mid \vec{x}_1 \dots \vec{x}_n) \qquad \log \hat{L} = \widehat{\log L}$$

$$\widehat{\log L} \quad \max_{\phi} \log L(\phi \mid \vec{x}_1 \dots \vec{x}_n)$$

$$= \log L(\hat{\phi} \mid \vec{x}_1 \dots \vec{x}_n)$$

IMPORTANTLY, $\hat{L}$ GIVES AN ASSESSMENT OF HOW WELL THE MODEL FITS THE DATA,

- LARGER $\hat{L}$ INDICATES A BETTER FIT
- TYPICALLY $\hat{L}$ CAN BE INCREASED BY ADDING MORE COMPONENTS

  * INCREASING K RISKS OVERFITTING !! (POOR GENERALIZABILITY)

  * PENALIZE MODELS W/ TOO MANY COMPONENTS (PARAMETERS)

AKAIKE INFORMATION CRITERION (AIC)

TO COMPARE MODELS (DIFFERENT CHOICES OF k), WE CAN COMPUTE AIC SCORES

$$AIC = -2 \log \hat{L} + 2p, \quad p = \# \text{ PARAMETERS}$$

WHERE LOW AIC IS BETTER.

PENALTY FOR OVERPARAMETERIZATION

(OTHERS DEFINE

$AIC = 2 \log \hat{L} - 2p$, SO HIGH AIC IS BETTER)

Bayes Information Criterion

Similar to AIC, but also incorporates <u>sample size</u>

$$BIC = -2\log \hat{L} + p \log N \quad , \quad p = \#\text{ parameters}$$

Again, small BIC is better

(In mclust package $BIC = 2\log\hat{L} - p\log N$
so we want maximal BIC)

<u>Note</u>: Both AIC & BIC are related to asymptotic normality of MLE $\hat{\phi}$ as $N \longrightarrow \infty$

- typically need N^d samples
- other assumptions required which do not hold w/o additional restrictions on $\hat{\phi}$
- relies on an approximation to Bayes factors using BIC

$$\frac{P(x_1 \ldots x_N \mid K \text{ clusters})}{P(x_1 \ldots x_N \mid K+1 \text{ clusters})} = \frac{P(K \text{ clusters} \mid x_1 \ldots x_N)\,P(K+1 \text{ clusters})}{P(K+1 \text{ clusters} \mid x_1 \ldots x_N)\,P(K \text{ clusters})}$$

requires integrating over parameters (with suitable priors)

<u>Ex</u>: Gaussian mixture in $\mathbb{R}^p$ w/ K components

- $(K-1)$ parameters for $p_1 \ldots p_K$
- Kp parameters for $\vec{\mu}_1 \ldots \vec{\mu}_K$
- $\frac{Kp(p+1)}{2}$ params. for $\Sigma_1 \ldots \Sigma_K$ if no restriction on $\Sigma_1 \ldots \Sigma_K$

  1 parameter for $\Sigma_1 \ldots \Sigma_K$ if we assume $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_K = \sigma^2 I$

<u>Ex</u>: NCI Data on Canvas > Files > Numerical Example > Clustering

ASSIGNING CLUSTER LABELS

ONCE $f_{mix}(x \mid \hat{\phi})$ HAS BEEN CHOSEN, ASSIGN x_i TO CLUSTER j FOR WHICH

$$\hat{p}_j f_j(x_i \mid \hat{\vec{\mu}}_j \, \hat{\Sigma}_j)$$

IS MAXIMIZED, i.e.

$$z_i = \underset{j=1,\dots K}{\operatorname{argmax}} \; p_j \, f_j(x_i \mid \hat{\mu}_j \, \hat{\Sigma}_j)$$

FINDING $\hat{L}$: THE EXPECTATION MAXIMIZATION ALGORITHM

IF WE KNEW ORIGINAL CLUSTER LABELS $z_1 \dots z_N$ THEN

$$L(\phi \mid x_1 \dots x_N, z_1 \dots z_N) = \prod_{i=1}^{N} \sum_{j=1}^{N} \mathbb{1}(z_i = j) \, f_j(\vec{x}_i \mid \vec{\mu}_j \, \Sigma_j)$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{K} \left[p_j f_j(x_i, \mu_j \, \Sigma_j) \right]^{\mathbb{1}(z_i = j)}$$

$$= \exp\left(\sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{1}(z_i = j) \left[\log(p_j) - \tfrac{1}{2} \log |\Sigma_j| - \tfrac{d}{2} \log(2\pi) - \tfrac{1}{2} (\vec{x}_i - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}_i - \mu_j) \right] \right)$$

IDEA BEHIND EM IS TO

E-STEP: ESTIMATE $P(z_i = j \mid \vec{x}_1 \dots \vec{x}_N, \phi)$ AND CONSTRUCT $E_{z \mid x, \phi}\left[\log L(\phi \mid x_1 \dots x_N, z_1 \dots z_N)\right]$

M-STEP: MAXIMIZE $E_{z \mid x, \phi}(\log(\phi \mid x, z))$

THEN ITERATE TO CONVERGENCE (TO A LOCAL OPTIMA)

Ex: Mixture of Two Gaussians

$$\phi = (p_1, p_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2)$$

$$\log L(\phi \mid x_1 \ldots x_N, z_1 \ldots z_N) = \sum_{i=1}^{N} \sum_{j=1}^{2} \mathbb{1}(z_i = j)\left[\log p_j - \tfrac{1}{2}\log|\Sigma_j| - \tfrac{1}{2}(\vec{x}_i - \vec{\mu}_j)^T \Sigma_j^{-1}(\vec{x}_i - \vec{\mu}_j) - \tfrac{d}{2}\log(2\pi)\right]$$

1) Initialize values for ϕ $(p_1, p_2 \neq 0)$

2) E-Step: Let $\phi^{(t)}$ be our current estimate for ϕ. Set

$$T_{i,j}^{(t)} = P\left(Z_i = j \mid \vec{x}_i, \phi^{(t)}\right) = \frac{p_j^{(t)} f_j(x_i, \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{j=1}^{2} p_j^{(t)} f_j(\vec{x}_i, \mu_j^{(t)} \Sigma_j^{(t)})} \propto p_j^{(t)} f_j(x_i, \mu_j^{(t)}, \Sigma_j^{(t)})$$

And let

$$E_{Z \mid x_1 \ldots x_N, \phi^{(t)}}\left[\log(\phi \mid x_1 \ldots x_N, Z_1 \ldots Z_N)\right]$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{2} T_{i,j}^{(t)}\left[\log p_j - \tfrac{1}{2}\log|\Sigma_j| - \tfrac{1}{2}(\vec{x}_i - \vec{\mu}_j)\Sigma_j^{-1}(\vec{x}_i - \mu_j) - \tfrac{d}{2}\log(2\pi)\right]$$

$$= Q(\phi \mid \phi^{(t)})$$

3) M-Step: maximize $Q(\phi \mid \phi^{(t)})$ over all ϕ. Closed form expressions in this case

$$p_j^{(t+1)} = \frac{\sum_{i=1}^{N} T_{ij}^{(t)}}{\sum_{i=1}^{N}\sum_{j=1}^{N} T_{ij}^{(t)}} \propto \sum_{i=1}^{N} T_{ij}^{(t)}$$

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{N} T_{ij}^{(t)} \vec{x}_i}{\sum_{i=1}^{N} T_{ij}^{(t)}}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{N} T_{ij}^{(t)} (\vec{x}_i - \hat{\mu}_j^{(t+1)})(\vec{x}_i - \hat{\mu}_j^{(t+1)})^T}{\sum_{i=1}^{N} T_{ij}^{(t)}}$$

4) Repeat 2-3 until convergence to a local optimum is reached.

Note: Restricting $\Sigma_1 = \Sigma_2 = \sigma^2 I$ and sending $\sigma^2 \to 0$ (known) has interesting connection to K-means clustering. See pset #7 problem 4.