Mixture models

Ref: Bishop's book
Chapter 9

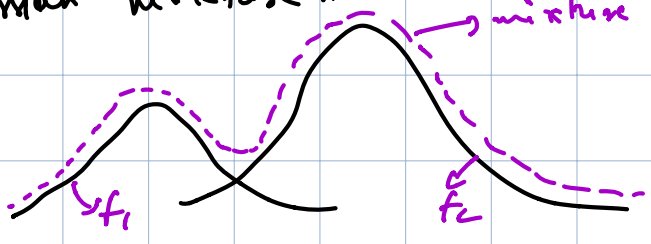
In practice, class conditional densities may not be in a standard form. E.g. multimodal densities.

Mixture model: $f(x) = \sum_{j=1}^{K} \lambda_j f_j(x)$,

$\lambda_j \geq 0$, $\sum_{j=1}^{K} \lambda_j = 1$

f_j, $j = 1 \cdots K$ is a valid density

e.g. Gaussian mixture model (GMM)



$f_i \sim N(\mu_i, \sigma_i^2)$

GMMs are useful to model multimodal densities

$$f(x|\theta) = \sum_{j=1}^{K} \lambda_j f_j(x), \quad f_j \sim N(\mu_j, \sigma_j^2)$$

$$\theta = (\lambda_j, j=1\cdots K, \mu_j, \sigma_j^2, j=1\cdots K)$$

→ this needs to be estimated from iid samples

Max-likelihood approach:

Dataset $D = \{x_1, \cdots x_n\}$ iid from f

$$L(\theta) = \prod_{i=1}^{n} \left(\sum_{j=1}^{K} \lambda_j f_j(x_i) \right)$$

$$\ell(\theta) = \sum_{i=1}^{n} \log \underbrace{\left(\sum_{j=1}^{k} \lambda_j f_j(x_i) \right)}$$

Sum inside a log

$\Rightarrow$ hard to optimize $\ell(\theta)$ even if f_j's are Gaussian

For ease of presentation, we focus on the case $k=2$

$$\theta_1 = (\mu_1, \sigma_1^2), \quad \theta_2 = (\mu_2, \sigma_2^2)$$

$$\phi(x|\theta_j) = \frac{1}{\sqrt{2\pi}\,\sigma_j} \exp\left(- \frac{(x-\mu_j)^2}{2\sigma_j^2} \right) \quad j = 1, 2$$

$$f(x|\theta) = \lambda_1 \phi(x|\theta_1) + \lambda_2 \phi(x|\theta_2)$$

$$\theta = (\underbrace{\theta_1, \theta_2}, \underbrace{\lambda_1, \lambda_2})$$

component density parameters — mixture co-efficients

$$\ell(\theta) = \sum_{i=1}^{n} \log \left(\lambda_1 \phi(x_i|\theta_1) + \lambda_2 \phi(x_i|\theta_2) \right)$$

Goal: $\max_{\theta} \ell(\theta)$

$$\frac{\partial \phi(x|\theta_j)}{\partial \mu_j} = \phi(x|\theta_j) \frac{(x-\mu_j)}{\sigma_j^2}$$

$$\frac{\partial \phi(x|\theta_j)}{\partial \sigma_j} = \phi(x|\theta_j)\left(\frac{(x-\mu_j)^2}{\sigma_j^3} - \frac{1}{\sigma_j}\right)$$

$$\frac{\partial l(\theta)}{\partial \mu_j} = \sum_{i=1}^{n}\left(\frac{\lambda_j \phi(x_i|\theta_j)}{\lambda_1\phi(x_i|\theta_1)+\lambda_2\phi(x_i|\theta_2)}\right)\left(\frac{x_i-\mu_j}{\sigma_j^2}\right)$$

letting $\alpha_{ij} = \left(\frac{\lambda_j \phi(x_i|\theta_j)}{\lambda_1\phi(x_i|\theta_1)+\lambda_2\phi(x_i|\theta_2)}\right)$, we have

$$\frac{\partial l(\theta)}{\partial \mu_j} = \sum_{i=1}^{n} \alpha_{ij}\left(\frac{x_i-\mu_j}{\sigma_j^2}\right)$$

$$\frac{\partial l(\theta)}{\partial \sigma_j} = \sum_{i=1}^{n} \alpha_{ij}\left(\frac{x_i-\mu_j}{\sigma_j^3} - \frac{1}{\sigma_j}\right)$$

ML-estimates satisfy

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n}\alpha_{ij}x_i}{\sum_{i=1}^{n}\alpha_{ij}}, \quad \hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n}\alpha_{ij}(x_i-\mu_j)^2}{\sum_{i=1}^{n}\alpha_{ij}},$$

$j = 1, 2$

Maximizing $l(\theta)$ w.r.t. λ_1 & λ_2:

$$\max_{\theta} \; l(\theta) \quad \text{s.t.} \quad \lambda_1 + \lambda_2 = 1$$

Lagrangian: $g(\theta, \eta) = \ell(\theta) + \eta(\lambda_1 + \lambda_2 - 1)$

↑ Lagrange multiplier

$$\nabla_{\lambda_1} g(\theta, \eta) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \frac{\phi(x_i|\theta_1)}{\lambda_1 \phi(x_i|\theta_1) + \lambda_2 \phi(x_i|\theta_2)} + \eta = 0$$

or, equivalently, $\sum_{i=1}^{n} \frac{\alpha_{i1}}{\lambda_1} + \eta = 0 \quad \text{— (1)}$

Similarly, $\sum_{i=1}^{n} \frac{\alpha_{i2}}{\lambda_2} + \eta = 0 \quad \text{— (2)}$

Summing (1) & (2), we have

$$\sum_{i=1}^{n} \alpha_{i1} + \sum_{i=1}^{n} \alpha_{i2} + \eta(\lambda_1 + \lambda_2) = 0$$

$$\eta = -\left(\sum_{i=1}^{n} \alpha_{i1} + \sum_{i=1}^{n} \alpha_{i2} \right)$$

$$= -n$$

So, ML estimate for λ_j satisfy

$$\hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \alpha_{ij}$$

So, ML estimation for θ involves solving the following system of equations:

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \alpha_{ij} x_i}{\sum_{i=1}^{n} \alpha_{ij}}, \quad \hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} \alpha_{ij} (x_i - \mu_j)^2}{\sum_{i=1}^{n} \alpha_{ij}}$$

$$\hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \alpha_{ij}$$

An iterative scheme for solving the above set of equations:-

$$\hat{\mu}_j^{(k+1)} = \frac{\sum_{i=1}^{n} \alpha_{ij}^{(k)} x_i}{\sum_{i=1}^{n} \alpha_{ij}^{(k)}}, \quad \hat{\sigma}_j^{2\,(k+1)} = \frac{\sum_{i=1}^{n} \alpha_{ij}^{(k)} (x_i - \mu_j^{(k)})^2}{\sum_{i=1}^{n} \alpha_{ij}^{(k)}}$$

$$\hat{\lambda}_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \alpha_{ij}$$

$$\alpha_{ij}^{(k+1)} = \frac{\lambda_j^{(k+1)} \phi(x_i | \theta_j^{(k+1)})}{\lambda_1^{(k+1)} \phi(x_i | \theta_1^{(k+1)}) + \lambda_2 \phi(x_i | \theta_2^{(k+1)})}$$

Prelude to EM algorithm:

$$f(x|\theta) = \lambda_1 \phi(x|\theta_1) + \lambda_2 \phi(x|\theta_2)$$

$x_i \sim f(x|\theta)$ can be generated as follows:

(i) $$z = \begin{cases} 1 & \text{w.p. } \lambda_1 \\ 2 & \text{w.p. } \lambda_2 \end{cases}$$

(ii) Generate x_i from $\phi(x|\theta_2)$

Dataset $D = \{x_i, i=1--n\}$

If we are given component information for each x_i, then the "parameter estimation" task is easy

Let z_{ij}, $i=1--n$, $j=1,2$ denote component information

$z_{ij}=1$ if $x_i \sim \phi(\cdot|\theta_j)$

$P(z_{ij}=1) = \lambda_j$, $i=1--n$, $j=1,2$

$f(x_i|z_{ij}=1) = \phi(x_i|\theta_j)$

$z_i = (z_{i1}, z_{i2})$

Complete dataset $\{(x_i, z_i), i=1--n\}$

Note: z_i is "missing" information

However, given z_i, the ML-estimates are

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} z_{i1} x_i}{\sum_{i=1}^{n} z_{i1}}, \quad \hat{\mu}_2 = \frac{\sum_{i=1}^{n} z_{i2} x_i}{\sum_{i=1}^{n} z_{i2}}$$

The ML-estimates for σ_j^2 can be worked out in a similar fashion.

EM algorithm:

Iterative procedure for ML estimation works with a regular dataset, i.e., Z_i (missing information) is not provided

$$D = \{ x_i, i=1\text{--}n \} \qquad \bar{D} = \{ x_i, z_i, i=1\text{--}n \}$$

$f(x, z|\theta) \rightarrow$ density for $\bar{D}$

$$\bar{l}(\theta) = \underbrace{\log\left(\prod_{i=1}^{n} f(x_i, z_i|\theta) \right)}$$

let $\log(f(\mathbf{x}, \mathbf{z})|\theta)$

denote $\log\left(\prod_{i=1}^{n} f(x_i, z_i|\theta) \right)$

$\mathbf{x} = (x_1 \text{--} x_n)$, $\mathbf{z} = (z_1 \text{--} z_n)$

EM algorithm:-

Expectation step (E-step):

$$Q(\theta, \theta^{(k)}) = E_{\mathbf{z}|\mathbf{x}, \theta^{(k)}} \log(f(\mathbf{x}, \mathbf{z}|\theta))$$

expectation of complete data log likelihood w.r.t. conditional distribution of missing data given incomplete data x & iterate $\theta^{(k)}$.

Maximization step (M-step):

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta, \theta^{(k)})$$

Compute next value of θ i.e., $\theta^{(k+1)}$ by maximizing the $Q(\cdot,\cdot)$ found in E-step.

Example:

GMMs with two components

$$f(x|\theta) = \lambda_1 \phi(x|\theta_1) + \lambda_2 \phi(x|\theta_2)$$

Recall $P(z_{ij}=1) = \lambda_j$, $f(x_i|z_{ij}=1) = \phi(x_i|\theta_j)$

$$z_i = (z_{i1}, z_{i2})$$

$$f(z_i|\theta) = (\lambda_1)^{z_{i1}} (\lambda_2)^{z_{i2}} = \prod_{j=1}^{2} (\lambda_j)^{z_{ij}}$$

$$f(x_i|z_i,\theta) = [\phi(x_i|\theta_1)]^{z_{i1}} [\phi(x_i|\theta_2)]^{z_{i2}}$$

$$f(x_i, z_i|\theta) = f(x_i|z_i,\theta) f(z_i|\theta)$$

$$= \prod_{j=1}^{2} \left[\lambda_j \phi(x_i|\theta_j)\right]^{z_{ij}}$$

$$f(\mathbf{x}, \mathbf{z}|\theta) = \prod_{i=1}^{n} \prod_{j=1}^{2} \left[\lambda_j \phi(x_i|\theta_j)\right]^{z_{ij}}$$

$$\log f(\mathbf{x}, \mathbf{z}|\theta) = \sum_{i=1}^{n} \sum_{j=1}^{2} z_{ij} \log\left(\lambda_j \phi(x_i|\theta_j)\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{2} z_{ij} \left(\log \lambda_j + \log\left(\phi(x_i|\theta_j)\right)\right)$$

Sum of log $\Rightarrow$ ML estimation is easier

↑ we got sum of log because we are given the missing info z_i in this case

E-step: $Q(\theta, \theta^{(k)}) = E_{\mathbf{z}|\mathbf{x},\theta^{(k)}} \log\left(f(\mathbf{x}, \mathbf{z}|\theta)\right)$

PTO

$$Q(\theta, \theta^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{2} E\left[z_{ij} \mid x, \theta^{(k)}\right] \log\left(\lambda_j \phi(x_i \mid \theta_j)\right)$$

$$E(z_{ij} \mid x, \theta) = P(z_{ij} = 1 \mid x, \theta)$$

$$= \frac{f(x_i \mid z_{ij}=1, \theta)\, P(z_{ij}=1)}{\sum_{k=1}^{2} f(x_i \mid z_{ik}=1, \theta)\, P(z_{ik}=1)}$$

$$= \frac{\lambda_j \phi(x_i \mid \theta_j)}{\sum_{k=1}^{2} \lambda_k \phi(x_i \mid \theta_k)} = \alpha_{ij}(\theta)$$

So,

$$Q(\theta, \theta^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{2} \alpha_{ij}(\theta^{(k)}) \log\left(\lambda_j \phi(x_i \mid \theta_j)\right)$$

M-step:

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta, \theta^{(k)})$$

$$Q(\theta, \theta^{(k)}) = \sum_{i=1}^{n} \sum_{j=1}^{2} \alpha_{ij}(\theta^{(k)}) \left[\log \lambda_j - \log(\sigma_j \sqrt{2\pi}) - \frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right]$$

Need to maximize the above w.r.t. θ

$$\frac{\partial Q}{\partial \mu_1} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_{i1}(\theta^{(k)}) \left(\frac{x_i - \mu_1}{\sigma_1^2}\right) = 0$$

$$\text{So, } \mu_1^{(k+1)} = \frac{\sum_{i=1}^{n} \alpha_{i1}(\theta^{k})\, x_i}{\sum_{i=1}^{n} \alpha_{i1}(\theta^{(k)})} \quad \text{— (*)}$$

$$\frac{\partial Q}{\partial \sigma_1} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_{i1}(\theta^{(k)}) \left(-\frac{1}{\sigma_1} + \frac{(x_i - \mu_1)^2}{\sigma_1^3}\right) = 0$$

$$\text{So, } (\sigma_1^2)^{(k+1)} = \frac{\sum_{i=1}^{n} \alpha_{i1}(\theta^{(k)}) (x_i - \mu_1^{(k)})^2}{\sum_{i=1}^{n} \alpha_{i1}(\theta^{(k)})} \quad \text{— (**)}$$

(*) and (**) are the same as that obtained earlier by a direct max-likelihood approach.

$$\alpha_{ij}^{(k+1)} = \frac{\lambda_j^{(k+1)} \phi(x_i \mid \theta_j^{(k+1)})}{\sum_{l=1}^{2} \lambda_l^{(k+1)} \phi(x_i \mid \theta_l^{(k+1)})} = \alpha_{ij}(\theta^{(k+1)})$$

H.W.

(1) Derive the EM algorithm for Bernoulli mixtures

(2) Work out the EM algorithm for GMMs with

two components, both having the same variance

K-means clustering

Given: $\{x_1, \ldots x_n\}$ $x_i \in \mathbb{R}^d$

Goal: Partition these points into K clusters, where K is given.

Let μ_j, $j=1,\ldots K$ be the cluster centers.

To restate the goal:

Find cluster centers μ_j, $j=1\ldots K$ such that the sum of square distances of each data point to closest μ_j is minimum

Notation:- $z_{ij} \in \{0,1\}$ $i=1\ldots n$, $j=1\ldots K$

z_{ij} = indicator that takes value 1 if x_i is mapped to cluster j

Objective $$J = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \|x_i - \mu_j\|^2$$

Need a clustering scheme that minimizes J, i.e.,

$$\min_{z_{ij}, \mu_j} J$$
$i=1\ldots n$
$j=1\ldots K$

K-means clustering is an iterative scheme with the following steps:

Step I: Fix $\mu_j, \forall j$ and $\min_{z_{ij}, \forall i,j} J$ } Fix cluster-heads & find the assignment of points to clusters

Step II: Fix $z_{ij}, \forall i,j$ and $\min_{\mu_j, \forall j} J$ } Fix the assignment (z_{ij}) & find the cluster-heads

Repeat Steps I & II until no change in the assignment, (or) after a large # of steps.

How to perform Step I:

J = linear function of z_{ij}

So, each data point can be optimized separately since J is in an additive form.

For point x_i, minimize $\sum_{j=1}^{K} z_{ij} \|x_i - \mu_j\|^2$

Solution: $z_{ij} = \begin{cases} 1 & \text{if } j = \arg\min_k \|x_i - \mu_k\|^2 \\ 0 & \text{else} \end{cases}$

PTO

Now to perform Step II: "z_{ij}'s are fixed"

J: quadratic in μ_j

$$\nabla_{\mu_j} J = 0 \Rightarrow 2 \sum_{i=1}^{n} z_{ij} (x_i - \mu_j) = 0$$

leading to $$\mu_j = \frac{\sum_{i=1}^{n} z_{ij} x_i}{\sum_{i=1}^{n} z_{ij}}$$

μ_j: average of all points in cluster j

$\sum_{i=1}^{n} z_{ij}$: # of points in cluster j

Note: Each round (= step I & II) of K-means clustering reduces J & hence K-means converges.

Connection of K-means to EM for GMMs:

GMM with a common covariance matrix ϵI for each component, where $\epsilon > 0$, I is identity matrix.

components = K.

$$\phi(x | \mu_j, \Sigma_j) = \frac{1}{(2\pi\epsilon)^{d/2}} \exp\left(-\frac{1}{2\epsilon} \|x - \mu_j\|^2\right)$$

$\Sigma_j = \epsilon I \ \forall j$

$$\gamma_{ij} = \frac{\lambda_j \exp\left(-\|x_i - \mu_j\|^2/2\epsilon\right)}{\sum_{k=1}^{K} \lambda_k \exp\left(-\|x_i - \mu_k\|^2/2\epsilon\right)}$$

As $\epsilon \to 0$,

$$\gamma_{ij} = \begin{cases} 1 & \text{if } j = \arg\min_k \|x_i - \mu_k\|^2 \\ 0 & \text{else} \end{cases}$$

↳ the assignment that one would obtain from K-means clustering.

Cluster head → averages of points assigned to a particular cluster.

γ_{ij} → hard assignment.

So, GMMs with common covariance ϵI would become the K-means clustering algorithm in the limit as $\epsilon \to 0$.

H.W.! Check the log likelihood of complete data becomes the sum of squares (J) in K-means i.e., as $\epsilon \to 0$