Mixture models and EM algorithm

(Unsupervised Learning Part II)

Cluster analysis partitions data sets into homogeneous subgroups, by finding "cutoff" linear boundaries between the resulting subgroups. However, often there is inherent uncertainty in the allocation to subgroups.

A more sophisticated method of classification is using mixture models, which can be viewed as soft K-means clustering.

Similar to cluster analysis, mixture analysis assumes that the data come from two or more homogeneous sub-populations, but it is not known which populations each observation was from. In mixture analysis,

- Each cluster is represented by a probability distribution, usually a parametric model, such as $N_p(\mu_i, \Sigma_i)$.
- The data set is represented as a mixture of cluster-specific component distributions.
- The unknown cluster membership is treated as a random variable.

A simple example

The simplest mixture is the case of two binary variables.
– Y is a variable of interest. First consider the case when Y is a binary variable.
– C denotes sub-populations or classes. Here $C = 1$ or 2.
The joint distribution of C and Y can be displayed in a 2×2 table.

	$Y=1$	$Y=2$	*row margin*
$C=1$	$\mathbb{P}(C=1, Y=1)$	$\mathbb{P}(C=1, Y=2)$	$\mathbb{P}(C=1)=p_1$
$C=2$	$\mathbb{P}(C=2, Y=1)$	$\mathbb{P}(C=2, Y=2)$	$\mathbb{P}(C=2)=p_2$
column margin	$\mathbb{P}(Y=1)$	$\mathbb{P}(Y=2)$	1

For any possible value of Y,

$$\begin{aligned}\mathbb{P}(Y=y) &= \mathbb{P}(C=1, Y=y) + \mathbb{P}(C=2, Y=y)\\ &= \mathbb{P}(C=1)\,\mathbb{P}(Y=y \mid C=1) + \mathbb{P}(C=2)\,\mathbb{P}(Y=y \mid C=2)\end{aligned}$$

From another point of view, for any observation of y, the observation is from class $C = k$ with probability $\mathbb{P}(C = k)$. In this example, the index of class $k = 1, 2$.

A more general example

Now consider a slightly more general case of a discrete random variable Y with integer outcomes.

	$Y=1$	$\cdots$	$Y=j$	$\cdots$	*row margin*
$C=1$	$P(C=1, Y=1)$	$\cdots$	$P(C=1, Y=j)$	$\cdots$	$P(C=1)=p_1$
$C=2$	$P(C=2, Y=1)$	$\cdots$	$P(C=2, Y=j)$	$\cdots$	$P(C=2)=p_2$
column margin	$P(Y=1)$	$\cdots$	$P(Y=j)$	$\cdots$	1

The column margin, the marginal probability of Y, can be written as

$$\mathbb{P}(Y=y) = \mathbb{P}(C=1)\,\mathbb{P}(Y=y \mid C=1) + \mathbb{P}(C=2)\,\mathbb{P}(Y=y \mid C=2)$$

The conditional distributions

$$P_i(y) = \mathbb{P}(Y=y \mid C=i), \qquad i = 1, 2,$$

are probability distributions with $\sum_y P_i(y) = 1$. The probability distribution of Y can be expressed as

$$\mathbb{P}(Y=y) = p_1 P_1(y) + p_2 P_2(y) = \begin{cases} P_1(y), & \text{with probability } p_1\\ P_2(y), & \text{with probability } p_2\end{cases}$$

with

$$p_1 + p_2 = 1, \qquad where \quad p_i = \mathbb{P}(C=i), \quad i = 1, 2.$$

The probability distribution of Y is expressed as a mixture of two probability distributions. Y has probability p_1 to be in population 1, with probability $p_2 = 1 - p_1$ to be in population 2.

If Y is a continuous random variable, probability density function is used in the place of probability mass function.

$$f_Y(y) = p_1 f_1(y) + p_2 f_2(y) = \begin{cases} f_1(y), & \text{with probability } p_1,\\ f_2(y), & \text{with probability } p_2,\end{cases} \qquad p_1 + p_2 = 1.$$

With a mixture distribution, an observation of Y has probability p_1 to be from population 1, with probability $p_2 = 1 - p_1$ to be from population 2.

Definition of finite mixture distributions

Suppose that random p-vector X has a probability density function of the form

$$f(\boldsymbol{x}) = p_1 f_1(\boldsymbol{x}) + \cdots + p_k f_k(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^p$$

where $f_i(\boldsymbol{x})$'s are probability density functions on $\mathbb{R}^p$,

$$f_i \neq f_j, \qquad for \quad i \neq j,\ i, j = 1, \cdots k,$$

and

$$p_1 + \cdots + p_k = 1, \qquad p_i > 0 \quad for \quad i = 1, \cdots k.$$

Then X is said to have a **finite mixture distribution** with k components.

The probability p_i's are **mixture proportions** or mixture weights, the density function f_i's are **mixture components**.

Parametric and Gaussian mixture distribution

Typically the approach of mixture analysis involves parametric models.

$$f(\boldsymbol{x}) = f(\boldsymbol{x}; \theta) = p_1 f_1(\boldsymbol{x}; \theta_1) + \cdots + p_k f_k(\boldsymbol{x}; \theta_k), \qquad \theta = (\theta_1, \cdots, \theta_k).$$

The most popular mixture model is the **finite normal mixture** model or finite **Gaussian mixture model**,

$$f(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i, i = 1, \cdots, k) = p_1 \phi_1(\boldsymbol{x}) + \cdots + p_k \phi_k(\boldsymbol{x}),, \qquad \boldsymbol{x} \in \mathbb{R}^p$$

where

$$\phi_i(\boldsymbol{x}) = \phi_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

is the density function of a p-variate normal distribution $N_p(\boldsymbol{\mu}_i, \Sigma_i)$, $\quad i = 1, \cdots, k$.

Remarks and warnings on mixture models

- Any skewed sampling distribution, even slightly, can fit to several small mixtures towards the tail, even if the distribution actually is not a mixture.

- Before fitting a mixture model, there should be scientific reasons to believe that the underlying distribution is indeed a mixture,
- A typical or common situation: there are two obvious humps in the density function, but a three-component mixture models fits much better than an obvious two-component mixture distribution. Likely the third component is small but connects the two larger components well.
- Mixture of normal components with covariance $\Sigma_i = \eta\Sigma$ is approximately the same as K-means.

1 Likelihood equation and MLE for mixture models

Review likelihood function and MLE

Recall that if $y_1, \cdots, y_n$ are independent observations of a random variable Y with probability density or mass function $f_Y(y, \theta)$, then the joint probability density or mass function of $(y_1, \cdots, y_n)$ is the product

$$f_Y(y_1, \theta) f_Y(y_2, \theta) \cdots f_Y(y_n, \theta) = \prod_{j=1}^{n} f_Y(y_j, \theta)$$

Given observed data value $y_1, \cdots, y_n$, the joint density is evaluated at $y_1, \cdots, y_n$, while parameter θ is unknown. Then the joint density given data is a function of unknown parameters, called the likelihood function.

$$L(\theta | y_1, \cdots, y_n) = \prod_{j=1}^{n} f_Y(y_j, \theta)$$

The logarithm of the likelihood function is

$$\ell(\theta | y_1, \cdots, y_n) = log\left(L(\theta | y_1, \cdots, y_n)\right) = \log\left(\prod_{j=1}^{n} f_Y(y_j, \theta)\right) = \sum_{j=1}^{n} \log f_Y(y_j, \theta)$$

Often log-likelihood has simpler form and better mathematical properties.

In data analysis, one of the main objectives is to estimate the true value of θ, which leads to the knowledge of the probability density (or probability) mass function f_Y.

The maximum likelihood estimator (MLE) of the unknown parameter, $\hat{\theta} = \hat{\theta}_{MLE}$, is the θ value that maximizes the likelihood function, given observed data. A common expression is

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta \mid y_1, \cdots, y_n)$$

which means

$$L(\hat{\theta}_{MLE} \mid y_1, \cdots, y_n) = \max_{\theta} L(\theta \mid y_1, \cdots, y_n)$$

Since the θ value maximizing likelihood function is the same as the θ value maximizing the logarithm of the likelihood function, we may write

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \ \log\left(L(\theta \mid y_1, \cdots, y_n)\right) = \arg\max_{\theta} \ell(\theta \mid y_1, \cdots, y_n)$$

Likelihood function of mixture distribution

Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be n independent observations of p-variate random vector $\boldsymbol{X} = (X_1, \cdots, X_p)$ of mixture density

$$f(\boldsymbol{x}) = p_1 f_1(\boldsymbol{x}) + \cdots + p_k f_k(\boldsymbol{x}), \quad \sum_{m=1}^{k} p_m = 1$$

The likelihood function of the n observations is

$$\prod_{j=1}^{n} f(\boldsymbol{x}_j) = \prod_{j=1}^{n} \left[p_1 f_1(\boldsymbol{x}_j) + \cdots + p_k f_k(\boldsymbol{x}_j)\right]$$

The logarithm of the likelihood function of the n observations can be written as

$$\ell(p_1, \cdots, p_k \mid \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) = \log \prod_{j=1}^{n} \left(p_1 f_1(\boldsymbol{x}_j) + \cdots + p_k f_k(\boldsymbol{x}_j)\right) = \sum_{j=1}^{n} \log\left(p_1 f_1(\boldsymbol{x}_j) + \cdots + p_k f_k(\boldsymbol{x}_j)\right)$$

The likelihood function above is written as a function of the mixture proportion parameters $p_1, \cdots, p_k$. In practice, most likely each mixture component f_i depends on some unknown parameter θ_i, which can be a real number or a vector or or a combination of vectors and matrices. Consequently often a mixture likelihood is a function of many parameters.

$$\ell = \ell(p_1, \cdots, p_k \mid \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) = \ell(p_1, \cdots, p_k; \theta_1, \cdots, \theta_k \mid \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$$

MLE of mixture proportions

To investigate the conditions that maximum likelihood estimators satisfy, first consider a simplified case that f_i's are known, so that they do not depend on any unknown parameters. In other words, first let's treat the parameters of f_i's as constants, by considering ℓ as a function of the p_m's, $\ell = \ell(p_1, \cdots, p_k)$.

We are to find an expression for the maximum likelihood estimator of the mixture proportion p_m, which can be viewed as the (unknown) **prior probability** of component m before observations.

To derive the MLEs of the p_m's for $\ell = \ell(p_1, \cdots, p_k)$ with the probability constraint

$$\sum_{m=1}^{k} p_m = 1$$

consider the method of Lagrange Multipliers. The optimization of deriving the MLEs is equivalent to finding the maximum of the Lagrangian

$$L = \ell(p_1, \cdots, p_k) - \lambda\left(\sum_{m=1}^{k} p_m - 1\right)$$

The MLEs of the p_m's occur at critical points of $L = L(p_1, \cdots, p_k, \lambda)$.

Take derivatives of the Lagrangian L with respect to the mixture weight parameters $p_1, \cdots, p_k$ and set them to zero,

$$\frac{\partial L}{\partial p_m} = \sum_{j=1}^{n} \frac{f_m(\boldsymbol{x}_j)}{\sum_{i=1}^{k} p_i f_i(\boldsymbol{x}_j)} - \lambda = 0, \qquad m = 1, \cdots, k. \tag{1}$$

Denote the **posterior probability** of $\boldsymbol{x}_j$ belonging to component m after observation $\boldsymbol{x}_j$ as

$$\pi_{mj} = \pi_{m|\boldsymbol{x}_j} = \frac{p_m f_m(\boldsymbol{x}_j)}{f(\boldsymbol{x}_j)} = \frac{p_m f_m(\boldsymbol{x}_j)}{\sum_{i=1}^{k} p_i f_i(\boldsymbol{x}_j)}. \tag{2}$$

which is also called responsibility of component m given observation $\boldsymbol{x}_j$.

For any observation $j = 1, \cdots, n$, the posterior probabilities must sum to 1,

$$\sum_{m=1}^{k} \pi_{mj} = 1.$$

To derive the value of λ, take a weighted sum of the k equations in (1), which is a sum of zeros, thus $= 0$.

$$\begin{aligned}
0 = \sum_{m=1}^{k} p_m \frac{\partial L}{\partial p_m} &= \sum_{m=1}^{k} \left(\sum_{j=1}^{n} \frac{p_m f_m(\boldsymbol{x}_j)}{\sum_{i=1}^{k} p_i f_i(\boldsymbol{x}_j)} \right) - \sum_{m=1}^{k} (p_m \lambda) \\
&= \sum_{j=1}^{n} \left(\sum_{m=1}^{k} \frac{p_m f_m(\boldsymbol{x}_j)}{\sum_{i=1}^{k} p_i f_i(\boldsymbol{x}_j)} \right) - \left(\sum_{m=1}^{k} p_m \right) \lambda \\
&= \sum_{j=1}^{n} \left(\sum_{m=1}^{k} \pi_{mj} \right) - 1 \times \lambda \\
&= \left(\sum_{j=1}^{n} 1 \right) - \lambda = n - \lambda = 0 \qquad \Rightarrow \quad \lambda = n.
\end{aligned}$$

To see the relationship between prior and posterior probabilities, bring $\lambda = n$ back to (1). For each $m = 1, \cdots, k$,

$$p_m \frac{\partial L}{\partial p_m} = \sum_{j=1}^{n} \pi_{mj} - p_m n = 0$$

We see that the maximum likelihood estimate of the mixture proportion parameter p_m satisfies

$$p_m = \frac{1}{n} \sum_{i=1}^{n} \pi_{mi}, \quad m = 1, \cdots, k. \tag{3}$$

Equation (3) shows that the maximum likelihood estimator of the **prior probability** for the mth mixture component is the **average of the posterior probabilities of all n observations** for this component.

From (1) we obtain the MLE equations

$$\sum_{j=1}^{n} \frac{f_m(\boldsymbol{x}_j)}{\sum_{i=1}^{k} p_i f_i(\boldsymbol{x}_j)} = n, \qquad m = 1, \cdots, k, \quad \sum_{m=1}^{k} p_m = 1. \tag{4}$$

To summarize, Equations (2), (3) and (4) yield the relationship between the MLEs of the priors and the posteriors,

$$\begin{cases}
p_m = \dfrac{1}{n} \displaystyle\sum_{j=1}^{n} \pi_{mj}, & \sum_{m=1}^{k} p_m = 1 \\[2ex]
\pi_{mj} = \dfrac{p_m f_m(\boldsymbol{x}_j)}{\sum_{i=1}^{k} p_i f_i(\boldsymbol{x}_j)}, & \sum_{m=1}^{k} \pi_{mj} = 1
\end{cases}$$

These equations form a complicated system. Numerical procedures have to be employed to solve for the p_m's.

In general there is no guarantee that there exists a solution for (4) such that all p_m are in the interval $[0, 1]$.

In addition, usually $f_m(\boldsymbol{x}) = f_m(\boldsymbol{x}; \theta)$ also depends on unknown parameters of interest. This makes the optimization procedure of deriving the MLE's even more challenging.

The most popular numerical procedure to solve for the MLEs is the EM algorithm and its variations.

2 EM algorithm for two-component Gaussian mixture

The Expectation-Maximization (EM) algorithm is an elegant and powerful method to simplify difficult maximum likelihood problems such as the one in (4), and to obtain parameter estimations for p_m's as well as for θ's in $f_m(x; \theta_m)$. EM has been used effectively in maximum likelihood estimation in various situations.

The EM algorithm approximates the ML estimates of the parameters **iteratively**.

Example

We illustrate the EM algorithm with an example of two-component ($k = 2$) one-dimensional Gaussian mixture,

$$f(x) = p_1 \phi_1(x; \mu_1, \sigma_1^2) + p_2 \phi_2(x; \mu_2, \sigma_2^2), \qquad p_1 + p_2 = 1.$$

where $\phi_i(x; \mu_i, \sigma_i^2)$ is the density of normal distribution of mean μ_i and variance σ_i^2.

Given n observations $x_1, \cdots, x_n$, the log-likelihood is a function of unknown parameters $p_1, \mu_1, \sigma_1, \mu_2, \sigma_2$ ($p_2 = 1 - p_1$).

$$\ell(p_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \mid x_1, \cdots, x_n) = \sum_{j=1}^{n} \log \left[p_1 \phi_1(x_j; \mu_1, \sigma_1^2) + p_2 \phi_2(x_j; \mu_2, \sigma_2^2) \right]$$

The objective is to estimate the MLEs of parameters $p_1, \mu_1, \sigma_1, \mu_2, \sigma_2$.

EM algorithm proceeds to approximate the MLEs iteratively.

- **Initialization**

  Take initial guesses for distribution parameters $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2$, and mixture proportions $\hat{p}_1, \hat{p}_2 = 1 - \hat{p}_1$.

- **Expectation Step**

  Compute the posterior membership probability (also "responsibilities") $\hat{\pi}_{1j}$ for each observation $j = 1, \cdots, n$. Using (2), we have

  $$\pi_{mj} = \pi_{m|x_j} = \frac{p_m f_m(x_j)}{\sum_{i=1}^{2} p_i f_i(x_j)} = \frac{p_m \phi_m(x_j; \mu_m, \sigma_m^2)}{\sum_{i=1}^{2} p_i \phi_i(x_j; \mu_i, \sigma_i^2)}, \qquad m = 1, 2.$$

  this posterior probability is a conditional expectation.

  Plug in the initial guess of $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{p}_1, \hat{p}_2$, we obtain current estimates of the posterior probabilities.

  $$\hat{\pi}_{1j} = \frac{\hat{p}_1 \phi_1(x_j; \hat{\mu}_1, \hat{\sigma}_1^2)}{\hat{p}_1 \phi_1(x_j; \hat{\mu}_1, \hat{\sigma}_1^2) + \hat{p}_2 \phi_2(x_j; \hat{\mu}_2, \hat{\sigma}_2^2)}, \qquad \hat{\pi}_{2j} = 1 - \hat{\pi}_{1j} \qquad j = 1, \cdots, n.$$

- **Maximization Step**

  Treating the current estimates of the posteriors (*a.k.a.* posterior probabilities, responsibilities) $\hat{\pi}_{ij}$ as known, this step maximize the likelihood function to produce a set of updated estimators $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2$ and $\hat{p}_1, \hat{p}_2$.

  The updated means and variances of the two components can be expressed as weighted sums:

$$\hat{\mu}_1 = \sum_{i=1}^{n} \frac{\hat{\pi}_{1i} x_i}{\sum_{j=1}^{n} \hat{\pi}_{1j}}, \qquad \hat{\sigma}_1^2 = \sum_{i=1}^{n} \frac{\hat{\pi}_{1i}(x_i - \hat{\mu}_1)^2}{\sum_{j=1}^{n} \hat{\pi}_{1j}}$$

$$\hat{\mu}_2 = \sum_{i=1}^{n} \frac{\hat{\pi}_{2i} x_i}{\sum_{j=1}^{n} \hat{\pi}_{2j}}, \qquad \hat{\sigma}_2^2 = \sum_{i=1}^{n} \frac{\hat{\pi}_{2i}(x_i - \hat{\mu}_2)^2}{\sum_{j=1}^{n} \hat{\pi}_{2j}}$$

The mixing proportions $\hat{p}_1$ and $\hat{p}_2$ can be expressed as

$$\hat{p}_1 = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_{1i}, \qquad \hat{p}_2 = 1 - \hat{p}_1.$$

These are the maximum likelihood estimations. The derivations are shown in the remarks below.

- Now we have a new set of parameter estimates $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{p}_1, \hat{p}_2$, likely different from their initial guesses.
- These updated parameter values can be used to compute new estimates of the posteriors $\hat{\pi}_{1j}, \hat{\pi}_{2j}$ as in the Expectation Step.
- The updated $\hat{\pi}_{1j}, \hat{\pi}_{2j}$ lead to the next round new estimates of the parameters $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{p}_1, \hat{p}_2$ in the Maximization Step.
- Iterate these steps until convergence, that is, the differences of parameter estimates between two consecutive iterations are small enough to be within desired tolerance bound.

Remarks on derivation details

The MLEs (maximum likelihood estimators) in the Maximization Step can be derived from partial derivatives of the likelihood function

$$\ell = \ell(p_1, \mu_1, \mu_2, \sigma_i^2, \sigma_2^2 \mid x_1, \cdots, x_n) = \sum_{j=1}^{n} \log\left[p_1\phi_1(x_j; \mu_1, \sigma_1^2) + p_2\phi_2(x_j; \mu_2, \sigma_2^2)\right]$$

where

$$\phi_i(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

- For the mean μ_i,

$$\begin{aligned}
\frac{\partial \ell}{\partial \mu_1} &= \sum_{j=1}^{n} \frac{p_1}{p_1\phi_1(x_j; \mu_1, \sigma_1^2) + p_2\phi_2(x_j; \mu_2, \sigma_2^2)} \cdot \frac{\partial}{\partial \mu_1}\phi_1(x_j; \mu_1, \sigma_1^2) \\
&= \sum_{j=1}^{n} \frac{p_1}{p_1\phi_1(x_j; \mu_1, \sigma_1^2) + p_2\phi_2(x_j; \mu_2, \sigma_2^2)} \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x_j-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{2(x_j - \mu_1)}{2\sigma_1^2} \\
&= \sum_{j=1}^{n} \frac{p_1\phi_1(x_j; \mu_1, \sigma_1^2)}{p_1\phi_1(x_j; \mu_1, \sigma_1^2) + p_2\phi_2(x_j; \mu_2, \sigma_2^2)} \cdot \frac{x_j - \mu_1}{\sigma_1^2} \\
&= \sum_{j=1}^{n} \pi_{1j} \frac{x_j - \mu_1}{\sigma_1^2}
\end{aligned}$$

where

$$\pi_{1j} = \frac{p_1\phi_1(x_j; \mu_1, \sigma_1^2)}{p_1\phi_1(x_j; \mu_1, \sigma_1^2) + p_2\phi_2(x_j; \mu_2, \sigma_2^2)}$$

is a function of the density parameters (including μ_1).

In the maximization step of the EM algorithm, π_{1j}'s current estimate $\hat{\pi}_{1j}$ (based on previous-step estimate of μ_i, σ_i^2, p_i) is treated as known. Thus by setting

$$\left.\frac{\partial \ell}{\partial \mu_1}\right|_{\hat{\pi}_{1j}, j=1,\cdots,n} = \sum_{j=1}^{n} \hat{\pi}_{1j} \frac{x_j - \mu_1}{\sigma_1^2} = 0$$

we obtain the updated estimate

$$\hat{\mu}_1 = \sum_{i=1}^{n} \frac{\hat{\pi}_{1i}}{\sum_{j=1}^{n} \hat{\pi}_{1j}} x_i$$

Analogous derivation yields

$$\hat{\mu}_2 = \sum_{i=1}^{n} \frac{\hat{\pi}_{2i}}{\sum_{j=1}^{n} \hat{\pi}_{2j}} x_i$$

- For the variance σ_i^2 (note that $\sigma_i^2 = \theta$ is treated as a parameter, not the square of a parameter),

$$\begin{aligned}
\frac{\partial \ell}{\partial \sigma_1^2} &= \sum_{j=1}^{n} \frac{p_1}{p_1\phi_1(x_j; \mu_1, \sigma_1^2) + p_2\phi_2(x_j; \mu_2, \sigma_2^2)} \cdot \frac{\partial}{\partial \sigma_1^2}\phi_1(x_j; \mu_1, \sigma_1^2) \\
&= \sum_{j=1}^{n} \frac{p_1\phi_1(x_j; \mu_1, \sigma_1^2)}{p_1\phi_1(x_j; \mu_1, \sigma_1^2) + p_2\phi_2(x_j; \mu_2, \sigma_2^2)} \left(-\frac{1}{2\sigma_1^2} + \frac{(x_j - \mu_1)^2}{2(\sigma_1^2)^2}\right) \\
&= \sum_{j=1}^{n} \pi_{1j} \frac{-\sigma_1^2 + (x_j - \mu_1)^2}{2(\sigma_1^2)^2}
\end{aligned}$$

Using the current estimate $\hat{\pi}_{1j}$ and $\hat{\mu}_i$, and setting

$$\frac{\partial \ell}{\partial \sigma_1^2} = \sum_{j=1}^{n} \hat{\pi}_{1j} \frac{-\sigma_1^2 + (x_j - \hat{\mu}_1)^2}{2(\sigma_1^2)^2} = 0$$

we obtain

$$\hat{\sigma}_1^2 = \sum_{i=1}^{n} \frac{\hat{\pi}_{1i}(x_i - \hat{\mu}_1)^2}{\sum_{j=1}^{n} \hat{\pi}_{1j}}$$

Analogously,

$$\hat{\sigma}_2^2 = \sum_{i=1}^{n} \frac{\hat{\pi}_{2i}(x_i - \hat{\mu}_2)^2}{\sum_{j=1}^{n} \hat{\pi}_{2j}}$$

- For the mixture proportion p_i, recall that we have already derived in (3) that

$$p_m = \frac{1}{n} \sum_{i=1}^{n} \pi_{mi}, \qquad m = 1, 2.$$

Using the current estimate $\hat{\pi}_{mi}$, we have

$$p_m = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_{mi}, \qquad m = 1, 2.$$

3 EM for missing values in multivariate normal

EM has many applications. A good illustration to show EM in action is EM for imputation.
Assume that there are missing values in observations that are from multivariate normal distribution.

- Objective: Impute missing component values of multivariate observations.
- Assumption: The values are missing randomly (termed *Missing at Random*).
- Method of Imputation: EM algorithm.

  EM Iterations between the Expectation (prediction) step and the Maximization (estimation) step.

E-step : For each data point with missing values, estimate using the mean of the conditional distribution

$$\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{x}_2 \ \sim N_k(\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2),\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

where the partition of the original observation vector is $\boldsymbol{X}=(\boldsymbol{X}_1',\boldsymbol{X}_2')'$,

- $\boldsymbol{X}_1$ denotes the missing components,
- $\boldsymbol{X}_2$ denotes the observed components.
- The order of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is for convenience, can be switched or relabelled.

M-step : Perform MLE estimation of parameters ($\boldsymbol{\mu}_i$'s and Σ_{ij}'s) based on the (imputed) complete data.

Iteration : Repeat E-step and M-step until the imputed values converge.

Examples of EM imputation steps are shown in class demo.

The process show that EM can be used for imputation (as in the examples) as well as for parameter estimation for the $\boldsymbol{\mu}_i$'s and Σ_{ij}'s.

The imputation examples also implies that the process can extend to non-normal distribution, such as binomial, trinomial, etc.

Notes

The materials in the notes below are to show the value and the widespread applications of the EM methods.
The contents will not be covered in the final exam.

4 EM approach for parametric mixture models* (Latent variable model II)

In this section we outline how a latent variable approach of EM works, and why it works.

Consider the mixture model setup:

A random vector X with parametric mixture density

$$f(\boldsymbol{x};\theta_1,\cdots,\theta_k)=p_1f_1(\boldsymbol{x};\theta_1)+\cdots+p_kf_k(\boldsymbol{x};\theta_k),\qquad p_1+\cdots+p_k=1.$$

Assume that we observe $X_i=\boldsymbol{x}_i, i=1,\cdots,n$, each belonging to one and only one component in the mixture. However the true membership is not observed.

Objectives

Estimate p_c, $c=1,\cdots,k$, with $\sum_{c=1}^k p_c=1$.

Then assign membership to new observations.

An alternative EM approach - via latent variables

In this approach, the key role is played by a latent, unobservable variable.

By introducing a latent variable Z_i for the ith observation, the bothersome sum

$$\sum_{m=1}^k p_mf_m(\boldsymbol{x}_i;\theta_m)$$

in the likelihood function changes into a more manageable product

$$\prod_{c=1}^k (p_cf_c(\boldsymbol{x}_i;\theta_c))^{z_{ic}}$$

The latent variable of membership

- For the ith observation $\boldsymbol{x}_i$, define $Z_i=(0,\cdots,0,1,0,\cdots,0)$, an unobserved random k-vector, where the single non-zero component indicates the true membership of the ith observation.
- A natural probability distribution for Z_i is multinomial. Let z_i be a realization of Z_i.

$$p(z_i)=P(Z_i=z_i)=p_1^{z_{i1}}\cdots p_k^{z_{ik}}=\prod_{c=1}^k p_c^{z_{ic}}\quad=p_j\quad \text{if}\ \ z_{ic}=\begin{cases}1, & c=j\\ 0 & c\neq j,\end{cases}$$

where

$$z_i=(z_{i1},\cdots,z_{ik}),\qquad p_1+\cdots+p_k=1.$$

For example, for $z_i=(0,1,0,\cdots,0)$, that the ith observation is $\boldsymbol{x}_i$ is from component $c=2$, such event has probability

$$P(Z_i=z_i)=\prod_{c=1}^k p_i^{z_{ic}}=p_1^0\ p_2^1\ p_3^0\cdots p_k^0=p_2$$

- The conditional density of X_i given the true membership Z_i is

$$f(\boldsymbol{x}_i|z_i) = \prod_{c=1}^{k} f_c(\boldsymbol{x}_i;\theta_c)^{z_{ic}}$$

For example, when $z_i = (0, 1, 0, \cdots, 0)$, that $\boldsymbol{x}_i$ is from component 2, then

$$f(\boldsymbol{x}_i|z_i) = \prod_{c=1}^{m} f_c(\boldsymbol{x}_i;\theta_c)^{z_{ic}} = f_2(\boldsymbol{x}_i;\theta_2)$$

- The joint likelihood function for (X_i, Z_i), called **complete data likelihood**, can be written as

$$L(\theta, p\ |\boldsymbol{x}_i, z_i) = f(\boldsymbol{x}_i, z_i) = f(\boldsymbol{x}_i|z_i)p(z_i) = \left(\prod_{c=1}^{k} f_c(\boldsymbol{x}_i;\theta_c)^{z_{ic}}\right)\left(\prod_{c=1}^{k} p_c^{z_{ic}}\right) = \prod_{c=1}^{k}\Big(p_c f_c(\boldsymbol{x}_i;\theta_c)\Big)^{z_{ic}}$$

The joint log-likelihood function for all $(X_i, Z_i), i = 1, \cdots n$ can be written as

$$\ell(\theta, p\ |\boldsymbol{x}, z) = \sum_{i=1}^{n} \log f(\boldsymbol{x}_i|z_i)p(z_i) = \sum_{i=1}^{n}\sum_{c=1}^{k} z_{ic}\log\big(p_c f_c(\boldsymbol{x}_i;\theta_c)\big)$$

For example, in the two-component case of $k = 2$, the joint log-likelihood function above becomes

$$\ell(\theta, p\ |\boldsymbol{x}, z) = \sum_{i=1}^{n}\big[z_{i1}\log(p_1 f_1(\boldsymbol{x}_i;\theta_1)) + z_{i2}\log(p_2 f_2(\boldsymbol{x}_i;\theta_2))\big]$$

EM steps of the latent variable approach

Regroup the joint log-likelihood function for all $(X_i, Z_i), i = 1, \cdots n$,

$$\begin{aligned}
\ell(\theta, p\ |\boldsymbol{x}, z) &= \sum_{i=1}^{n} \log f(\boldsymbol{x}_i|z_i)p(z_i) \\
&= \sum_{i=1}^{n}\sum_{c=1}^{k} z_{ic}\log\big(p_c f_c(\boldsymbol{x}_i;\theta_c)\big) \\
&= \sum_{i=1}^{n}\sum_{c=1}^{k} z_{ic}\log(p_c) + \sum_{i=1}^{n}\sum_{c=1}^{k} z_{ic}\log f_c(\boldsymbol{x}_i;\theta_c) \\
&= \sum_{i=1}^{n}\sum_{z_{ci}=1}\log(p_c) + \sum_{i=1}^{n}\sum_{z_{ci}=1}\log f_c(\boldsymbol{x}_i;\theta_c)
\end{aligned}$$

Notice that the two parts depend on parameters in disjoint parameter subspaces, which means each part can be estimated individually.

It follows (omitted) that the MLE for the mixture proportion is the natural estimator

$$p_c = \frac{n_c}{n}$$

which does not explicitly depend on the specific form of f or the data $\boldsymbol{x}$ directly (note that n_c is not known).

The number of members in population c, n_c, is random. its posterior expectation given data and parameters is

$$\begin{aligned}
\mathbb{E}[n_c|\boldsymbol{x};\theta] &= \mathbb{E}\left[\sum_{i=1}^{n} 1_{\{i\text{th observation is from component } c\}}|\boldsymbol{x};\theta\right] \\
&= \sum_{i=1}^{n} P\{i\text{th observation is from component } c|\boldsymbol{x};\theta\} \\
&= \sum_{i=1}^{n} \pi_{ci}
\end{aligned}$$

Under the latent variable approach, the EM steps can be restated as follows:

- Take initial guesses for the parameters $\hat{\theta}_c, \hat{p}_c, c = 1, \cdots, k$.
- **Expectation step**

  The expected class membership Z_{ic} given data $\boldsymbol{x}$ and parameter θ_c's is the **conditional expectation**

$$\mathbb{E}[Z_{ic}|\boldsymbol{x};\theta] = \mathbb{E}[Z_{ic}|\boldsymbol{x}_i;\theta] = \mathbb{P}(Z_{ic} = 1|\boldsymbol{x}_i;\theta) = \frac{p_c f_c(\boldsymbol{x}_i;\theta_c)}{f(\boldsymbol{x}_i;\theta)} = \pi_{ci}$$

  The posteriors $\hat{\pi}_{ci}$ for $i = 1, \cdots, n$, can be estimated using (2),

$$\hat{\pi}_{ci} = \frac{\hat{p}_c f_c(x_i;\hat{\theta}_c)}{\hat{p}_1 f_1(x_i;\hat{\theta}_1) + \cdots + \hat{p}_k f_k(x_i;\hat{\theta}_k)}$$

- **Maximization step**

  Compute the updated mixing proportions and parameters.

$$\hat{p}_c \leftarrow \frac{\hat{n}_c}{n} = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{ci},$$

  The update of all current parameters leads to the next iteration.

$$\hat{\theta}^{(n+1)} \leftarrow \hat{\theta}^{(n)}$$

- Iterate these steps until convergence.

Example of p variate normal mixture

For example, in the case that population c is of normal distribution $N_p(\mu_c, \Sigma_c)$, $f_c(\boldsymbol{x};\theta_c) = \phi(\boldsymbol{x};\mu_c,\Sigma_c)$,

$$\hat{\mu}_c \ \leftarrow\ \sum_{i=1}^{n}\frac{\hat{\pi}_{ci}}{\sum_{j=1}^{n}\hat{\pi}_{cj}}\boldsymbol{x}_i$$

and

$$\hat{\Sigma}_c \ \leftarrow\ \frac{1}{\hat{n}_c}\sum_{i=1}^{n}\hat{\pi}_{ci}(\boldsymbol{x}_i - \hat{\mu}_c)(\boldsymbol{x}_i - \hat{\mu}_c)'$$

General parameters, such as the ones in the covariance matrix, will depend on the parameterization.

Notice that the membership assignments occurred in the form of posterior expectation, i.e., posterior probability.

5 Why EM works*

Each point in the data $X = X(\theta)$ should belong to a class c, but the class label is unknown — missing.

The goal is to find the class probability $P(c \mid X, \theta)$ by finding the MLE θ_{mle} that maximizes the log-likelihood

$$L(\theta) = L(\theta \mid X) = \ln P(X \mid \theta) = \ln \sum_c P(X, c \mid \theta)$$

The EM algorithm is an iterative procedure, which produces $L(\theta_n)$ after the nth iteration. In order to find the next iteration $L(\theta_{n+1}) > L(\theta_n)$, we try to maximize $L(\theta) - L(\theta_n)$.

$$\begin{aligned}
L(\theta) - L(\theta_n) &= \ln \sum_c P(X, c \mid \theta) \;-\; \ln P(X \mid \theta_n) \;=\; \ln \sum_c \frac{P(X, c \mid \theta)}{P(X \mid \theta_n)} \\
&= \ln \sum_c P(c \mid X, \theta_n) \; \frac{P(X, c \mid \theta)}{P(c \mid X, \theta_n) P(X \mid \theta_n)} \\
&= \ln \sum_c P(c \mid X, \theta_n) \; \frac{P(X, c \mid \theta)}{P(X, c \mid \theta_n)} \\
&\geq \sum_c P(c \mid X, \theta_n) \; \ln \frac{P(X, c \mid \theta)}{P(X, c \mid \theta_n)} \\
&= E_{c|X,\theta_n} \left[\ln P(X, c \mid \theta) - \ln P(X, c \mid \theta_n) \right] = \Delta(\theta \mid \theta_n)
\end{aligned}$$

The "$\geq$" is an application of Jensen's inequality, pertaining to the **convexity** (concavity) property of $\ln x$, that

$$\ln \left(\sum_i t_i x_i \right) \geq \sum_i t_i \ln x_i$$

for any finite sum $\sum_c t_c = 1$, $t \in (0, \infty)$. Here $t_c = P(c|X, \theta_n)$.

Consider the function

$$L_n(\theta) = L_n(\theta \mid \theta_n) = L(\theta_n) + \Delta(\theta \mid \theta_n)$$

Since $L(\theta) - L(\theta_n) \geq \Delta(\theta \mid \theta_n)$, we have

$$L_n(\theta) \leq L(\theta)$$

with

$$L_n(\theta_n) = L(\theta_n).$$

As a function of θ, the function $L_n(\theta)$ approaches the log-likelihood function $L(\theta)$ from below and agrees with $L(\theta)$ at $\theta = \theta_n$. The idea of EM algorithm is to find the next approximation θ_n such that

$$L_n(\theta_{n+1}) = L_n(\theta_{n+1} \mid \theta_n) = \max_\theta L_n(\theta \mid \theta_n)$$

Since $\sum_c P(c \mid X, \theta_n) \ln P(X, c \mid \theta_n)$ and $L(\theta_n)$ are not functions of θ,

$$\begin{aligned}
\theta_{n+1} &= \arg\max_\theta L_n(\theta) = \arg\max_\theta L_n(\theta \mid \theta_n) \\
&= \arg\max_\theta \left\{ \sum_c P(c \mid X, \theta_n) \ln P(X, c \mid \theta) - \sum_c P(c \mid X, \theta_n) \ln P(X, c \mid \theta_n) + L(\theta_n) \right\} \\
&= \arg\max_\theta \left\{ \sum_c P(c \mid X, \theta_n) \ln P(X, c \mid \theta) \right\} \qquad \text{(the part related to } \theta\text{)} \\
&= \arg\max_\theta E_{c|X,\theta_n} \left[\ln P(X, c \mid \theta) \right]
\end{aligned}$$

Note the expectation and maximization in the $\theta_n \to \theta_{n+1}$ step.

The two key steps in the iteration $\theta_n \to \theta_{n+1}$:

- **Expectation**

  Find the expectation of the joint likelihood of data X and class c w.r.t. posterior class probability $P(c|X, \theta_n)$, $a.k.a.$ w.r.t. the class probability conditioned on data X and the nth iteration parameter θ_n. In this step, one constructs a distribution function of c,

  $$L_n(\theta \mid \theta_n) = E_{c|X,\theta_n} \ln P(X, c \mid \theta) + L(\theta_n)$$

- **Maximization**

  Maximize the expectation, which also achieving maximization of $L_n(\theta \mid \theta_n)$, w.r.t. θ. In this step, one finds $\theta = \theta_{n+1}$ which maximizes the joint probability

  $$E_{c|X,\theta_n} \ln P(X, c \mid \theta)$$

  and furthermore, maximizes $L_n(\theta)$.

Remarks on EM methods

- Now maximizing $L(\theta)$ becomes maximizing $L_n(\theta \mid \theta_n)$, estimating $P(c \mid X, \theta_n)$ on the way.
- If one just wants to maximize $L(\theta)$, the class label type of parameters can be invented to aid the maximization process.
- EM method can be used in missing data **imputation**, thanks to the latent variable approach.

  (See examples in class)
- Convergence property of the probability distribution function and the data matters.
- EM as an chicken-and-egg problem solver
  - If we knew the membership Z, we can estimate population parameters (μ, σ^2).
  - If we knew the population parameters, we can estimate membership by classification method, such as assignment membership to the closest centroid.

  The problem: We don't know either.

  Remember the objective is to estimate either the membership probability or the parameters.

  EM will get the MLE of the parameters and membership estimate at the same time.