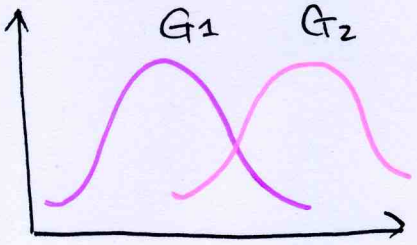
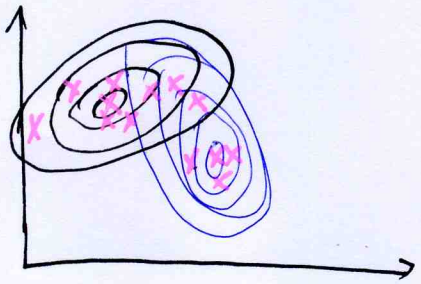
Mixture Models and EM - Expectation Maximization

① mixture model Revisited



$$X \sim \pi_1 G_1 + \pi_2 G_2$$
$$G_i \sim N(\mu_i, \sigma_i)$$
$$\Sigma_i \pi_i = 1$$



with the probability of π_i to select the cluster i, with the probability density function (pdf) f_1 and f_2 to sample the data.

HMM can be considered as a dynamic mixture model. (see previous lecture notes)

② K-Means Algorithm Revisited

Identifying groups (clusters) of data points in a multi-dimensional space. Given K as the cluster number. we hope to minimize ~~the~~ an objective function (also called distortion measure)

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2$$

where r_{nk} is an indicator that: ~ (belonging to the cluster or not)

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|x_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

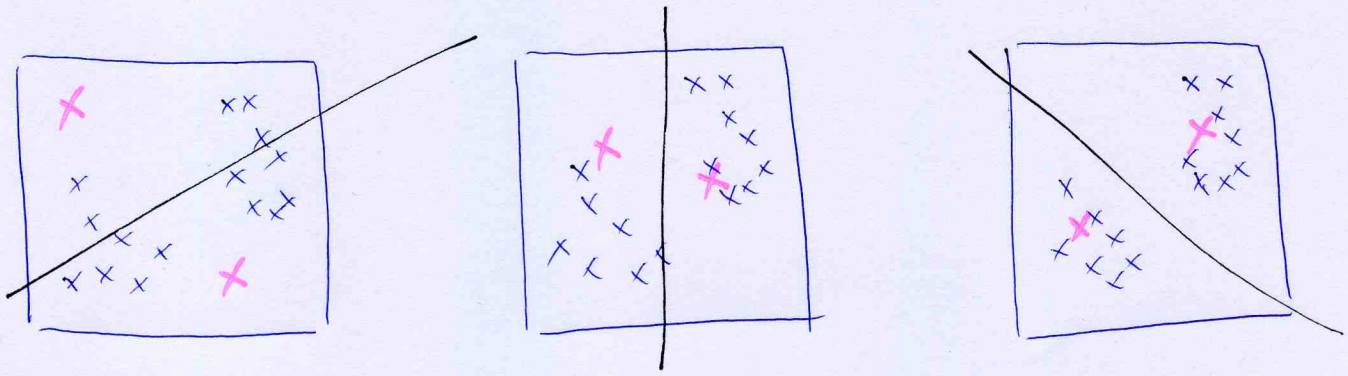
If r_{nk} ~~xxx~~ is fixed. (the clusters have existed!)

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}}$$

$\sum_{n=1}^{N} r_{nk}$: the number of elements in cluster k.

$\sum_{n=1}^{N} r_{nk} x_n$: all the elements in cluster k



Above figures are some schematic illustration of K-Means. In calculating M, we call it E-step (Expectation) given a set of new centers, we need to decide r_{nk} (the way of assigning each data point x_n) in order to maximize the objective function J. This is called M-step (Maximization).

③ Given a general form of Gaussian Mixture Model

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

where $N(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right\}$

$$P(z_k=1|x) = \frac{P(z_k=1)P(x|z_k=1)}{\sum_{k=1}^{K} P(z_k=1)P(x|z_k=1)}$$

$$= \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)} = r(z_k)$$

$z_k : k = 1 \cdots K$ can be regarded as a random variable

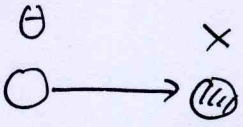
$z_k = \begin{cases} 0 \\ 1 \end{cases}$ for the data point x belonging to cluster k.

z is a multinomial random variable. e.g.

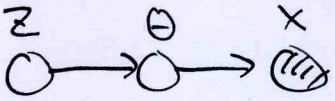
$z = [0\,1\,0\,0\,0\,0]$, $z = [0\,0\,0\,1\,0\,0]$ for $K=6$.

latent variables

④ Bayesian and Frequentist



We can regard θ as the parameter. It is just a parameter estimation problem



θ in Bayesian can also be treated as a random variable, which can be conditional on ~~on~~ another random variable. Here, we refer~~red~~ them as "latent variables", or hidden variables. Because they cannot be directly observed.

⑤ maximum Likelihood of Gaussian Mixture Model (see previous lecture notes). In multi-dimensional case:

$$\mathcal{L} = \ln p(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right\}$$

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = -\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_k,\Sigma_k)} \Sigma_k (x_n - \mu_k)$$

$$= -\sum_{n=1}^{N} \gamma(z_{nk}) \Sigma_k (x_n - \mu_k) = 0$$

Constant

See Appendix

$$\Rightarrow \sum_{n=1}^{N} \gamma(z_{nk}) = \sum_{n=1}^{N} \Sigma_k (x_n - \mu_k)$$

$$= \sum_k \left(-\sum_{n=1}^{N} \gamma(z_{nk}) x_n + \sum_{n=1}^{N} \gamma(z_{nk}) \mu_k\right) = 0$$

$$\Rightarrow \mu_k = \left(\sum_{n=1}^{N} \gamma(z_{nk}) x_n\right) \cdot \frac{1}{\sum_{n=1}^{N} \gamma(z_{nk})} \longrightarrow \frac{1}{N_k}$$

$$\Rightarrow \boxed{\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) = \sum_{n=1}^{N} p(z=k|x_n)$$

$$\frac{\partial \mathcal{L}}{\partial \Sigma_k} = 0 \Rightarrow \Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$$

Now, Let's consider the third parameter π_i with the constraint of $\sum_i \pi_i = 1$. Using lagrage multiplier

$$\mathcal{L} = \ln p(X|\pi,\mu,\Sigma) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi} = 0 \Rightarrow \sum_{n=1}^{N}\frac{\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} + \lambda = 0$$

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

⑥ In summary. The EM Algorithm for the Gaussian Mixture Model is as the following:

1. Initialize, μ_k, Σ_k and mixing coefficient π_k to evalure the initial value of log liklihood.

2. E-step. evaluate the responsibilies using the current paramete values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j \mathcal{N}(x_n|\mu_j,\Sigma_j)} \quad \text{for each } x_n,\ n = 1\cdots N$$

3. M-step: Re-estimate the parameter using current responsibilities

$$\mu_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n$$

$$\Sigma_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n - \mu_k^{new})(x_n - \mu_k^{new})^T$$

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

where $N_k = \sum_{n=1}^{N}\gamma(z_{nk})$

4. Evaluate the log likelihood

$$\ln p(x|\mu,\Sigma,\pi) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)\right\}$$

where $\mathcal{N}(x_n|\mu,\Sigma)$ is the pdf of multi-dim gaussian.

$X \sim N(\mu_k, \Sigma_k)$

$$P(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right\}$$

$$\ln P(X_1 \cdots X_N) = \ln P(X_1 \cdots X_N \mid \mu_k, \Sigma_k)$$

$$= \ln \prod_n P(X_n \mid \mu_k \Sigma_k) = \sum_{n=1}^{N} \ln P(X_n \mid \mu_k, \Sigma_k)$$

$$= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma_k| - \frac{1}{2}\sum_{n=1}^{N}(X_n-\mu_k)^T \Sigma_k^{-1}(X_n-\mu_k)$$

$$\frac{\partial \ln}{\partial \mu_k} = \left(-\frac{ND}{2}\ln(2\pi)\right)' - \left(\frac{N}{2}\ln|\Sigma_k|\right)' - \left(\frac{1}{2}\sum_{n=1}^{N}(X_n-\mu_k)^T\Sigma_k^{-1}(X_n-\mu_k)\right)'$$

$$= -\sum_{n=1}^{N}(X_n-\mu_k) = 0 \Rightarrow \mu_k = \frac{1}{N}\sum_{n=1}^{N} X_n$$

For mixture model where $k = 1 \cdots k$.

$$\ln = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{k} \pi_k N(X_n \mid \mu_k, \Sigma_k)\right\} \quad P(X \mid \pi, \mu, \Sigma) =$$

$$F = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{k} \pi_k f_k\right\} \qquad \boxed{\frac{\partial F}{\partial f_k} = \frac{\pi_k}{\pi_k f_k}}$$

$$= \sum_{n=1}^{N} \ln\{\pi_1 f_1 + \pi_2 f_2 \cdots \pi_k f_k\}$$

$$\frac{\partial F}{\partial f_k} = \frac{\pi_k}{\pi_1 f_1 + \cdots + \pi_k f_k}$$

$$\frac{f_k}{\partial \mu} \cdots \frac{\partial F}{\partial \mu} = \frac{\pi_k f_k}{\pi_1 f_1 \cdots \pi_k f_k} \frac{df_k}{\partial \mu}$$

$$= \gamma(\ldots)\Sigma_k(X_n - \mu_k)$$

Appendix 1