Mixture Models (and/towards EM)

So far - inference in models with missing and known variables whose distribution is known.
But -- dist's only linear Gaussian & discrete

might want more complex distribution over observed variables

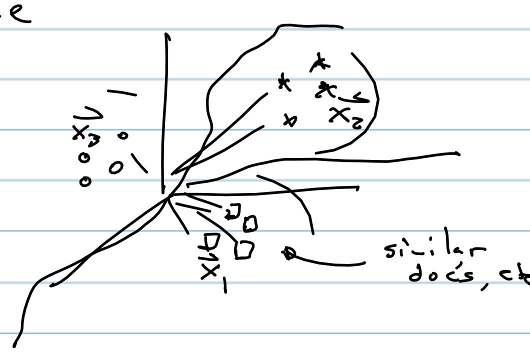
"If we define a joint dist. over observed and latent variables, the corresponding dist. on the obs. var's alone is obtained via marginalization. This allows complex marginal dists over observed vars to be expressed in terms of more tractable joint dist's over the extended space of observed and latent var's." The intro. of latent var's thereby allows complicated dist's to be formed from simpler components.

Mixture models ⟺ discrete latent vars
- useful for clustering data

Clustering
- Information retrieval, cluster text docs
- Image " " images
- Stock trajectories
- Customers based on preferences / choices / features, etc.

K-means (simple and intuitive)

One way to identify clusters of data points in a high dimensional space



Given

Data $\{\vec{x}_1, \dots, \vec{x}_N\}$, $\vec{x}_i \in \mathbb{R}^D$

\# obs N

Goal

Identify K clusters

"Formally" find groups of vectors/points whose inter-point distances are "smaller" in-cluster than out of cluster

"Parameters" $\{\vec{\mu}_k\}$ $\vec{\mu}_k \in \mathbb{R}^D$, "prototype" for cluster K also "center"

$r_{nk} \in \{0, 1\}$ indicator of $\vec{x}_n$ in cluster k

i.e. $r_{nk} = 1$ if $\vec{x}_n$ is associated with cluster center $\vec{\mu}_k$

Objective Function to Minimize

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\vec{x}_n - \vec{\mu}_k\|^2$$

Goal,

Identify $\{\vec{\mu}_k\}$ and $\{r_{nk}\}$ s.t. J is minimized

Algorithm

Two step proc. Chose some $\{\vec{\mu}_k\}$

1. Minimize J w.r.t. $\{r_{nk}\}$ with $\{\vec{\mu}_k\}$ fixed
2. Minimize J w.r.t. $\{\vec{\mu}_k\}$ " $\{r_{nk}\}$ "

Expectation Maximation

Step 1 (E)

- J is linear in r_{nk}
  - n terms independent, can optize each ind.

Solution

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\vec{x}_n - \vec{\mu}_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

Interpretation

choose r_{nk} by assigning $\vec{x}_n$ to nearest cluster center

Step 2 (M)

- J is quadratic in $\vec{\mu}_k$
- take deriv. and set equal to zero

$$\frac{\partial J}{\partial \vec{\mu}_k} = \sum_{n=1}^{N} \frac{\partial}{\partial \vec{\mu}_k}\left(r_{nk} \|\vec{x}_n - \vec{\mu}_k\|^2\right) = 0$$

$$= 2\sum_{n=1}^{N} r_{nk}(\vec{x}_n - \vec{\mu}_k) = 0$$

$$\Rightarrow \vec{\mu}_k = \frac{\sum_n r_{nk}\vec{x}_n}{\sum_n r_{nk}}$$

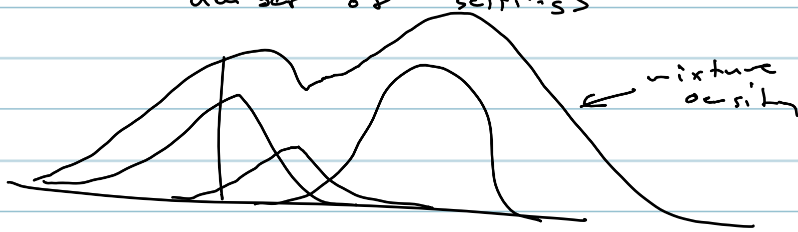
Interpretation

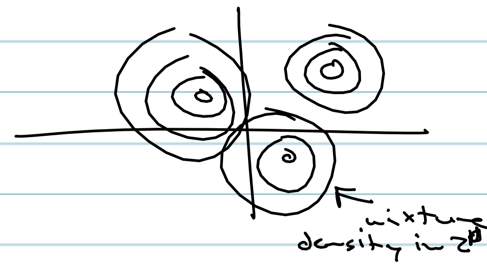
- $\sum_n r_{nk}$ is # points assigned to cluster k
- $\vec{\mu}_k$ is average of points assigned to cluster k

- Repeat til convergence
- Convergence to local minimal only

Mixtures of Gaussians

- Generalization of k-means (probabilistic)
- Useful for density estimation & clustering
- Quite useful in practice in an enormous number of settings





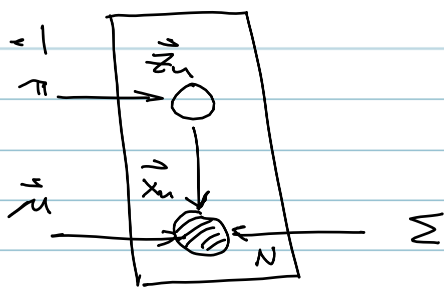
Notation

$$p(\vec{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\vec{x} \mid \vec{\mu}_k, \Sigma_k)$$

$\vec{z}$ as before, $\begin{array}{ccccccc} 1 & & & k & & & K \\ [0 & 0 & 0 & 1 & 0 & 0 & 0] \end{array}$ ← if $\vec{x}$ assigned to cluster k

$z_k \in \{0, 1\}$, $\sum_k z_k = 1$

Graphical Model



Joint Distribution

$$p(\vec{x}, \vec{z}) = p(\vec{z})\, p(\vec{x} \mid \vec{z})$$

where

$$p(z_k = 1) = \pi_k, \quad 0 \le \pi_k \le 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

can write $p(\vec{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$

and $p(\vec{x} \mid z_k = 1) = \mathcal{N}(\vec{x} \mid \vec{\mu}_k, \Sigma_k)$

can write $p(\vec{x} \mid \vec{z}) = \prod_{k=1}^{K} \mathcal{N}(\vec{x} \mid \vec{\mu}_k, \Sigma_k)^{z_k}$

Since the joint dist. is $p(\vec{x}|\vec{z})\,p(\vec{z})$ we can write

$$p(\vec{x}) = \sum_{\vec{z}} p(\vec{z})\,p(\vec{x}|\vec{z}) = \sum_{k=1}^{N} \pi_k \mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k)$$

This is for a single data point, for N datapoints $\vec{x}_n$ there is a corresponding $\vec{z}_n$

- Note, because of joint dist. EM possible.

Will need conditional dist of $\vec{z}\,|\,\vec{x}$
This is given by

$$\gamma(z_k) \equiv p(z_k=1|\vec{x}) = \frac{p(z_k=1)\,p(\vec{x}|z_k=1)}{\sum_{j=1}^{K} p(z_j=1)\,p(\vec{x}|z_j=1)}$$

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

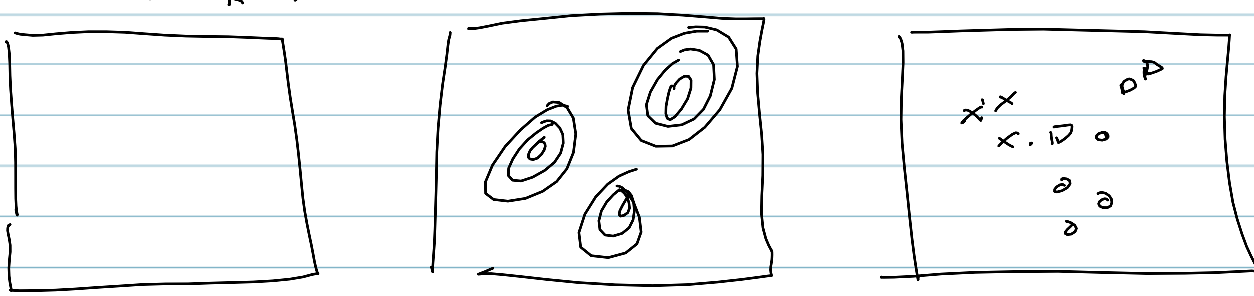
$\gamma(z_k)$ is the "responsibility" that component k takes for explaining the observation $\vec{x}$

Generating from a Gaussian mixture model

Ancestral sampling.

$z_1 \quad z_2 \quad z_3 \quad z_4 \quad . \quad ..$

$1 \quad 2 \quad 1 \quad k \quad 3 \quad 2 \quad '$



Maximum Likelihood

Given $\{\vec{x}_1, \dots, \vec{x}_N\}$ which we wish to model using a Gaussian mixture.

Represent

$$\underline{X} = \begin{bmatrix} \leftarrow & \vec{x}_1^T & \rightarrow \\ \vdots & \vec{x}_2^T & \vdots \\ \vdots & \vec{x}_3^T & \vdots \\ & \vdots & \\ \leftarrow & \vec{x}_N^T & \rightarrow \end{bmatrix} \qquad Z = \begin{bmatrix} \leftarrow & \vec{z}_1^T & \rightarrow \\ - & & \cdot \\ & \ddots & \\ \leftarrow & \vec{z}_N^T & \rightarrow \end{bmatrix}$$

- Graphical model from before
- Log likelihood

(1) $$\ln p(\underline{X} \mid \vec{\pi}, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\vec{x}_n \mid \mu_k, \Sigma_k) \right\}$$

Wish to max. w.r.t. $\vec{\pi}, \mu, \Sigma$

Notable complications (even with $\Sigma_k = \sigma_k^2 I$)

a) - $\vec{\mu}_j = \vec{x}_n$ fit, likelihood has term that can go to infinity

$$\mathcal{N}(x_n \mid x_n, \sigma_j^2 I) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}$$

$$\lim_{\sigma_j \to 0} \frac{1}{\sigma_j} \to \infty$$

maximizing (1) can cause problems like this

b) Identifiability, $K!$ different equivalent modes.

EM for Gaussian Mixtures

- EM has broad applicability
- Generalizations possible, including variational inference

To start let's look at conditions at ML solution for G.M.

$$\frac{\partial}{\partial \mu_k} \ln p(\underline{X} \mid \vec{\pi}, \mu, \Sigma) = 0$$

$$\Rightarrow \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\vec{x}_n \mid \vec{\mu}_k, \Sigma_k) \right\} = 0$$

$$\Rightarrow \cancel{\sum_{k=1}^{K} \pi_k \mathcal{N}(\vec{x}_n \mid \vec{\mu}_k, \Sigma_k)}$$

$$= \sum_{n=1}^{N} \left[\left(\frac{1}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\vec{x} \mid \vec{\mu}_j, \Sigma_j)} \right) \pi_k \mathcal{N}(\vec{x}_n \mid \vec{\mu}_k, \Sigma_k) \Sigma_k^{-1} (x_n - \mu_k) \right]$$

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

Multiplying both sides by (non-singular) Σ_k^{-1} we obtain

$$\sum_{n=1}^{N} \gamma(z_{nk}) (\vec{x}_n - \vec{\mu}_k) = 0 \qquad \left\{ \sum_{n=1}^{N} \gamma(z_{nk}) \vec{x}_n = \sum_{n=1}^{N} \gamma(z_{nk}) \vec{\mu}_k \right.$$

which implies

$$\mu_k = \frac{1}{\sum_{n=1}^{N} \gamma(z_{nk})} \cdot \sum_{n=1}^{N} \gamma(z_{nk}) \vec{x}_n \quad \text{or} \quad \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \vec{x}_n$$

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

"num points assigned to cluster k"

looks like weighted average (don't know $\gamma(z_{nk})$'s)

$$\frac{\partial}{\partial \Sigma_k} \ln p(\cancel{X} | \pi, \mu, \Sigma)$$

exponent

$$= \sum_{n=1}^{N} \left[\frac{1}{\sum_{j=1}^{K} \pi_j N(x_n | \mu_j, \Sigma_j)} \cdot \pi_k N(x_n | \mu_k, \Sigma_k) \cdot \cancel{(x_n - \mu_k)(x_n - \mu_k)} \right.$$

$$= \cancel{\sum_{n=1}^{N} r(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T = 0}$$

$$\frac{\partial}{\partial \Sigma} \ln \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

$$= -\Sigma \Sigma^{-1}$$

$$= \frac{\partial}{\partial \Sigma} \left(-\frac{1}{2} \ln |\Sigma| - \frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$

plugging back in

$$= -\frac{1}{2} (\Sigma^{-1}) \qquad + \frac{1}{2} \Sigma^{-1} (x-\mu)(x-\mu)^T \Sigma^{-1}$$

$$\Rightarrow$$

$$\cancel{\frac{1}{2} (\Sigma^{-1}) = \frac{1}{2} \Sigma^{-1} (x-\mu}$$

$$\sum_{n=1}^{N} r(z_{nk}) \Sigma^{-1} = \sum_{n=1}^{N} r(z_{nk}) \Sigma^{-1} (x-\mu)(x-\mu)^T \Sigma^{-1}$$

multiplying through by Σ twice and

$$\Sigma_k = \frac{1}{\sum_{n=1}^{N} r(z_{nk})} \left(\sum_{n=1}^{N} (x_n - \mu_k)(x - \mu_k)^T \right)$$

used $\dfrac{\partial \ln |\Sigma|}{\partial \Sigma} = (\Sigma^{-1})^T$ and $= (\Sigma^T)^{-1}$ | but since Σ is ~~po~~ symmetric $= \Sigma^{-1}$

used $\dfrac{\partial a^T X^{-1} b}{\partial X} = -X^{-T} a b^T X^{-T}$

Maximizing for Σ_k yields

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\vec{x}_n - \vec{\mu}_k)(\vec{x}_n - \vec{\mu}_k)^T$$

which looks like a weighted ML covariance matrix estimate.

Last - max $p(X \mid \pi, \mu, \Sigma)$ w.r.t. π_k
under constraint that

$$\sum_{k=1}^{K} \pi_k = 1$$ - sol'n: use lagrange mult.

and maximize

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

$$\frac{\partial}{\partial \pi_k} \ln P(X \mid \pi, \mu, \Sigma) + \lambda = 0$$

$$= \sum_{n=1}^{N} \frac{N(x_n \mid \mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n \mid \mu_j, \Sigma_j)} + \lambda = 0$$

to eliminate lagrange mult.

$$\cancel{= \sum_{n=1}^{N} \gamma(z_{nk}) + \lambda = 0}$$

$$\cancel{= \pi_k \sum_{n=1}^{N} \gamma(z_{nk}) = -\pi_k \lambda}$$

$$\cancel{= \sum_{k=1}^{K} \pi_k \left(\sum_{n=1}^{N} \gamma(z_{nk}) \right) = \sum_{k=1}^{K} \pi_k \lambda}$$

$$\cancel{= \frac{\sum_{k=1}^{K} \pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_j \pi_j N(x_n \mid \mu_j, \Sigma_j)}} = 1 = 1 \cdot \lambda$$

Solving for π_k at optimal using Lagrange Multiplier

$$\frac{\partial}{\partial \pi_k}\left[p(\underline{X} \mid \pi, \mu, \Sigma) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right] = 0$$

$$\sum_{n=1}^{N} \frac{\partial}{\partial \pi_k} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k) + \lambda = 0$$

$$\sum_{n=1}^{N} \frac{\mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{N} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)} = -\lambda$$

Trick, multiply both sides by π_k and sum over k

$$\sum_{k=1}^{K} \pi_k \sum_{n=1}^{N} \frac{\mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{N} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)} = -\lambda \sum_{k=1}^{K} \pi_k$$

rearrange

$$\sum_{n=1}^{N} \underbrace{\frac{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{N} \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}}_{1} = -\lambda \quad \Rightarrow \quad N = -\lambda$$

using the same trick we get

$$\pi_k N = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n \mid \mu_j, \Sigma_j)}$$

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

$$\pi_k = N_k / N$$

Final Product : EM for Gaussian Mixtures

1. Initialize means $\vec{\mu}_k$ and covariances Σ_k and mixing coefficients π_k

2. E step

compute responsibilities

* old μ_k's and Σ_k's

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\vec{x}_n | \vec{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\vec{x}_n | \vec{\mu}_j, \Sigma_j)}$$

3 M step

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

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

$$\pi_k^{new} = N_k / N \quad \text{where} \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

4 Evaluate log lik. and check for convergence (params, log lik, etc.)

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

Repeat!

- Note, can take a long time to come to convergence

- Will converge to a local maximum