9 Mixture Models and EM

9.1 K-means Clustering

Definition. ***Concepts***

1. ***Clustering***
2. ***K-means*** *Problem with K-means is that at every step, every point is assigned to one and only one cluster centers, but maybe many points which are halfway between cluster centers. It is beneficial to have soft assignment of such points by adopting a probabilistic viewpoint*
3. ***Application*** *in image segmentation and compression*

9.2 Mixture of Gaussians

Definition. ***Mixtures of Gaussians*** *Given*

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

Reformulate linear mixture models as Gaussian mixtures in terms of discrete latent variables. Goal is to define random variable $\mathbf{x}$ *and* $\mathbf{z}$ *such that the distribution of* $\mathbf{x}$ *is equivalent to linear basis model with Gaussian kernel*

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z}) p(\mathbf{z})$$

Define $\mathbf{z}$: *1-of-K encoding* $\rightarrow [0, 1]$ *as a soft assignment to cluster centers. then marginal distribution over* $\mathbf{z}$ *given by*

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \qquad \textit{such that} \qquad p(z_k) = \pi_k$$

Define conditional distribution of $\mathbf{x}$ *over* $\mathbf{z}$ *as Gaussian*

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right)^{z_k} \qquad \textit{such that} \qquad p(\mathbf{x}|z_k = 1) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

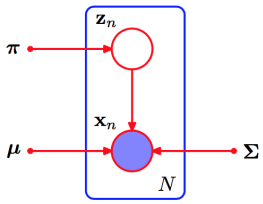
Therefore

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

The posterior distribution of $\mathbf{z}$ *given observation* $\mathbf{x}$*, responsibility, is given by*

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{k=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

Graphical representation of a Gaussian mixture model for a set of N i.i.d. data points $\{\mathbf{x}_n\}$, with corresponding latent points $\{\mathbf{z}_n\}$, where $n = 1, \ldots, N$.



Definition. ***MLE for Gaussian Mixtures***
Assume $N \times D$ *matrix* $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ *and* $N \times K$ *matrix* $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^N$, *then we want to maximize log likelihood*

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln \prod_{n=1}^{N} p(\mathbf{x}_n) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\}$$

Maximum likelihood has some problems (here)

1. *overfits because of existence of singularity*
2. *identifiability problem*

Naively, we solve MLE by taking derivative with respect to $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, *and* $\boldsymbol{\pi}$, *the solution is not a closed form solution as* N_k *invovles parameters that we want to estimate*

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{K} \gamma(z_{nk}) \mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left\|\mathbf{x}_n - \boldsymbol{\mu}_k\right\|^2 \qquad \pi_k = \frac{N_k}{N}$$

where $N_k = \sum_{n=1}^{N} \gamma(r_{nk})$ *represents the* ***effective number of points*** *assigned to cluster* k. *MLE estimator* $\boldsymbol{\mu}_k$ *for* k*-th Gaussian component is obtained by taking a weighted mean of all points in the dataset in which the weight factor for each data point* $\mathbf{x}_n$ *is given by the posterior probability* $\gamma(z_{nk}) = p(z_{nk} = 1|\mathbf{X})$ *tha component* k *was responsible for generating* $\mathbf{x}_n$. π_k *the* k*-th* ***mixing coefficient*** *is given by the average responsibility that component takes for explaining the dataset*

Definition. ***EM for Gaussian Mixtures***
The goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *for which there is no closed form solution. (detailed here)*

1. *Pick some initial value for* $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$, *and alternate between E step and M step*
2. ***E step*** *use value of parameter to evaluate the responsibility* $\gamma(r_{nk})$, *i.e. soft assignment*
3. ***M step*** *re-estimate* $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}$ *using maximum likelihood formulas*

4. *Evaluate log likelihood by*

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\}$$

check for convergence, return to step 2 if convergence criterion not satisfied

9.3 An alternative view of EM

Definition. ***Expectation-Maximization***

1. ***Goal****: find maximum likelihood solutions for models having latent variables*
2. ***EM model*** *Given observed data* $\mathbf{X}$ *and latent variables* $\mathbf{Z}$*, and set of model parameters* $\boldsymbol{\Theta}$*, we want to maximize the log likelihood function*

$$\ln p(\mathbf{X}|\boldsymbol{\Theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) \right\}$$

   $\{\mathbf{X}, \mathbf{Z}\}$ *is the commplete dataset, and he actual observed* $\{\mathbf{X}\}$ *is incomplete dataset.*

3. ***EM algorithm sketch*** *Idea is we cannot compute complete-data likelihood* $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta})$ *but we instead approximate it with posterior of latent variable* $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\Theta})$ *(E step). The expectation is given by*

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{old}) = \mathbb{E}_{\mathbf{Z} \sim p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\Theta}^{old})} \left\{ \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta}) \right\} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\Theta}^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\Theta})$$

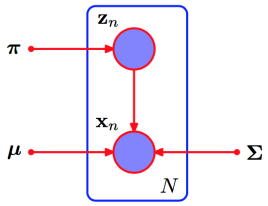
   Then we maximize this expectation with respect to the parameters (M step).

$$\boldsymbol{\Theta}^{new} = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{old})$$

   Detailed algorithm here

Definition. ***Gaussian Mixture Revisited***
If given complete data $\{\mathbf{X}, \mathbf{Z}\}$*, then the graphical model is as follows*

This shows the same graph as in Figure 9.6 except that we now suppose that the discrete variables $\mathbf{z}_n$ are observed, as well as the data variables $\mathbf{x}_n$.



the likelihood function is simply the joint distribution of $\mathbf{X}, \mathbf{Z}$ *over deterministic model parameters*

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \overset{graphical}{=} \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)^{z_{nk}}$$

Note this is different from previous log likelihood where only have incomplete data $\{\mathbf{X}\}$ *in that we have to consider joint distribution of both* $\mathbf{X}$ *and* $\mathbf{Z}$ *since both are observed variables. We note that this likelihood fucntion have closed form solution and so EM algorithm works trivially*

$$\mathcal{L}_{complete} = \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n}\sum_{k} z_{nk}\left(\ln \pi_k + \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right)$$

If give only incomplete data $\{\mathbf{X}\}$*, then using Bayes rule to get posterior distribution of* $\mathbf{Z}$

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto p(\mathbf{Z}, \mathbf{X}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right)^{z_{nk}}$$

where $\{\mathbf{z}_n\}$ *are independent in this case. Then we can derive expectation that we want to maximize*

$$\mathcal{L}_{incomplete} = \mathbb{E}_{\mathbf{Z}}\left\{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})\right\} = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})(\ln \pi_k + \ln \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right))$$

which has no closed form solution, so we use EM algorithm exactly the same as before.

Definition. ***Relation to K-means***

1. ***Comparison*** *K-means performs* ***hard assignment*** *of data potins to clusters. The EM algorithm makes* ***soft assignment*** *based on posterior probabilities*

2. ***Idea*** *A Gaussian mixture model in which covariance matrix are given by* ϵI *where* $\epsilon \to 0$ *is equivalent to K-means, i.e.*

$$\gamma(z_{nk}) \overset{\epsilon \to 0}{\longrightarrow} r_{nk}$$

$$\mathbb{E}_{\mathbf{Z}}\left\{\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})\right\} \overset{\epsilon \to 0}{\longrightarrow} -\sum_{n}\sum_{k} r_{nk}\left\|\mathbf{x}_n - \boldsymbol{\mu}_k\right\|^2$$

Also note that K-means does not estimate covariances of clusters but only cluster means.

9.4 The EM Algorithm in General

Definition. ***Kullback-Leibler Divergence*** *$KL(P||Q)$ from Q to P is the information gain if Q is used instead of P. In application, P typically represents the true distribution, while Q typically represents a model, or an approximation of P. In order to find a distribution Q that is closest to P, we can minimize KL divergence and compute an information projection.*

$$KL(P||Q) = \mathbb{E}_{i\sim P}\left\{\log\frac{P(i)}{Q(i)}\right\} = -\sum_i P(i)\log\frac{Q(i)}{P(i)}$$

EM Algorithm in General

1. Given probabilistic model $p(\mathbf{X},\mathbf{Z}|\mathbf{\Theta})$, goal is to maximize the likelihood fucntion given by

$$p(\mathbf{X}|\mathbf{\Theta}) = \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\mathbf{\Theta})$$

Idea is optimizing $p(\mathbf{X}|\mathbf{\Theta})$ directly is difficult, but optimizing $p(\mathbf{X},\mathbf{Z}|\mathbf{\Theta})$ is significantly easier. For any distribution $q(\mathbf{Z})$ over latent variables, we have following decomposition

$$\ln p(\mathbf{X}|\mathbf{\Theta}) = \mathcal{L}(q,\mathbf{\Theta}) + KL(q||p)$$

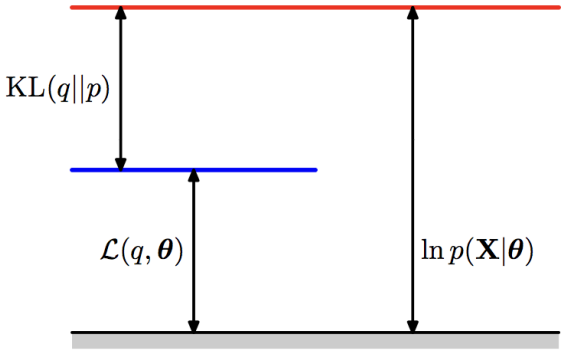
where

$$\mathcal{L}(q,\mathbf{\Theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z})\ln\left\{\frac{p(\mathbf{X},\mathbf{Z}|\mathbf{\Theta})}{q(\mathbf{Z})}\right\} \qquad KL(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z})\ln\left\{\frac{p(\mathbf{X}|\mathbf{Z},\mathbf{\Theta})}{q(\mathbf{Z})}\right\}$$

Proof.

$$\begin{aligned}
\mathcal{L}(q,\mathbf{\Theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z})\left(\ln p(\mathbf{X},\mathbf{Z}|\mathbf{\Theta}) - \ln q(\mathbf{Z})\right)\\
&= \sum_{\mathbf{Z}} p(\mathbf{Z})\left(\ln p(\mathbf{Z}|\mathbf{X},\mathbf{\Theta}) + \ln p(\mathbf{X}|\mathbf{\Theta}) - \ln q(\mathbf{Z})\right)\\
&= \sum_{\mathbf{Z}} p(\mathbf{Z})\left(\ln p(\mathbf{Z}|\mathbf{X},\mathbf{\Theta}) - \ln q(\mathbf{Z})\right) + \sum_{\mathbf{Z}} p(\mathbf{Z})\ln p(\mathbf{X}|\mathbf{\Theta})\\
&= \sum_{\mathbf{Z}} p(\mathbf{Z})\left(\frac{p(\mathbf{Z}|\mathbf{X},\mathbf{\Theta})}{p(\mathbf{Z})}\right) + \ln p(\mathbf{X}|\mathbf{\Theta})\sum_{\mathbf{Z}} p(\mathbf{Z})\\
&= -KL(p||q) + \ln p(\mathbf{X}|\mathbf{\Theta})
\end{aligned}$$

Note $KL(q||p)$ is kullback-leibler divergence between $q(\mathbf{Z})$ and the posterior $p(\mathbf{Z}|\mathbf{X},\mathbf{\Theta})$. Note $KL(q||p) \geq 0$ with equality if and only if $q(\mathbf{Z}) - p(\mathbf{Z}|\mathbf{X},\mathbf{\Theta})$. Hence $\mathcal{L}(q,\mathbf{\Theta}) \leq \ln p(\mathbf{X}|\mathbf{\Theta})$, i.e. $\mathcal{L}(q,\mathbf{\Theta})$ is a lower bound on $\ln p(\mathbf{X}|\mathbf{\Theta})$


KL(q||p)
$\mathcal{L}(q, \boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

□