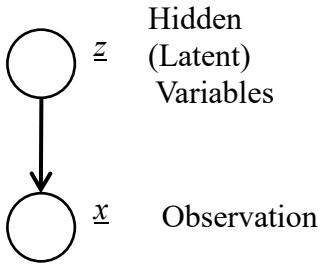
Lecture 12:

Why Gaussian mixtures?

Parametric, e.g., Gaussian ….Lecture 3….fast but limited

Non Parametric…. Lectures 4-5, kNN, Parzen, PNN …. general but slow (require lot of data)

Mixture Models: RBF, Mixture of exeprts,….



$$z_j \in \{0,1\} \text{ and } \sum_{j=1}^{M} z_j = 1$$

$$P(z_j = 1) = P_j \Rightarrow P(\underline{z}) = \prod_{j=1}^{M} P_j^{z_j}$$

$$p(\underline{x} \mid \underline{z}) = \prod_{j=1}^{M} [N(\underline{x}; \underline{\mu}_j, \Sigma_j)]^{z_j}$$

$$\Rightarrow p(\underline{x}) = \sum_{\underline{z}} p(\underline{x}, \underline{z}) = \sum_{\underline{z}} P(\underline{z})\, p(\underline{x} \mid \underline{z})$$

$$= \sum_{\underline{z}} \prod_{j=1}^{M} [P_j N(\underline{x}; \underline{\mu}_j, \Sigma_j)]^{z_j} = \sum_{j=1}^{M} P_j N(\underline{x}; \underline{\mu}_j, \Sigma_j)$$

pdf of $\underline{x}$ is a Gaussian Mixture

We have access to only $\underline{x}$. Given incomplete data

$$D = \left\{ \underline{x}^1 \ \underline{x}^2 \ \ldots . \underline{x}^N \right\}, \textit{ find the ML estimates of } \left\{ P_j, \underline{\mu}_j, \Sigma_j \right\}_{j=1}^{M}$$

$$\text{Let } \underline{\theta} = \left\{ P_j, \underline{\mu}_j, \Sigma_j \right\}_{j=1}^{M}$$

$$\min_{\underline{\theta}} J \quad \textit{where} \quad J = -\ln p(D \mid \underline{\theta})$$

If we had complete data:

$$D_c = \left\{ (\underline{x}^1, \underline{z}^1), \ (\underline{x}^2, \underline{z}^2) \ldots . (\underline{x}^N, \underline{z}^N) \right\}$$

$$\Rightarrow -\ln p(D_c \mid \underline{\theta}) = \sum_{n=1}^{N} \sum_{j=1}^{M} z_j^n \{ -\ln P_j + \frac{p}{2} \ln 2\pi + \frac{1}{2} \ln |\Sigma_j| + \frac{1}{2} \| \underline{x}^n - \underline{\mu}_j \|^2_{\Sigma_j^{-1}} \}$$

If had complete data, estimation is trivial. Similar to Gaussian case, except that we estimate with subsets of data that are assigned to each mixture component. Did this in Lecture 4.

In EM, replace each latent variable by its expectation with respect to the posterior density during the **E-step**

$$z_j^n \rightarrow E(z_j^n \mid \underline{x}^n, \underline{\theta}) = P(z_j^n = 1 \mid \underline{x}^n, \underline{\theta}) = \gamma_j^n \text{....responsibilities}$$

$$P(z_j^n = 1 \mid \underline{x}^n, \underline{\theta}) = \frac{P_j N(\underline{x}^n; \underline{\mu}_j, \Sigma_j)}{\sum_{k=1}^{M} P_k N(\underline{x}^n; \underline{\mu}_k, \Sigma_k)} = \gamma_j^n$$

In EM, minimize the expected value of the negative complete-data log likelihood during the **M-step**

$$Q(\underline{\theta}, \underline{\theta}^{old}) = E_{\underline{Z}}\{-\ln p(D_c \mid \underline{\theta})\} = \sum_{n=1}^{N}\sum_{j=1}^{M} \gamma_j^n \{-\ln P_j + \frac{p}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma_j| + \frac{1}{2}\| \underline{x}^n - \underline{\mu}_j \|^2_{\Sigma_j^{-1}}\}$$

$$\textit{subject to } \sum_{j=1}^{M} P_j = 1; P_j \geq 0$$

1. Initialize the means $\{\underline{\mu}_j\}_{j=1}^M$, covariances $\{\Sigma_j\}_{j=1}^M$, and mixing coefficients $\{P_j\}_{j=1}^M$.

   Evaluate J = -ln $p(\underline{x} \mid \underline{\theta}) = -\sum_{n=1}^{N} \ln\{\sum_{j=1}^{M} P_j N(\underline{x}^n; \underline{\mu}_j, \Sigma_j)\}$

2. E-step: Evaluate the responsibilities using the current parameter values

$$\gamma_j^n = \frac{P_j N(\underline{x}^n; \underline{\mu}_j, \Sigma_j)}{\sum_{k=1}^{M} P_k N(\underline{x}^n; \underline{\mu}_k, \Sigma_k)}; \; j = 1, 2, .., M; n = 1, 2, .., N$$

$$N_j = \sum_{n=1}^{N} \gamma_j^n; \; j = 1, 2, .., M$$

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

$$\underline{\mu}_j^{new} = \frac{1}{N_j}\sum_{n=1}^{N} \gamma_j^n \underline{x}^n$$

$$\Sigma_j^{new} = \frac{1}{N_j}\sum_{n=1}^{N} \gamma_j^n (\underline{x}^n - \underline{\mu}_j^{new})(\underline{x}^n - \underline{\mu}_j^{new})^T$$

$$P_j^{new} = \frac{N_j}{N}$$

4. Evaluate the negative log likelihood and check for convergence of parameters or the likelihood. If not converged, go to step 2.

Show slide 22.

What happens as $\Sigma_j \approx \varepsilon I$

$$\gamma_j^n = \frac{P_j N(\underline{x}^n; \underline{\mu}_j, \varepsilon I)}{\sum_{k=1}^{M} P_k N(\underline{x}^n; \underline{\mu}_k, \varepsilon I)}; \; j = 1,2,..,M; n = 1,2,..,N$$

$$\Rightarrow \qquad \gamma_j^n = \frac{P_j e^{-\|\underline{x}^n - \underline{\mu}_j\|^2/2\varepsilon}}{\sum_{k=1}^{M} P_k e^{-\|\underline{x}^n - \underline{\mu}_k\|^2/2\varepsilon}}$$

As $\varepsilon \to 0$

$$\gamma_j^n \to 1 if \;\; j = \arg\min_k \| \underline{x}^n - \underline{\mu}_k \|; \text{ the rest go to zero as long as none of the } P_j \text{ is zero}$$

K-means clustering algorithm:

K-means clustering to select K and the centers

a. Initialization

Choose initial center at random. Let n1 be the data point.

For k= 2,..,K

For n=1,2,.., N & n ≠ ni , i=1,2,..,k-1

$$D_n = \min_{1 \le i \le k-1} \left\| \underline{x}^n - \underline{\mu}_i \right\|_2^2$$

End

Select $\underline{\mu}_k = \underline{x}^{n_k}$ probabilistically $p(\underline{x}^{n_k}) = D(\underline{x}^{n_k}) \left[\sum_{\substack{n=1 \\ n \ne n^i; i=1,2,..,k-1}}^{N} D(\underline{x}^n) \right]^{-1}$

b. For n=1, 2 , …, N

Assign n to cluster $\mathbf{C}_j$ if $j = \arg\min_{1 \le k \le K} \left\| \underline{x}^n - \underline{\mu}_k \right\|_2$

End.

c. Recompute means $\underline{\mu}_j = \frac{1}{N_j} \sum_{n \in \mathbf{C}_j} \underline{x}^n$

d. If centers have changed, go to b, else stop

How to select K?

$$BIC \triangleq -2 \ln p(D \mid K, \underline{\mu}) + (Kp+1) \ln N$$

$$PE = \frac{2}{N} \sum_{j=1}^{K} \sum_{n \in C_j} \left\| \underline{x}^n - \underline{\mu}_j \right\|^2 + \frac{2Kp}{N} \sigma^2 \qquad \text{Prediction Error}$$

$$K_T = \arg\min_K \left\{ \frac{1}{Kp} \sum_{j=1}^{K} \sum_{i=1}^{p} \left(\frac{1}{|C_j|} \sum_{n \in \mathbf{C}_j} \left(\frac{x_i^n - \mu_{ji}}{\sigma_{ij}} \right)^4 - 3 \right) \right\}$$ Excess Kurtosis

$$J(D,K) = \frac{1}{|D|} \sum_{i \in D} \| \underline{x}_i - \hat{\underline{x}}_i \|_2^2$$ Knee of the MSE

$$\hat{\underline{x}}_i = \underline{\mu}_k, \text{ where } k = \arg\min_j \| \underline{x}_i - \underline{\mu}_j \|_2^2$$

There is a lot more to EM than this.

$\underline{x}$: *data*

$\underline{z}$: *hidden* var *iables* (*mixture*)

$\underline{\theta}$: *parameters*

$q(\underline{z})$ = *any arbitrary distribution approximating* $p(\underline{z} \mid \underline{x}, \underline{\theta})$

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

$$\underbrace{-\ln p(\underline{x}, \underline{z} \mid \underline{\theta})}_{complete\ NLL} = -\ln p(\underline{z} \mid \underline{x}, \underline{\theta}) - \ln p(\underline{x} \mid \underline{\theta})$$

$$\underbrace{-\ln p(\underline{x} \mid \underline{\theta})}_{surprise\ or\ NLL} = -\ln p(\underline{x}, \underline{z} \mid \underline{\theta}) + \ln q(z) + \ln p(\underline{z} \mid \underline{x}, \underline{\theta}) - \ln q(z)$$

$$= -\underbrace{E_q[\ln \frac{p(\underline{x}, \underline{z} \mid \underline{\theta})}{q(\underline{z})}]}_{\ln L(q, \underline{\theta})} + \underbrace{E_q[\ln \frac{p(\underline{z} \mid \underline{x}, \underline{\theta})}{q(\underline{z})}]}_{-KL(q(\underline{z}) \| p(\underline{z} \mid \underline{x}, \underline{\theta}))}$$

$$\Rightarrow J = -\ln L(q, \underline{\theta}) - KL(q(\underline{z}) \| p(\underline{z} \mid \underline{x}, \underline{\theta}))$$

$$J \leq \underbrace{-\ln L(q, \underline{\theta})}_{Expected\ Free\ energy} \quad \because KL(q(\underline{z}) \| p(\underline{z} \mid \underline{x}, \underline{\theta})) \geq 0$$

$$E - step : q(\underline{z}) = p(\underline{z} \mid \underline{x}, \underline{\theta}^{old})$$

$$M - step : \underline{\theta}^{new} = \min_{\underline{\theta}} [-\ln L(q, \underline{\theta})] \ldots . \min imize \exp ected \ \ free\ energy$$

$$= \min_{\underline{\theta}} -E_q[\ln p(\underline{x}, \underline{z} \mid \underline{\theta})]$$

$$= \min_{\underline{\theta}} \tilde{Q}(\underline{\theta}, \underline{\theta}^{old})$$

$$Note : -\ln L(q, \underline{\theta}) = Q(\underline{\theta}, \underline{\theta}^{old}) = \tilde{Q}(\underline{\theta}, \underline{\theta}^{old}) - H_q(\underline{z}, \underline{\theta}^{old})$$

Still another way of looking at it:

$p(\underline{z} \mid \underline{x}, \underline{\theta})$ *is difficult to compute, but* $p(\underline{x}, \underline{z} \mid \underline{\theta})$ *is easier*

$q(\underline{z})$ = *any arbitrary distribution approximating* $p(\underline{z} \mid \underline{x}, \underline{\theta})$

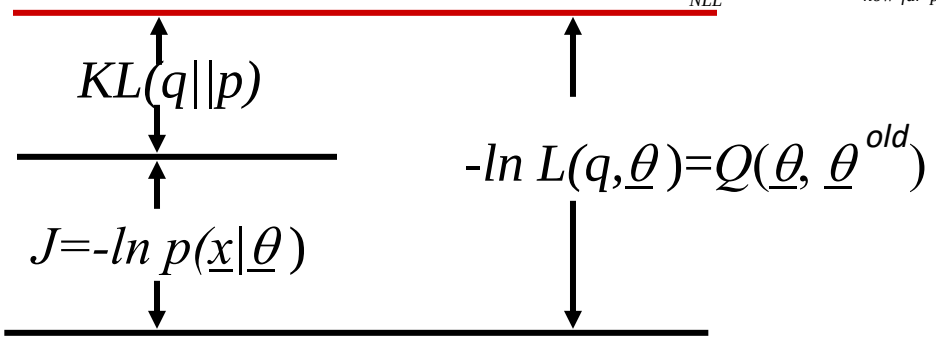
$$\underbrace{-\ln L(q, \underline{\theta})}_{Variational\ Free\ Energy} = KL(q(\underline{z}) \| p(\underline{x}, \underline{z} \mid \underline{\theta})) = -\ln p(\underline{x} \mid \underline{\theta}) + KL(q(\underline{z}) \| p(\underline{z} \mid \underline{x}, \underline{\theta})) \geq -\ln p(\underline{x} \mid \underline{\theta})$$

or $\ln L(q, \underline{\theta}) \leq \ln p(\underline{x} \mid \underline{\theta})$....*Evidence lower bound* $\Rightarrow \max imize \ln L(q, \underline{\theta})$ *or* $\min imize -\ln L(q, \underline{\theta})$

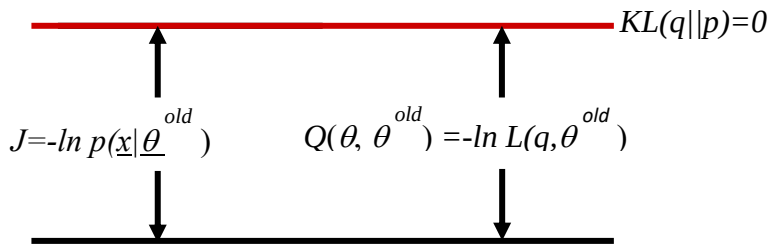
One more way:

$$\underbrace{-\ln L(q,\underline{\theta})}_{\text{Variational Free Energy}} = KL(q(\underline{z}) \,||\, p(\underline{x},\underline{z}\,|\,\underline{\theta})) = -H(q(z)) + E_{q(z)}[-\ln p(\underline{x},\underline{z}\,|\,\underline{\theta})]$$

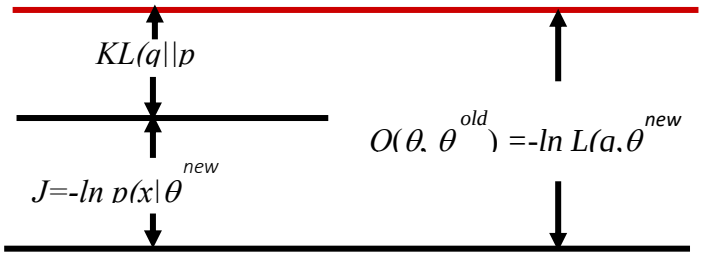
$$= E_{q(z)}\{\ln q(z) - \ln p(\underline{z}\,|\,\underline{\theta}) - \ln p(\underline{x}\,|\,\underline{z},\underline{\theta})\} = \underbrace{E_{q(z)}\{-\ln p(\underline{x}\,|\,\underline{z},\underline{\theta})\}}_{NLL} + \underbrace{KL(q(z) \,||\, p(\underline{z}\,|\,\underline{\theta}))}_{\text{how far posterior approx. from prior of latent}}$$



E-step:



M-step:



Variational Inference: What if q(z) is complicated and is expressed as

$$q(\underline{z}) = \prod_{j=1}^{K} q_j(\underline{z}_j); \{\underline{z}_j\} \text{ are disjoint groups}$$

Minimize the upper bound -lnL(q(z)) with respect to $q_j(\underline{z}_j)$ while keeping $\{q_i(\underline{z}_i) : i \neq j\}$ constant (a la Gauss-Seidel or coordinate descent iteration!). Very useful in control and inference applications.

$$-\ln L(q(\underline{z})) = -\int q(\underline{z}) \ln \left\{ \frac{p(\underline{x},\underline{z})}{q(\underline{z})} \right\} d\underline{z} = -\int \prod_{i=1}^{K} q_i(\underline{z}_i) \{\ln p(\underline{x},\underline{z})\} d\underline{z} - \sum_{i=1}^{K} H_{q_i}(\underline{z}_i)$$

$$= -\int q_j(\underline{z}_j) \underbrace{\left\{ \ln p(\underline{x},\underline{z}) \prod_{\substack{i=1 \\ i\neq j}}^{K} q_i(\underline{z}_i) d\underline{z}_i \right\}}_{E_{i\neq j}[\ln p(\underline{x},\underline{w})]} d\underline{z}_j - H_{q_j}(\underline{z}_j) - \sum_{\substack{i=1 \\ i\neq j}}^{K} H_{q_i}(\underline{z}_i)$$

$$\frac{\partial[-\ln L(q(\underline{z}))]}{\partial q_j(\underline{z}_j)} = -E_{i\neq j}[\ln p(\underline{x},\underline{z})] + 1 + \ln\left[q_j(\underline{z}_j) \right] = 0$$

$$\ln\left[q_j(\underline{z}_j) \right] \propto E_{i\neq j}[\ln p(\underline{x},\underline{z})]$$

$$\Rightarrow q_j(\underline{z}_j) = \frac{e^{E_{i\neq j}[\ln p(\underline{x},\underline{z})]}}{\int e^{E_{i\neq j}[\ln p(\underline{x},\underline{z})]} d\underline{z}_j}$$

Log of the optimal q_j is the expectation of the log of joint distribution with respect to all of the other factors $\{q_i(\underline{z}_i) : i \neq j\}$. This idea is used in loopy belief propagation and expectation propagation also.

Discuss applications to Gaussian mixture and logistic regression applications of variational Bayes.

Discuss Information-theoretic co-clustering

Discuss LVQ