Lecture 11: Mixture Models, EM, K-Means, Variational Bayes, LVQ & Information-theoretic Co-clustering

Prof. Krishna R. Pattipati
Dept. of Electrical and Computer Engineering
University of Connecticut
Contact: krishna@engr.uconn.edu **(860) 486-2890**

Spring 2021
April 23, 2021

Copyright © 2001-2021 K.R. Pattipati

Lecture Outline

- Mixture Models
- Expectation Maximization (EM)
- K-Means Algorithm
- Variational Bayes EM
- Variational Logistic Regression
- Information-Theoretic Co-clustering
- Learning Vector Quantization
- Summary

Copyright © 2001-2021 K.R. Pattipati

Why Gaussian Mixtures?

❑ Why Gaussian Mixtures?

- Parametric $\rightarrow$ fast but limited
- Non Parametric $\rightarrow$ general but slow (require lot of data)
- Mixture Models
  - RBF
  - Conditional Density Estimation (function approx.)
  - Mixture of experts models

$$p(\underline{x}) = \sum_{j=1}^{M} p(\underline{x} \mid j) P_j$$

$$\sum_{j=1}^{M} P_j = 1 \quad ; \; 0 \leq P_j \leq 1$$

$$\sum_{j=1}^{M} \int_{\underline{x}} p(\underline{x} \mid j) P_j \, d\underline{x} = 1$$

Similar technique for $p(.) = p(\underline{x} \mid k)$

$k = 1, 2, \ldots, C$

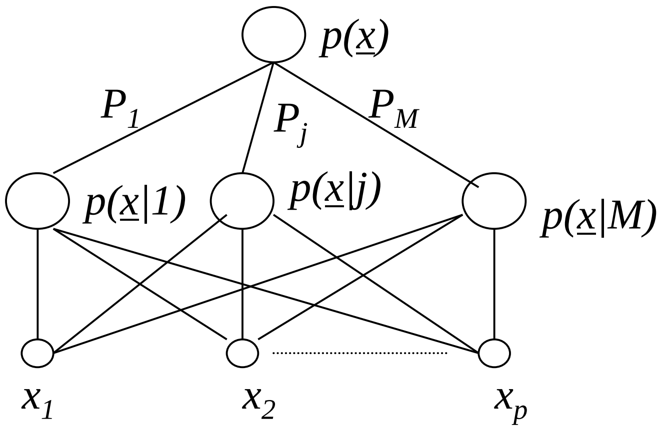
Copyright © 2001-2021 K.R. Pattipati

GMM Learning Problem

$$p(\underline{x} \mid j) = N(\underline{\mu}_j, \Sigma_j)$$

$$= N(\underline{\mu}_j, \sigma_j^2 I) \quad \text{typically}$$

$$= \frac{1}{(2\pi\sigma_j^2)^{p/2}} e^{\left(\frac{-\|(\underline{x}-\underline{\mu}_j)\|_2^2}{2\sigma_j^2} \right)}$$



Problem: Given data,

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

$$\text{Let } \underline{\theta} = \left\{ P_j, \underline{\mu}_j, \sigma_j \right\}$$

$$L = \max_{\underline{\theta}} p(D|\underline{\theta}) \quad \Rightarrow \max_{\underline{\theta}} l = \left[\ln p(D \mid \underline{\theta}) \right] \quad \Rightarrow \min_{\underline{\theta}} \left[-\ln p(D \mid \underline{\theta}) \right] = J$$

Copyright © 2001-2021 K.R. Pattipati

ML Estimation of GMM Parameters - 1

$$J = -\sum_{i=1}^{N} \ln\, p(\underline{x}^i, \underline{\theta}) = -\sum_{i=1}^{N} \ln\left(\sum_{j=1}^{M} p(\underline{x}^i \mid j) P_j\right)$$

$$\min \quad J$$

$$s.t. \sum_{j=1}^{M} P_j = 1; \quad 0 \le P_j \le 1$$

Lagrangian:

$$l = J + \lambda \sum_{i=1}^{M} P_j - \lambda$$

$$\frac{\partial J}{\partial \underline{\mu}_j} = -\sum_{i=1}^{N} \frac{1}{\sum_{k=1}^{M} p(\underline{x}^i|k) P_k} P_j \frac{\partial p(\underline{x}^i|j)}{\partial \underline{\mu}_j}$$

$$\frac{\partial p(\underline{x}^i|j)}{\partial \underline{\mu}_j} = -\frac{1}{(2\pi\sigma_j^2)^{p/2}} e^{\left(\frac{-\|(\underline{x}^i - \underline{\mu}_j)\|^2}{2\sigma_j^2}\right)} \frac{\underline{\mu}_j - \underline{x}^i}{\sigma_j^2} = -p(\underline{x}^i|j) \frac{\underline{\mu}_j - \underline{x}^i}{\sigma_j^2}$$

$$So,\ \frac{\partial J}{\partial \underline{\mu}_j} = \sum_{i=1}^{N} P(j|\underline{x}^i) \frac{\underline{\mu}_j - \underline{x}^i}{\sigma_j^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

posterior

Note the Simplicity of Gradient

Copyright © 2001-2021 K.R. Pattipati

ML Estimation of GMM Parameters - 2

$$\frac{\partial J}{\partial \underline{\sigma}_j} = -\sum_{i=1}^{N} \frac{1}{\sum_{k=1}^{M} p(\underline{x}^i|k)P_k} P_j \frac{\partial p(\underline{x}^i|j)}{\partial \sigma_j}$$

Dimension of feature vector

$$= \sum_{i=1}^{N} P(j\,|\,\underline{x}^i)\left\{\frac{p}{\sigma_j} - \frac{||\underline{x}^i - \underline{\mu}_j||^2}{\sigma_j^{\ 3}}\right\} \quad \text{. . . . .(2)}$$

$$\frac{\partial l}{\partial P_j} = -\sum_{i=1}^{N} \frac{1}{\sum_{k=1}^{M} p(\underline{x}^i\,|\,k)P_k} p(\underline{x}^i\,|\,j) + \lambda$$

$$= -\sum_{i=1}^{N} \frac{P(j|\underline{x}^i)}{P_j} + \lambda \quad \Rightarrow \frac{1}{P_j}\left[-\sum_{i=1}^{N} P(j|\underline{x}^i) + \lambda P_j\right] \quad \ldots\ldots\ldots\ldots\ldots(3)$$

Copyright © 2001-2021 K.R. Pattipati

ML Estimation of GMM = Coupled Nonlinear Equations

From (1), $$\hat{\underline{\mu}}_j = \frac{\sum_{i=1}^{N} P(j \mid \underline{x}^i)\underline{x}^i}{\sum_{i=1}^{N} P(j \mid \underline{x}^i)}$$

Necessary Conditions of Optimality:
Set Gradients Equal to Zero

$$\hat{\underline{\sigma}}_j^{\,2} = \frac{1}{p} \frac{\sum_{i=1}^{N} P(j \mid \underline{x}^i) \| \underline{x}^i - \hat{\underline{\mu}}_j \|^2}{\sum_{i=1}^{N} P(j \mid \underline{x}^i)}$$

General Case:

$$\Sigma_j = \frac{\sum_{i=1}^{N} P(j \mid \underline{x}^i)(\underline{x}^i - \hat{\underline{\mu}}_j)(\underline{x}^i - \hat{\underline{\mu}}_j)^T}{\sum_{i=1}^{N} P(j \mid \underline{x}^i)}$$

noting that, $\sum_{j=1}^{M} P(j \mid \underline{x}^i) = 1$ and $\sum_{j=1}^{M} P_j = 1$ we have $\lambda = \mathrm{N}$ $\Rightarrow \hat{P}_j = \frac{1}{N} \sum_{i=1}^{N} P(j \mid \underline{x}^i)$

These are coupled non-linear equations

Responsibility

Copyright © 2001-2021 K.R. Pattipati

Methods of Solution: NLP

- Nonlinear Programming (NLP) Techniques

$$\underline{\theta}_0 \rightarrow \underline{\theta}_1 \rightarrow \ldots\ldots\ldots\ldots \underline{\theta}^*$$

$$\underline{\theta}_{k+1} \rightarrow \underline{\theta}_k - \eta H \nabla_{\underline{\theta}} l$$

$$H = \begin{cases} I & \Rightarrow \text{SD or Gradient Method} \\ \left[\nabla^2 J\right]^{-1} & \Rightarrow \text{Newton's Method} \\ \left[\nabla^2 J + \varepsilon I\right]^{-1} & \Rightarrow \text{Levenberg-Marquardt Method} \\ \left[\nabla_{\underline{\theta}} J \nabla_{\underline{\theta}} J^T + \varepsilon I\right]^{-1} & \Rightarrow \text{Levenberg-Marquardt version of Gauss Newton Method} \\ & \text{Various versions of Quasi-Newton Method} \\ & \text{Various versions of Conjugate Gradient method} \end{cases}$$

Best to compute Hessian using finite Difference method

Copyright © 2001-2021 K.R. Pattipati

EM Algorithm

❑ EM Algorithm

Gauss-Seidel view of EM

M-step

$$\hat{\mu}_j^{new} = \frac{\sum_{i=1}^{N} \hat{P}^{old}(j \mid \underline{x}^i)\underline{x}^i}{\sum_{i=1}^{N} \hat{P}^{old}(j \mid \underline{x}^i)}$$

How did we get these equations and Why?…. Later

- By setting gradient to zero (M-step)
- Evaluating posterior Probabilities/Responsibilities (E-step)

$$\hat{\sigma}_j^{new^2} = \frac{1}{p} \frac{\sum_{i=1}^{N} \hat{P}^{old}(j \mid \underline{x}^i) \| \underline{x}^i - \underline{\hat{\mu}}_j^{new} \|^2}{\sum_{i=1}^{N} \hat{P}^{old}(j \mid \underline{x}^i)}$$

E-step

$$\hat{P}_j^{new} = \frac{1}{N} \sum_{i=1}^{N} \hat{P}^{old}(j \mid \underline{x}^i)$$

$$\hat{P}^{new}(j \mid \underline{x}^i) = \frac{p(\underline{x}^i \mid j)\hat{P}_j^{new}}{\sum_{m=1}^{M} p(\underline{x}^i \mid m)\hat{P}_m^{new}}$$

Copyright © 2001-2021 K.R. Pattipati

Sequential Estimation -1

❑ Sequential Estimation ~ Stochastic Approximation

$$\underline{\hat{\mu}}_j^{n+1} = \frac{\sum_{i=1}^{n+1} P(j \mid \underline{x}^i)\underline{x}^i}{\sum_{i=1}^{n+1} P(j \mid \underline{x}^i)}$$

$$= \frac{\sum_{i=1}^{n} P(j \mid \underline{x}^i)}{\sum_{i=1}^{n+1} P(j \mid \underline{x}^i)} \underline{\hat{\mu}}_j^{n} + \frac{P(j \mid \underline{x}^{n+1})}{\sum_{i=1}^{n+1} P(j \mid \underline{x}^i)} \underline{x}^{n+1}$$

$$= \underline{\hat{\mu}}_j^{n} + \underbrace{\frac{P(j \mid \underline{x}^{n+1})}{\sum_{i=1}^{n+1} P(j \mid \underline{x}^i)}}_{\eta_j^{n+1}} \left[\underline{x}^{n+1} - \underline{\hat{\mu}}_j^{n}\right]$$

Note:

$$\frac{1}{\eta_j^{n+1}} = \frac{\sum_{i=1}^{n+1} P(j \mid \underline{x}^i)}{P(j \mid \underline{x}^{n+1})} = 1 + \frac{\sum_{i=1}^{n} P(j \mid \underline{x}^i)}{P(j \mid \underline{x}^{n+1})}$$

$$= 1 + \frac{\sum_{i=1}^{n} P(j \mid \underline{x}^i)}{P(j \mid \underline{x}^{n+1})} \cdot \frac{P(j \mid \underline{x}^n)}{P(j \mid \underline{x}^n)}$$

$$= 1 + \frac{P(j \mid \underline{x}^n)}{P(j \mid \underline{x}^{n+1})} \cdot \frac{1}{\eta_j^n}$$

Copyright © 2001-2021 K.R. Pattipati

Sequential Estimation ~ Stochastic Approximation -2

- Sometimes replace, $\underline{\eta}_j^{n+1} = \dfrac{P(j\,|\,\underline{x}^{n+1})}{(n+1)\hat{P}_j^{n+1}}$ or, $\dfrac{1}{\underline{\eta}_j^{n+1}} = \dfrac{P(j\,|\,\underline{x}^{n})}{P(j\,|\,\underline{x}^{n+1})}\dfrac{1}{\underline{\eta}_j^{n}} + 1$

$$\boxed{\underline{\eta}_j^{n+1} = \frac{\underline{\eta}_j^{n} P(j\,|\,\underline{x}^{n+1})}{\underline{\eta}_j^{n} P(j\,|\,\underline{x}^{n+1}) + P(j\,|\,\underline{x}^{n})}}$$

Similarly, $\hat{\underline{\sigma}}_j^{2\,n} = \dfrac{1}{p}\dfrac{\sum_{i=1}^{n} P(j\,|\,\underline{x}^i)\,\|\,\underline{x}^i - \hat{\underline{\mu}}_j^{n}\,\|^2}{\sum_{i=1}^{n} P(j\,|\,\underline{x}^i)}$

$$\hat{\underline{\sigma}}_j^{2\,n+1} = \frac{1}{p}\frac{\sum_{i=1}^{n+1} P(j\,|\,\underline{x}^i)\,\|\,\underline{x}^i - \hat{\underline{\mu}}_j^{n+1}\,\|^2}{\sum_{i=1}^{n+1} P(j\,|\,\underline{x}^i)} = \frac{1}{p}\frac{\sum_{i=1}^{n+1} P(j\,|\,\underline{x}^i)\,\|\,\underline{x}^i + \hat{\underline{\mu}}_j^{n} - \hat{\underline{\mu}}_j^{n} - \hat{\underline{\mu}}_j^{n+1}\,\|^2}{\sum_{i=1}^{n+1} P(j\,|\,\underline{x}^i)}$$

$$= \frac{1}{p}\frac{\sum_{i=1}^{n+1} P(j\,|\,\underline{x}^i)\left\{\|\,\underline{x}^i - \hat{\underline{\mu}}_j^{n}\,\|^2 + 2\left(\underline{x}^i - \hat{\underline{\mu}}_j^{n}\right)^T\left(\hat{\underline{\mu}}_j^{n} - \hat{\underline{\mu}}_j^{n+1}\right) + \|\,\hat{\underline{\mu}}_j^{n} - \hat{\underline{\mu}}_j^{n+1}\,\|^2\right\}}{\sum_{i=1}^{n+1} P(j\,|\,\underline{x}^i)}$$

$$= \hat{\underline{\sigma}}_j^{2\,n}\frac{\sum_{i=1}^{n} P(j\,|\,\underline{x}^i)}{\sum_{i=1}^{n+1} P(j\,|\,\underline{x}^i)} + \frac{1}{p}\frac{P(j\,|\,\underline{x}^{n+1})}{\sum_{i=1}^{n+1} P(j\,|\,\underline{x}^i)}\left(\|\,\underline{x}^{n+1} - \hat{\underline{\mu}}_j^{n}\,\|^2\right) - \frac{1}{p}\left(\|\,\hat{\underline{\mu}}_j^{n} - \hat{\underline{\mu}}_j^{n+1}\,\|^2\right)$$

Copyright © 2001-2021 K.R. Pattipati

Sequential Estimation ~ Stochastic Approximation -3

$$= \underline{\hat{\sigma}}_j^{2\,n} + \frac{P(j\,|\,\underline{x}^{n+1})}{\sum_{i=1}^{n+1} P(j\,|\,\underline{x}^{i})}\left(\frac{\|\,\underline{x}^{n+1} - \underline{\hat{\mu}}_j^{n}\,\|^2}{p} - \underline{\hat{\sigma}}_j^{2\,n}\right) - \frac{1}{p}\left(\|\,\underline{\hat{\mu}}_j^{n} - \underline{\hat{\mu}}_j^{n+1}\,\|^2\right)$$

$$\boxed{\underline{\hat{\sigma}}_j^{2\,n+1} = \underline{\hat{\sigma}}_j^{2\,n} + \eta_j^{n+1}\left[\frac{1}{p}\|\,\underline{x}^{n+1} - \underline{\hat{\mu}}_j^{n}\,\|^2 - \underline{\hat{\sigma}}_j^{2\,n}\right] - \frac{1}{p}\left(\|\,\underline{\hat{\mu}}_j^{n} - \underline{\hat{\mu}}_j^{n+1}\,\|^2\right)}$$

$$\boxed{= \underline{\hat{\sigma}}_j^{2\,n} + \eta_j^{n+1}\left[\frac{1}{p}\left(1 - \eta_j^{n+1}\right)\|\,\underline{x}^{n+1} - \underline{\hat{\mu}}_j^{n}\,\|^2 - \underline{\hat{\sigma}}_j^{2\,n}\right]}$$

Re*call*

$$\underline{\hat{\mu}}_j^{n+1} - \underline{\hat{\mu}}_j^{n} = \eta_j^{n+1}\left[\underline{x}^{n+1} - \underline{\hat{\mu}}_j^{n}\right]$$

- Similarly,

$$\boxed{\hat{P}_j^{n+1} = \hat{P}_j^{n} + \frac{1}{n+1}\left[P(j\,|\,\underline{x}^{n+1}) - \hat{P}_j^{n}\right]}$$

Copyright © 2001-2021 K.R. Pattipati

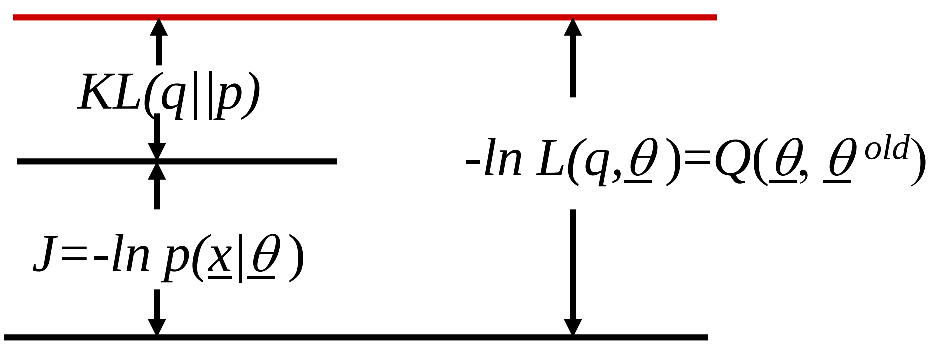
Probabilistic Interpretation of EM

❑ Key ideas of EM as applied to Gaussian Mixture Problem

$$J = -\sum_{i=1}^{N} \ln p(\underline{x}^i) \quad = -\sum_{i=1}^{N} \ln\left(\sum_{j=1}^{M} p(\underline{x}^i \mid j) P_j \right)$$

$$J^{new} - J^{old} = -\sum_{i=1}^{N} \ln\left[\frac{p^{new}(\underline{x}^i)}{p^{old}(\underline{x}^i)} \right]$$

$$= -\sum_{i=1}^{N} \ln\left[\frac{\sum_{j=1}^{M} P_j^{new} p^{new}(\underline{x}^i | j) \frac{P^{old}(j|\underline{x}^i)}{P^{old}(j|\underline{x}^i)}}{p^{old}(\underline{x}^i)} \right]$$



Idea:

$\underline{x}$: data

$\underline{z}$: hidden variables (mixture)

$\underline{\theta}$: parameters

$q(\underline{z})$ = any arbitrary distribution

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

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

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

$$J \le -\ln L(q, \underline{\theta}) \because KL(q(\underline{z}) \| p(\underline{z} \mid \underline{x}, \underline{\theta})) \ge 0$$

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

$$M-step: \underline{\theta}^{new} = \min_{\underline{\theta}}[-\ln L(q, \underline{\theta})]$$

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

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

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

Copyright © 2001-2021 K.R. Pattipati

EM Algorithm as a Minimization of a Bound on NLL

- For convex functions,

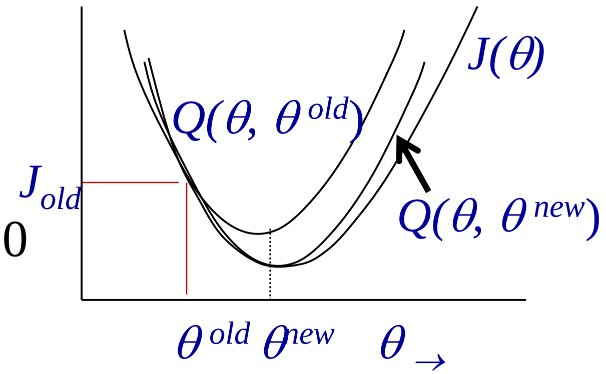
$$-\ln\left[\sum \lambda_i x_i\right] \le -\sum \lambda_i \ln x_i \qquad \text{where } \sum \lambda_i = 1,\ \lambda_i \ge 0$$

$$\Rightarrow J^{new} - J^{old} \le -\sum_{i=1}^{N}\sum_{j=1}^{M} P^{old}(j\,|\,\underline{x}^i)\ln\left[\frac{P_j^{new}\,p^{new}(\underline{x}^i\,|\,j)}{p^{old}(\underline{x}^i)\ P^{old}(j\,|\,\underline{x}^i)}\right]$$

$$= -\sum_{i=1}^{N}\sum_{j=1}^{M} P^{old}(j\,|\,\underline{x}^i)\ln\left[\frac{p^{new}(\underline{x}^i, j)}{p^{old}(\underline{x}^i, j)}\right]$$

$$\Rightarrow J^{new} \le -\sum_{i=1}^{N}\sum_{j=1}^{M} \underbrace{P^{old}(j\,|\,\underline{x}^i)}_{q(\underline{z},\theta^{old})}\ln p^{new}(\underline{x}^i, j) = Q(\theta,\theta^{old}) = -\ln L(q,\theta)$$

$\Rightarrow$ Minimizing l will lead to a decrease in $J(\theta)$



$$Note: at\ \theta^{old},\ J^{old}(\theta^{old}) = Q(\theta^{old},\theta^{old}) \Rightarrow Force\ KL = 0$$

$$J^{new}(\theta^{new}) \le Q(\theta^{new},\theta^{old}) \Rightarrow KL \ge 0$$

$Q(\theta,\theta^{old})$ and $J(\theta)$ have the same gradient at θ^{old}

Copyright © 2001-2021 K.R. Pattipati

Lower Bound Optimization Problem (M-step)

Dropping terms that depend on old parameters, we get

$$Q = -\sum_{i=1}^{N}\sum_{j=1}^{M} P^{old}(j \mid \underline{x}^i)\ln\left[P_j^{new}\, p^{new}(\underline{x}^i \mid j)\right] = -\sum_{i=1}^{N}\sum_{j=1}^{M} P^{old}(j \mid \underline{x}^i)\ln\left[p^{new}(\underline{x}^i, j)\right]$$

For Gaussian conditional probability density functions

$$Q = -\sum_{i=1}^{N}\sum_{j=1}^{M} P^{old}(j \mid \underline{x}^i)\left\{\ln P_j^{new} - p\ln\sigma_j^{new} - \frac{\|\underline{x}^i - \underline{\mu}_j^{new}\|^2}{2\sigma_j^{2\,new}}\right\}$$

- Optimization problem:
  - min Q
  - s.t. $\sum_{j=1}^{M} P_j^{new} = 1; \qquad P_j^{new} \geq 0; \qquad j = 1,2,\ldots\ldots..M$

Copyright © 2001-2021 K.R. Pattipati

Solution of the M-Step

$$\underline{\mu}_j^{new} = \frac{\sum_{i=1}^{N} P^{old}(j|\underline{x}^i)\underline{x}^i}{\sum_{i=1}^{N} P^{old}(j|\underline{x}^i)}$$

$$\sigma_j^{2\,new} = \frac{1}{p}\frac{\sum_{i=1}^{N} P^{old}(j\,|\,\underline{x}^i)\,\|\,\underline{x}^i - \underline{\mu}_j^{new}\,\|^2}{\sum_{i=1}^{N} P^{old}(j\,|\,\underline{x}^i)}$$

General Case:

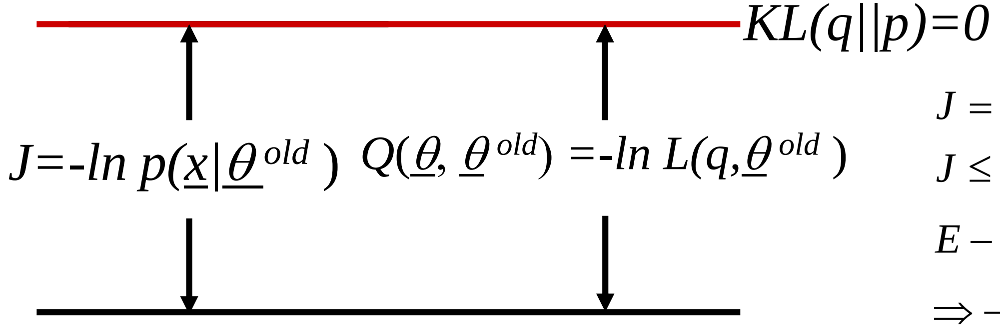
$$\Sigma_j^{new} = \frac{\sum_{i=1}^{N} P^{old}(j\,|\,\underline{x}^i)(\underline{x}^i - \underline{\hat{\mu}}_j^{new})(\underline{x}^i - \underline{\hat{\mu}}_j^{new})^T}{\sum_{i=1}^{N} P^{old}(j\,|\,\underline{x}^i)}$$

$$P_j^{new} = \frac{1}{N}\sum_{i=1}^{N} P^{old}(j\,|\,\underline{x}^i)$$

Copyright © 2001-2021 K.R. Pattipati

Graphical Illustration of E and M Steps

- **E-step**



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

$$J \le -\ln L(q,\underline{\theta}) \because KL(q(\underline{z}) \,||\, p(\underline{z}\,|\,\underline{x},\underline{\theta})) \ge 0$$

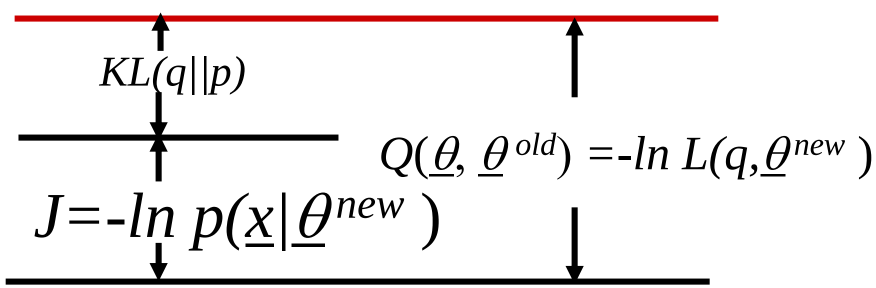
$$E-step: q(\underline{z}) = p(\underline{z}\,|\,\underline{x},\underline{\theta}^{old})$$

$$\Rightarrow -\ln L(q,\underline{\theta}^{old}) = -\ln p(\underline{x}\,|\,\underline{\theta}^{old})$$

Why?

$$\because KL(q(\underline{z}) \,||\, p(\underline{z}\,|\,\underline{x},\underline{\theta}^{old})) = 0$$

- **M-step**



$$M-step: \underline{\theta}^{new} = \arg\min_{\underline{\theta}}[-\ln L(q,\underline{\theta})]$$

$$\Rightarrow -\ln L(q,\underline{\theta}^{new}) \ge -\ln p(\underline{x}\,|\,\underline{\theta}^{new})$$

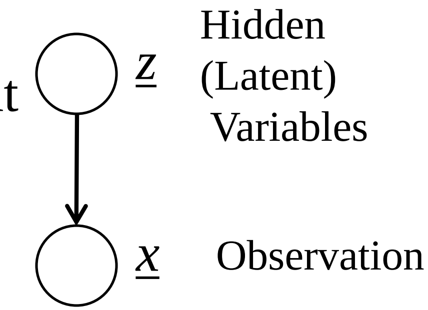
why?

$$\because KL(q(\underline{z}) = p(\underline{z}\,|\,\underline{x},\underline{\theta}^{old}) \,||\, p(\underline{z}\,|\,\underline{x},\underline{\theta}^{new})) \ge 0$$

Note: EM is a Maximum Likelihood Algorithm. Is there a Bayesian Version? Yes: If you assume priors on $(\{\mu_j, \sigma_j^2, P_j\})$ called *Variational Bayesian Inference.*

Copyright © 2001-2021 K.R. Pattipati

An Alternate View of EM for Gaussian Mixtures - 1



- $\underline{z}$ is a M-dimensional binary random vector such that

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

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

- $\underline{x}$ is a p-dimensional random vector such that

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

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

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

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

only possible $\underline{z}$ vectors:

$\underline{z} \in \{\underline{e}_i : i = 1, 2, .., M\}$

$\underline{e}_i = i^{th}$ unit vector

Copyright © 2001-2021 K.R. Pattipati

An Alternate View of EM for Gaussian Mixtures - 2

- If have several observations $\{\underline{x}^n: n=1,2,..,N\}$, each data point will have a corresponding latent vector $\underline{z}_n$.

Note the generality

Problem: Given incomplete (partial) data,

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

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

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

Complete Data:

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

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

Copyright © 2001-2021 K.R. Pattipati

An Alternate View of EM for Gaussian Mixtures - 3

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

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

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

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

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

Copyright © 2001-2021 K.R. Pattipati

EM Algorithm for Gaussian Mixtures -4

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

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

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

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

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

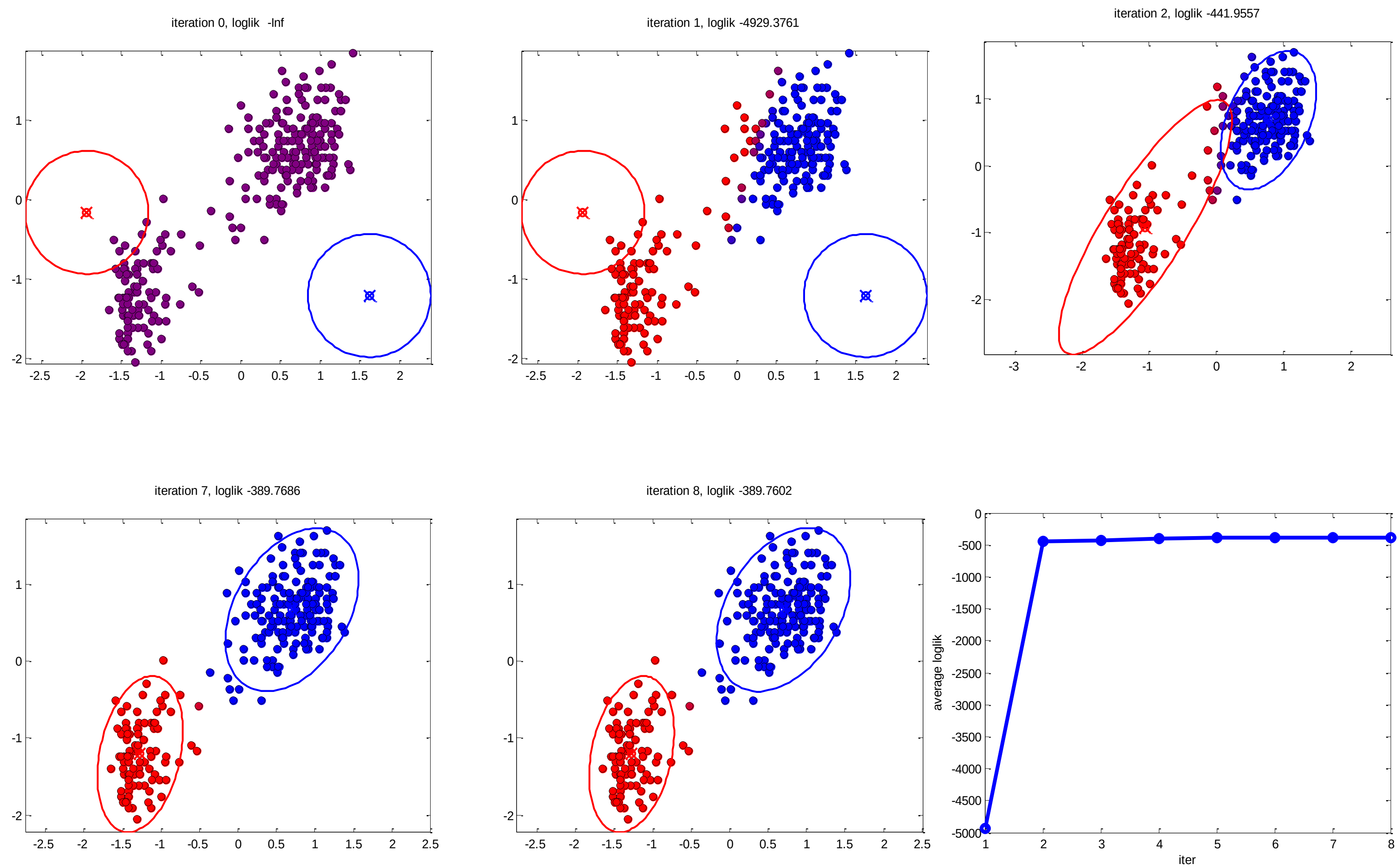
For unbiased estimate of covariance,
Divide by $\dfrac{1}{(N_j - \dfrac{\sum_{j=1}^{N}(\gamma_j^n)^2}{N_j})}$

Goes to $1/(N_j - 1)$ for (0-1) case

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

Copyright © 2001-2021 K.R. Pattipati

Illustration of EM Algorithm for Gaussian Mixtures



Murphy, Page 353, mixGaussDemoFaithful

Copyright © 2001-2021 K.R. Pattipati

Relation of Gaussian Mixtures to K - means

Suppose $\Sigma_j = \varepsilon I$ for $j = 1, 2, .., M$

Then

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

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

As $\varepsilon \to 0$

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

The expected value of negative log likelihood of complete-data is

$$E_{\underline{Z}}\{-\ln p(D_c \mid \underline{\theta})\} = \frac{1}{2\varepsilon} \sum_{n=1}^{N} \sum_{j=1}^{M} \gamma_j^n \| \underline{x}^n - \underline{\mu}_j \|^2 + \text{constant}$$

So, K-means minimizes $\frac{1}{2} \sum_{n=1}^{N} \sum_{j=1}^{M} \gamma_j^n \| \underline{x}^n - \underline{\mu}_j \|^2$

Copyright © 2001-2021 K.R. Pattipati

K-Means Algorithm

- *K-means clustering to select K and the centers*

a. Initialization

- Choose initial center at random. Let n_1 be the data point.
- For $k = 2,..,K$

For $n=1,2,..,N$ & $n \neq n_i$, $i=1,2,..,k-1$

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

End

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

b. For $n=1, 2, \ldots, N$

Assign n to cluster $\boldsymbol{C}_j$ 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

Copyright © 2001-2021 K.R. Pattipati

Model Selection

- BIC

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

- Prediction Error

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

- Excess Kurtosis-based Measure

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

- Knee or kink in the squared reconstruction error on a test set

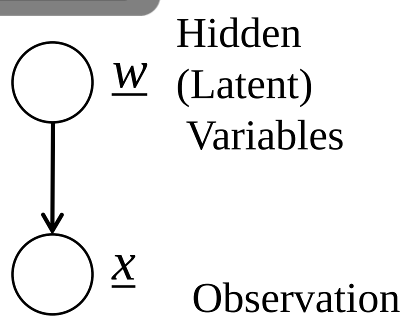
$$J(D,K) = \frac{1}{|D|}\sum_{i\in D}\|\,\underline{x}_i - \hat{\underline{x}}_i\,\|_2^2$$

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

Copyright © 2001-2021 K.R. Pattipati

Variational Bayesian Inference - 1

- $\underline{w}$ is a latent vector (continuous or discrete)
  - Mixture vector (discrete), $\underline{z}$
  - Parameters $(\{\mu_j, \Sigma_j, P_j\})$



- $\underline{x}$ is a p-dimensional random vector
- Recall

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

$$-\ln L(q(\underline{w})) = -\int q(\underline{w}) \ln \left\{ \frac{p(\underline{x}, \underline{w})}{q(\underline{w})} \right\} d\underline{w} = -E_{q(\underline{w})}\left(\ln p(\underline{x}, \underline{w})\right) - H_q(\underline{w})$$

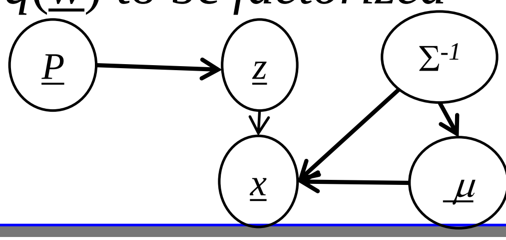
$$KL(q(\underline{w}) \,\|\, p(\underline{w} \mid \underline{x})) = -\int q(\underline{w}) \ln \left\{ \frac{p(\underline{w} \mid \underline{x})}{q(\underline{w})} \right\} d\underline{w} = -E_{q(\underline{w})}\left(\ln p(\underline{w} \mid \underline{x})\right) - H_q(\underline{w})$$

$$J = -\ln p(\underline{x}) \le -\ln L(q(\underline{w})) \because KL(q(\underline{w}) \,\|\, p(\underline{w} \mid \underline{x})) \ge 0$$

- Variational inference typically assumes $q(\underline{w})$ to be *factorized*

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

Example: $q(\underline{w}) = q(\underline{z})\, q(\{\underline{\mu}_j, \Sigma_j, P_j\})$



Copyright © 2001-2021 K.R. Pattipati

Variational Bayesian Inference - 2

- Minimize the upper bound -ln$L(q(\underline{w}))$ with respect to q_j ($\underline{w}_j$) **while keeping {q_i ($\underline{w}_i$) : $i \neq j$} constant** (*a la* Gauss-Seidel)

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

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

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

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

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

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 ($\underline{w}_i$) : $i \neq j$}. This idea is used in loopy belief propagation and expectation propagation also.

- Iterative algorithm for finding the factors {q_j ($\underline{w}_j$)}

Copyright © 2001-2021 K.R. Pattipati

Application to Gaussian Mixtures - 1

- Here $\underline{w}$ involves mixture variables and component parameters

$$q(\underline{w}) = q(\underline{z})\, q(\{\underline{\mu}_j, \Sigma_j, P_j\}_{j=1}^{M})$$

$\underline{z}$ is a binary random vector of dimension M

- Model assumptions

Mixture Distribution: $p(\{\underline{z}^n\}_{n=1}^{N} \mid \{P_j\}_{j=1}^{M}) = \prod_{n=1}^{N}\prod_{j=1}^{M} P_j^{z_j^n}$

Data Likelihood given latent variables:

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

We also assume *priors* on $\{P_j, \underline{\mu}_j, \Sigma_j\}_{j=1}^{M} \Rightarrow$ Bayesian approach

$$p(\underline{P}) = Dirichlet(\underline{P} \mid \underline{\alpha}) = \frac{\Gamma(M\alpha_0)}{\left(\Gamma(\alpha_0)\right)^M}\prod_{j=1}^{M} P_j^{\alpha_0 - 1};\ \text{conjugate prior to multinomial}$$

$$\boxed{\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt;\ \Gamma(\alpha+1) = \alpha\,\Gamma(\alpha);\ \Gamma(n) = (n-1)!\ \text{for integers}}$$

Copyright © 2001-2021 K.R. Pattipati

Application to Gaussian Mixtures - 3

- Variational Bayes M-step (VBM-step)…. It is easier to see M-step first

$$\ln q(\{P_j,\underline{\mu}_j,\Sigma_j\}_{j=1}^M) = E_{q(\{\underline{z}^n\}_{n=1}^N)}\left(\ln p(\{\underline{x}^n\}_{n=1}^N,\{\underline{z}^n\}_{n=1}^N,\{P_j,\underline{\mu}_j,\Sigma_j\}_{j=1}^M)\right)$$

$$= E_{q(\{\underline{z}^n\}_{n=1}^N)}\left(\begin{array}{l}\ln \prod_{n=1}^N\prod_{j=1}^M [P_j N(\underline{x}^n;\underline{\mu}_j,\Sigma_j)]^{z_j^n}\ . \\ \dfrac{\Gamma(M\alpha_0)}{(\Gamma(\alpha_0))^M}\left(\prod_{j=1}^M P_j^{\alpha_0-1}.N(\underline{\mu}_j;\underline{m}_0,\dfrac{1}{\beta_0}\Sigma_j).\ Wishart(\Sigma_j^{-1};\nu_0,W_0)\right)\end{array}\right)+cons\,t\,.$$

$$=\left\{\sum_{J=1}^M[(\alpha_0-1)+\underbrace{\sum_{n=1}^N E\left[z_j^n\right]}_{N_j}]\ln P_j\right\}+\sum_{j=1}^M\left\{\begin{array}{l}\ln N(\underline{\mu}_j;\underline{m}_0,\dfrac{1}{\beta_0}\Sigma_j) \\ +\ln Wishart(\Sigma_j^{-1};\nu_0,W_0)\end{array}\right\}$$

$$+\sum_{n=1}^N\sum_{j=1}^M \underbrace{E\left[z_j^n\right]}_{\gamma_j^n}\ln N(\underline{x}^n;\underline{\mu}_j,\Sigma_j)+cons\tan t$$

$$\boxed{\begin{array}{l} q(\{P_j,\underline{\mu}_j,\Sigma_j\}_{j=1}^M)=q(\underline{P})\,q(\{\underline{\mu}_j,\Sigma_j\}_{j=1}^M) \\ q(\underline{P})=Dirichlet(\underline{P};\{\alpha_0+N_j=\alpha_j\}_{j=1}^M) \\ q(\underline{\mu}_j,\Sigma_j)=Gaussian-Wishart\end{array}}$$

Copyright © 2001-2021 K.R. Pattipati

Application to Gaussian Mixtures - 4

- Updated factorized distribution after M-step

$$q(\{P_j,\underline{\mu}_j,\Sigma_j\}_{j=1}^{M}) = q(\underline{P})\,q(\{\underline{\mu}_j,\Sigma_j\}_{j=1}^{M})$$

$$q(\underline{P}) = Dirichlet(\underline{P};\{\alpha_0+N_j=\alpha_j\}_{j=1}^{M}) \Rightarrow E(P_j) = \frac{\alpha_j}{\sum_{k=1}^{M}\alpha_k} = \frac{\alpha_0+N_j}{M\alpha_0+N}$$

$$q(\underline{\mu}_j,\Sigma_j) = Gaussian-Wishart$$

$$=N(\underline{\mu}_j;\underline{m}_j,\frac{1}{\beta_j}\Sigma_j).Wishart(\Sigma_j^{-1};\nu_j,W_j)$$

$$\beta_j=\beta_0+N_j;\ N_j=\sum_{n=1}^{N}\gamma_j^n$$

$$\underline{m}_j = \frac{1}{\beta_j}\left(\beta_0\underline{m}_0+N_j\bar{\underline{x}}_j\right);\ \bar{\underline{x}}_j = \frac{1}{N_j}\sum_{n=1}^{N}\gamma_j^n\underline{x}^n$$

$$W_j^{-1}=W_0^{-1}+N_jS_j+\frac{\beta_0N_j}{\beta_0+N_j}(\bar{\underline{x}}_j-\underline{m}_0)(\bar{\underline{x}}_j-\underline{m}_0)^T$$

$$\text{where } S_j=\frac{1}{N_j}\sum_{n=1}^{N}\gamma_j^n(\underline{x}^n-\bar{\underline{x}}_j)(\underline{x}^n-\bar{\underline{x}}_j)^T$$

$$\nu_j=\nu_0+N_j$$

Updates for $\{N_j,\bar{\underline{x}}_j,S_j\}$ are similar to ML

Sequential VBEM?

Copyright © 2001-2021 K.R. Pattipati

Application to Gaussian Mixtures - 5

- Variational Bayes E-step (VBE-step)

$$\ln q(\{\underline{z}^n\}_{n=1}^N) = E_{q(\{P_j,\underline{\mu}_j,\Sigma_j\}_{j=1}^M)}\left(\ln p(\{\underline{x}^n\}_{n=1}^N,\{\underline{z}^n\}_{n=1}^N,\{P_j,\underline{\mu}_j,\Sigma_j\}_{j=1}^M)\right)$$

$$= E_{q(\{P_j,\underline{\mu}_j,\Sigma_j\}_{j=1}^M)}\left(\ln \prod_{n=1}^N\prod_{j=1}^M [P_j N(\underline{x}^n;\underline{\mu}_j,\Sigma_j)]^{z_j^n}\right) + constant$$

$$= E_{q(\{\underline{\mu}_j,\Sigma_j\}_{j=1}^M)} \sum_{n=1}^N\sum_{j=1}^M z_j^n \left(\ln \underbrace{p(\{\underline{x}^n\}_{n=1}^N \mid \{\underline{z}^n\}_{n=1}^N,\{\underline{\mu}_j,\Sigma_j\}_{j=1}^M)}_{N(\underline{x}^n;\underline{\mu}_j,\Sigma_j)}\right) +$$

$$E_{q\{\underline{P}\}}\left(\ln \underbrace{p(\{\underline{z}^n\}_{n=1}^N \mid \{P_j\}_{j=1}^M)}_{P_j}\right) + constant$$

$$= \sum_{n=1}^N\sum_{j=1}^M z_j^n \ln \rho_j^n$$

where $\ln\rho_j^n = E_{P_j}\left[\ln P_j\right] + \frac{1}{2}E_{\Sigma_j^{-1}}\left[\ln|\Sigma_j^{-1}|\right] - \frac{p}{2}\ln(2\pi) - \frac{1}{2}E_{\underline{\mu}_j,\Sigma_j^{-1}}\left(\|\underline{x}^n-\underline{\mu}_j\|^2_{\Sigma_j^{-1}}\right)$

$\Rightarrow q(\{\underline{z}^n\}_{n=1}^N) = \prod_{n=1}^N\prod_{j=1}^M [\gamma_j^n]^{z_j^n}$ where $\gamma_j^n = \dfrac{\rho_j^n}{\sum_{k=1}^M \rho_k^n}$ .... responsibilities

Copyright © 2001-2021 K.R. Pattipati

Application to Gaussian Mixtures - 6

- Variational Bayes E-step (VBE-step) … continued
  - Evaluation of responsibilities
  - Recall

$$\ln \rho_j^n = E_{P_j}\left[\ln P_j\right] + \frac{1}{2} E_{\Sigma_j^{-1}}\left[\ln |\Sigma_j^{-1}|\right] - \frac{p}{2}\ln(2\pi) - \frac{1}{2} E_{\underline{\mu}_j, \Sigma_j^{-1}}\left(\| \underline{x}^n - \underline{\mu}_j \|^2_{\Sigma_j^{-1}}\right)$$

$$E_{P_j}\left[\ln P_j\right] = \psi(\alpha_j) - \psi(\sum_{k=1}^{M} \alpha_k);\ \psi(\alpha) = \frac{d}{d\alpha}\ln\Gamma(\alpha)\text{....}\textit{digamma}\text{ function}$$

$$E_{\Sigma_j^{-1}}\left[\ln |\Sigma_j^{-1}|\right] = \sum_{i=1}^{p} \psi(\frac{\nu_j + 1 - i}{2}) + p\ln 2 + \ln |W_j|$$

See Bishop Chapter 10

$$E_{\underline{\mu}_j, \Sigma_j^{-1}}\left(\| \underline{x}^n - \underline{\mu}_j \|^2_{\Sigma_j^{-1}}\right) = \frac{p}{\beta_j} + \nu_j (\underline{x}^n - \underline{m}_j)^T W_j (\underline{x}^n - \underline{m}_j)$$

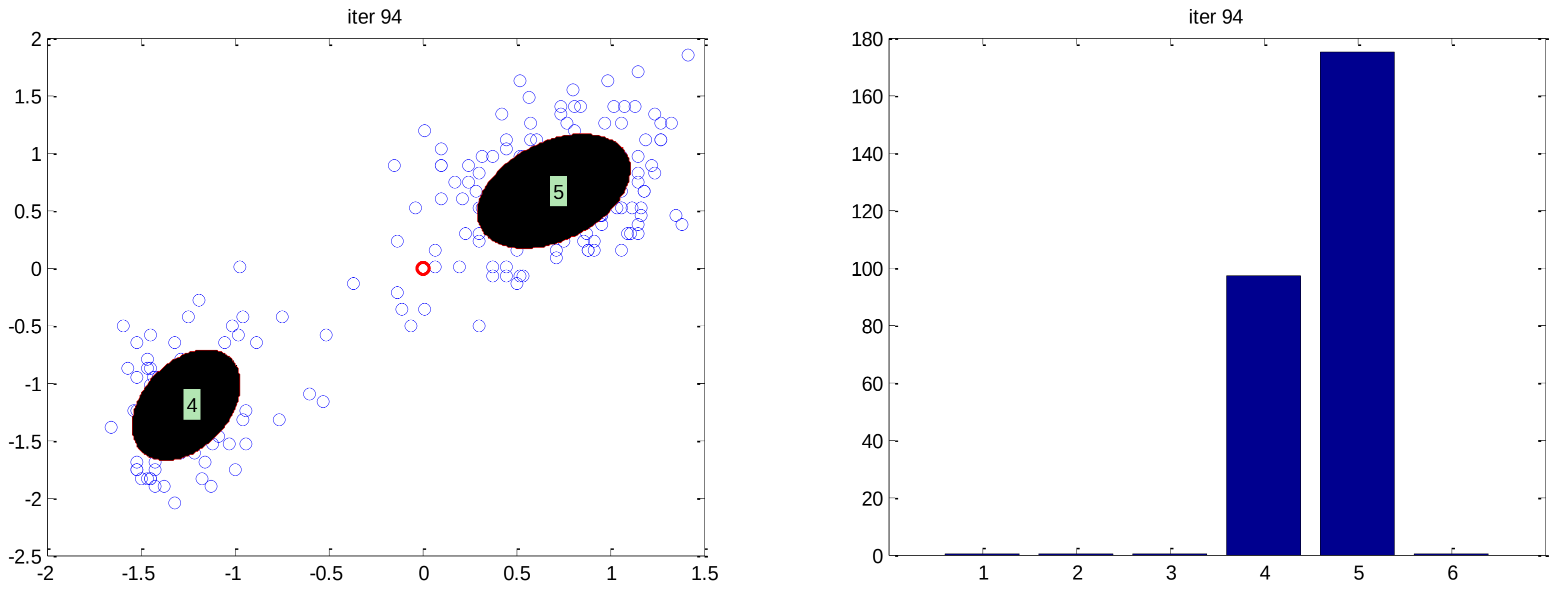
$\textit{Since } \gamma_j^n \propto \rho_j^n$

$$\gamma_j^n \propto |W_j| \exp\left(\psi(\alpha_j) + \frac{1}{2}\sum_{i=1}^{p}\psi(\frac{\nu_j + 1 - i}{2}) - \frac{p}{2\beta_j} - \frac{\nu_j}{2}(\underline{x}^n - \underline{m}_j)^T W_j (\underline{x}^n - \underline{m}_j)\right)$$

Copyright © 2001-2021 K.R. Pattipati

VB Approximation of Gaussian Gamma

- In VBEM start with large M and very small α_0 <<1 (≈0.001)
- It automatically prunes clusters with very few members ("rich get richer")
- In this example, we start with 6 clusters, but only 2 remain at the end



mixGaussVbDemoFaithful from Murphy, Page 755

Copyright © 2001-2021 K.R. Pattipati

Variational Logistic Regression -1

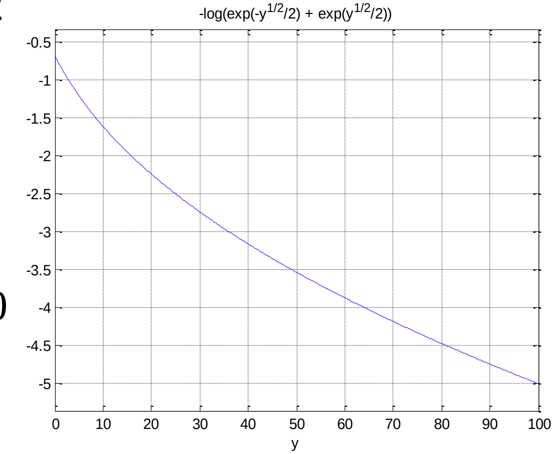
- Lower bound on sigmoid function

$$Consider\ \ln g(x) = -\ln(1+e^{-x}) = -\ln[e^{-x/2}(e^{x/2}+e^{-x/2})] = \frac{x}{2} - \ln(e^{x/2}+e^{-x/2})$$

$$f(x) = -\ln(e^{x/2}+e^{-x/2})\ is\ convex\ in\ x^2. Why?$$

$$Let\ x=\sqrt{y} \Rightarrow f(y) = -\ln(e^{\sqrt{y}/2}+e^{-\sqrt{y}/2}); y \geq 0$$

$$\frac{df}{dy} = \frac{-1}{4\sqrt{y}}\tanh(\frac{\sqrt{y}}{2}) < 0;\ \frac{d^2 f}{dy^2} = \frac{\frac{\sinh(\sqrt{y})}{y^{3/2}} - \frac{1}{y}}{8(\cosh(\sqrt{y})+1)} > 0\ \forall y \geq 0$$



$Recall\ for\ convex\ functions$

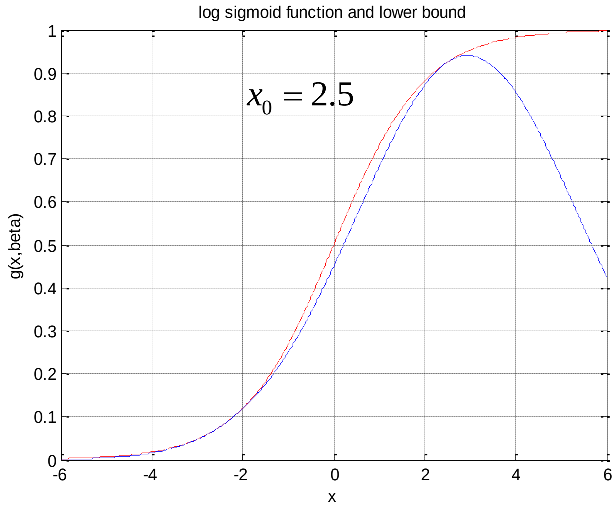
$$f(y) \geq f(y_0) + \frac{df}{dy}|_{y=y_0}(y-y_0)\ \forall y_0 = x_0^2$$

$$So,\ f(y) \geq -\ln(e^{\sqrt{y_0}/2}+e^{-\sqrt{y_0}/2}) - \frac{1}{4\sqrt{y_0}}\tanh(\frac{\sqrt{y_0}}{2})(y-y_0)$$

$$\Rightarrow f(x) \geq -\ln(e^{x_0/2}+e^{-x_0/2}) - \underbrace{\frac{1}{4x_0}\tanh(\frac{x_0}{2})}_{\lambda(x_0)}(x^2-x_0^2)$$

$$\ln g(x) \geq \frac{x-x_0}{2} + \underbrace{\frac{x_0}{2} - \ln(e^{x_0/2}+e^{-x_0/2})}_{\ln g(x_0)} - \lambda(x_0)(x^2-x_0^2)$$

$$So, g(x) \geq g(x_0)\exp\{\frac{x-x_0}{2} - \lambda(x_0)(x^2-x_0^2)\}$$



Copyright © 2001-2021 K.R. Pattipati

Variational Logistic Regression - 2

- Binary (Two class) Case using local variational lower bound

Posterior distribution of z for a given x

$$y = \underline{w}^T \underline{x} \text{ or } y = \underline{w}^T \phi(\underline{x})$$

$$p(z \mid \underline{w}) = g(y)^z (1-g(y))^{1-z} = \left(\frac{1}{1+e^{-y}}\right)^z \left(\frac{e^{-y}}{1+e^{-y}}\right)^{1-z}$$

$$= e^{yz}\left(\frac{e^{-y}}{1+e^{-y}}\right) = e^{yz}\left(\frac{1}{1+e^{y}}\right) = e^{yz} g(-y)$$

$$\text{Re}\,call\ \ g(y) \geq g(y_0)\exp\{(y-y_0)/2 - \lambda(y_0)(y^2 - y_0^2)\}$$

$$\lambda(y_0) = \frac{1}{4y_0}\tanh(\frac{y_0}{2}) = \frac{1}{4y_0}\left(\frac{1-e^{-y_0}}{1+e^{-y_0}}\right) = \frac{1}{2y_0}[\frac{1}{2} - g(-y_0)] = \frac{1}{2y_0}[g(y_0) - \frac{1}{2}]$$

$$g(-y) \geq g(y_0)\exp\{-(y+y_0)/2 - \lambda(y_0)(y^2 - y_0^2)\}\forall y_0$$

$$p(z \mid \underline{w}) = e^{yz} g(-y) \geq e^{yz} g(y_0)\exp\{-(y+y_0)/2 - \lambda(y_0)(y^2 - y_0^2)\}$$

$$-\ln p(z \mid \underline{w}) \leq -z\underline{w}^T \underline{x} - \ln g(y_0) + (\underline{w}^T \underline{x} + y_0)/2 + \lambda(y_0)[(\underline{w}^T \underline{x})^2 - y_0^2]$$

Quadratic function in $\underline{w} \Rightarrow$ Gaussian posterior

$$Given\ Data, D = \{\underline{x}^n, z^n\}_{n=1}^N \ and\ \ prior\ \ p(\underline{w}) = N(\underline{w}; \underline{w}_0, \Sigma_0)$$

$$-\ln p(\underline{w} \mid D) \leq -\ln p(\underline{w}) - \sum_{n=1}^{N}\left\{\ln(g(y_{0n})) + z^n \underline{w}^T \underline{x}^n - (\underline{w}^T \underline{x}^n + y_{0n})/2 - \lambda(y_{0n})[(\underline{w}^T \underline{x}^n)^2 - y_{0n}^2]\right\}$$

Copyright © 2001-2021 K.R. Pattipati

Variational EM for Logistic Regression

- Variational EM for minimizing the upper bound on NLL

$$Variational\ E-step:$$

$$q(\underline{w}) = p(\underline{w} \mid \{y_{0n}\}^{old}) = N(\underline{w}; \underline{w}_N, \Sigma_N)$$

$$(\Sigma_N)^{-1} = (\Sigma_0)^{-1} + 2\sum_{n=1}^{N} \lambda(y_{0n}^{old})\underline{x}^n \underline{x}^{n^T} = (\Sigma_0)^{-1} + \sum_{n=1}^{N} \frac{1}{y_{0n}}[g(y_{0n}) - \frac{1}{2}]\underline{x}^n \underline{x}^{n^T}$$

$$\underline{w}_N = \Sigma_N \left((\Sigma_0)^{-1} \underline{w}_0 + \sum_{n=1}^{N} (z^n - 1/2)\underline{x}^n \right)$$

$$Variational\ M-step: decouples\ \ for\ each\ y_{0n}$$

$$Q_i(y_{on}, y_{on}^{old}) = E\{\ln(g(y_{0n})) - y_{0n}/2 - \lambda(y_{0n})[(\underline{w}^T \underline{x}^n)^2 - y_{0n}^2]\}$$

$$\frac{dQ_i(y_{on}, y_{on}^{old})}{dy_{on}} = 0 \Rightarrow \frac{1}{g(y_{0n})} g(y_{0n})[1 - g(y_{0n})] - \frac{1}{2} + 2y_{0n}\lambda(y_{0n}) - \frac{d\lambda(y_{0n})}{dy_{on}}[E(\underline{w}^T \underline{x}^n)^2 - y_{0n}^2]$$

$$= [\frac{1}{2} - g(y_{0n})] + [g(y_{0n}) - \frac{1}{2}] - \frac{d\lambda(y_{0n})}{dy_{on}}[E(\underline{w}^T \underline{x}^n)^2 - y_{0n}^2] = 0$$

$$\Rightarrow E(\underline{w}^T \underline{x}^n)^2 - y_{0n}^2 = 0 \Rightarrow \left(y_{0n}^{new}\right)^2 = \left(\underline{x}^n\right)^T (\Sigma_N + \underline{w}_N \underline{w}_N^T)\underline{x}^n$$

This is still too much work! Are there simpler algorithms? Perceptrons and MLPs

Copyright © 2001-2021 K.R. Pattipati

Information Theoretic Co-clustering

- Most clustering algorithms seek to cluster one dimension of the matrix (e.g., documents or columns) based on similarities along the second dimension (e.g., word distribution of documents or rows).

- For sparse, noisy, and high-dimensional data, *simultaneous clustering* ("co-clustering", "bi-clustering") of both rows and columns is beneficial.
  - Example: given a term-document matrix, co-clustering in two dimensions simultaneously clusters terms and documents
  - Other Examples: Marketing, Dimensionality Reduction, Currency Exchange,……
  - More robust to sparsity than traditional single dimensional (e.g., terms or documents) clustering.
  - Co-clustering can be used as a pre-processor for supervised classification or as a classifier in its own right

Copyright © 2001-2021 K.R. Pattipati

Key Idea of Co-clustering

- Co-clustering Problem: Find maps

$$R(X):\{x_1,x_2,..,x_m\}\rightarrow\{\hat{x}_1,\hat{x}_2,...,\hat{x}_k\} \qquad C(Y):\{y_1,y_2,..,y_n\}\rightarrow\{\hat{y}_1,\hat{y}_2,...,\hat{y}_l\}$$

to minimize $\min_{\hat{X},Y}[I(X;Y)-I(\hat{X};\hat{Y})]\Rightarrow\max_{\hat{X},\hat{Y}} I(\hat{X};\hat{Y})$

- $\hat{X}=R(X)$ and $\hat{Y}=C(Y)$ $\Longrightarrow$ $H(\hat{X}\mid X)=H(\hat{Y}\mid Y)=0.$

$$\begin{aligned} I(X;Y)-I(\hat{X};\hat{Y}) &= [H(X)-H(\hat{X})]+[H(Y)-H(\hat{Y})]+[H(\hat{X},\hat{Y})-H(X,Y)] \\ &= H(X\mid\hat{X})+H(Y\mid\hat{Y})+H(\hat{X},\hat{Y})-H(X,Y) \\ &= E_{p(x,y)}[\log_2\frac{p(x,y)}{p(x\mid\hat{x})\,p(\hat{x},\hat{y})\,p(y\mid\hat{y})}]=D(p(x,y)\,\|\,q(x,y)) \end{aligned}$$

$$q(x,y)=p(x\mid\hat{x})\,p(\hat{x},\hat{y})\,p(y\mid\hat{y}) \text{ where } x\in\hat{x}\;,\; y\in\hat{y}.$$

$\boxed{\text{Decomposition of pmf } p(x,y) \text{ into a product of three matrices}}$

Copyright © 2001-2021 K.R. Pattipati

Illustration of Co-clustering

$$\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

$$p(x,y)$$

$$\begin{bmatrix} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{bmatrix} \begin{bmatrix} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{bmatrix} \begin{bmatrix} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{bmatrix} = \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 028 & .028 & .036 & .036 \\ .036 & .036 & .028 & .028 & .036 & .036 \end{bmatrix}$$

$$p(\hat{x},\hat{y}) = \sum_{x \in \hat{x}} \sum_{y \in \hat{x}} p(x,y)$$

$$p(y \mid \hat{y}) = \frac{p(y)}{p(\hat{y})}$$

$$q(x,y)$$

$$p(x \mid \hat{x}) = \frac{p(x,\hat{x})}{p(\hat{x})} = \frac{p(x)}{p(\hat{x})}$$

$$Note: q(x,y) = p(x \mid \hat{x})\, p(\hat{x},\hat{y})\, p(y \mid \hat{y}) = p(x) \underbrace{p(\hat{y} \mid \hat{x})\, p(y \mid \hat{y})}_{q(y \mid \hat{x})} = q(x) q(y \mid \hat{x}) = p(y) \underbrace{p(x \mid \hat{x})\, p(\hat{x} \mid \hat{y})}_{q(x \mid \hat{y})} = q(y) q(x \mid \hat{y})$$

- #parameters that determine q are: $(m\text{-}k)+(kl\text{-}1)+(n\text{-}l)$

Copyright © 2001-2021 K.R. Pattipati

Co-clustering Algorithm

- **Step 1**: Set iteration *i=1*. Start with initial cluster maps (R_i, C_i). Compute the pmfs $q^{(i,i)}(\hat{x},\hat{y}), q^{(i,i)}(x\,|\,\hat{x}), q^{(i,i)}(y\,|\,\hat{y}), q^{(i,i)}(y\,|\,\hat{x})$

$$q^{(i,i)}(y\,|\,\hat{x}) = \sum_{\hat{y}} q^{(i,i)}(y\,|\,\hat{y})q^{(i,i)}(\hat{y}\,|\,\hat{x}) = \sum_{\hat{y}} q^{(i,i)}(y\,|\,\hat{y})\frac{q^{(i,i)}(\hat{x},\hat{y})}{p(\hat{x})}$$

- **Step 2:** For every row *x*, assign it to the cluster that minimizes the K-L divergence $D(p(y\,|\,x)\,\|\,q^{(i,i)}(y\,|\,\hat{x}))$. The result is (R_{i+1}, C_i)
- **Step 3:** Compute the pmfs $q^{(i+1,i)}(\hat{x},\hat{y}), q^{(i+1,i)}(x\,|\,\hat{x}), q^{(i+1,i)}(y\,|\,\hat{y}), q^{(i+1,i)}(x\,|\,\hat{y})$

$$q^{(i+1,i)}(x\,|\,\hat{y}) = \sum_{\hat{x}} q^{(i+1,i)}(x\,|\,\hat{x})q^{(i+1,i)}(\hat{x}\,|\,\hat{y}) = \sum_{\hat{x}} q^{(i+1,i)}(x\,|\,\hat{x})\frac{q^{(i+1,i)}(\hat{x},\hat{y})}{p(\hat{y})}$$

- **Step 4:** For every column *y*, assign it to the cluster that minimizes the K-L divergence $D(p(x\,|\,y)\,\|\,q^{(i+,i)}(x\,|\,\hat{y}))$. The result is (R_{i+1}, C_{i+1})
- **Step 5:** Compute the pmfs $q^{(i+1,i+1)}(\hat{x},\hat{y}), q^{(i+1,i+1)}(x\,|\,\hat{x}), q^{(i+1,i+1)}(y\,|\,\hat{y}), q^{(i+1,i+1)}(y\,|\,\hat{x})$

Set *i=i+1*. Iterate Steps 2-5 until the K-L divergence converges.

Copyright © 2001-2021 K.R. Pattipati

1-D versus 2-D Clustering

Confusion Matrix

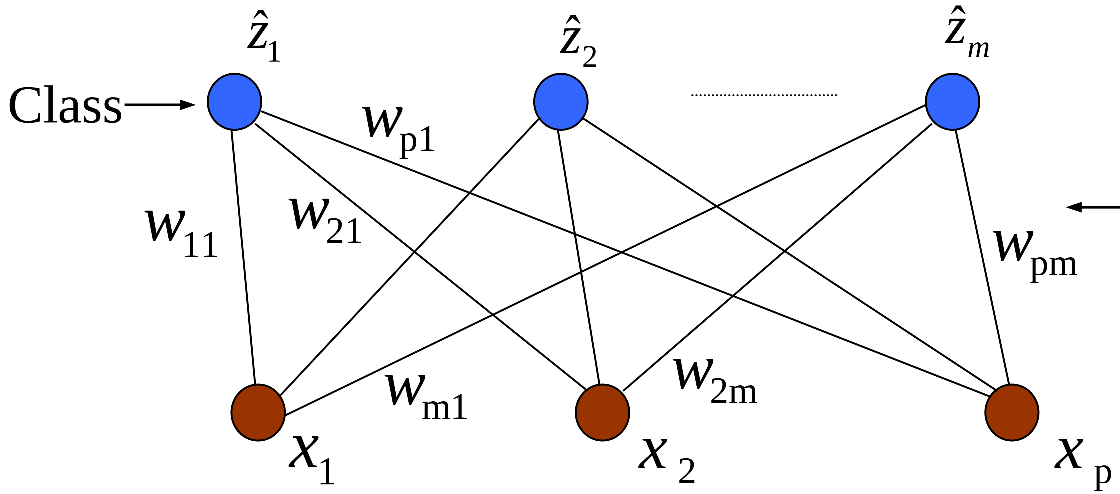
Co-Clustering (0.9835)			1-D Clustering (0.821)		
992	4	8	847	142	44
40	1452	7	41	954	405
1	4	1387	275	86	1099

I. Dhillon, 2003: CLASSIC 3 dataset
ftp://ftp.cs.cornell.edu/pub/smart

Copyright © 2001-2021 K.R. Pattipati

Learning Vector Quantization

- Learning Vector Quantization (Supervised Clustering)



Note the change: $\underline{w}_i$ for output i is a column vector

- Each output unit represents a class.
- *Several outputs may represent the same class*.
- The weight vector for an output unit is called a *Code book vector*.
- Initial *Code book vector* from *K*-means or any other clustering algorithm.

Copyright © 2001-2021 K.R. Pattipati

LVQ Algorithm -1

- <u>LVQ algorithm</u> can download LVQ-PAK from Helsinki Univ. of Technology

Step1: Initialize codebook vectors

Initialize learning rate, $\eta(0) \approx 0.03 \left(< 0.1\right)$

$$t = 0$$

Step2: While stopping condition is false, do steps 3-7

Step3: For each training input vector $\underline{x}^n$, do steps 4-5

Step4: Find J so that $\left\| \underline{x}^n - \underline{w}_J \right\|$ is a minimum

Copyright © 2001-2021 K.R. Pattipati

LVQ Algorithm -2

Step5: Update $\underline{w}_J$ as follows:

If $z^n = C_J$, then $\left(z^n \text{ is the correct class of } \underline{x}^n \right)$

$$\underline{w}_J^{(new)} = \underline{w}_J^{(old)} + \eta(t)\left[\underline{x}^n - \underline{w}_J^{(old)}\right]$$

If $z^n \neq C_J$, then

$$\underline{w}_J^{(new)} = \underline{w}_J^{(old)} - \eta(t)\left[\underline{x}^n - \underline{w}_J^{(old)}\right]$$

Step6: $\eta\left(t+1\right) = \varepsilon \cdot \eta\left(t\right)$; $t = t+1$

$$\varepsilon^{30\cdot\#\ of\ Codebooks} \approx e^{-38}$$

Step7: Test stopping condition:

$$\Rightarrow \varepsilon = e^{-38/(30\cdot\#\ of\ Codebooks\ )}$$

$t = \left(30 - 200\right) \times$ number of codebook vectors.

or, $\eta < 10^{-38}$

- Works well in practice.

Copyright © 2001-2021 K.R. Pattipati

Summary

- Mixture Models
- Expectation Maximization (EM)
- K-Means Algorithm
- Variational Bayes EM
- Variational Logistic Regression
- Learning Vector Quantization
- Information-Theoretic Co-clustering

Copyright © 2001-2021 K.R. Pattipati