PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 9: MIXTURE MODELS AND EM

Outlines

- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

K-means Clustering (1/3)

- Problem of identifying groups, or clusters, of data points in a multidimensional space
  - Partitioning the data set into some number K of clusters
  - Cluster: a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster
  - Goal: an assignment of data points to clusters such that the sum of the squares of the distances to each data point to its closest vector (the center of the cluster) is a minim

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

K-means Clustering (2/3)

- Two-stage optimization
  - In the 1st stage: minimizing $\mathcal{J}$ with respect to the r_{nk}, keeping the μ_k fixed

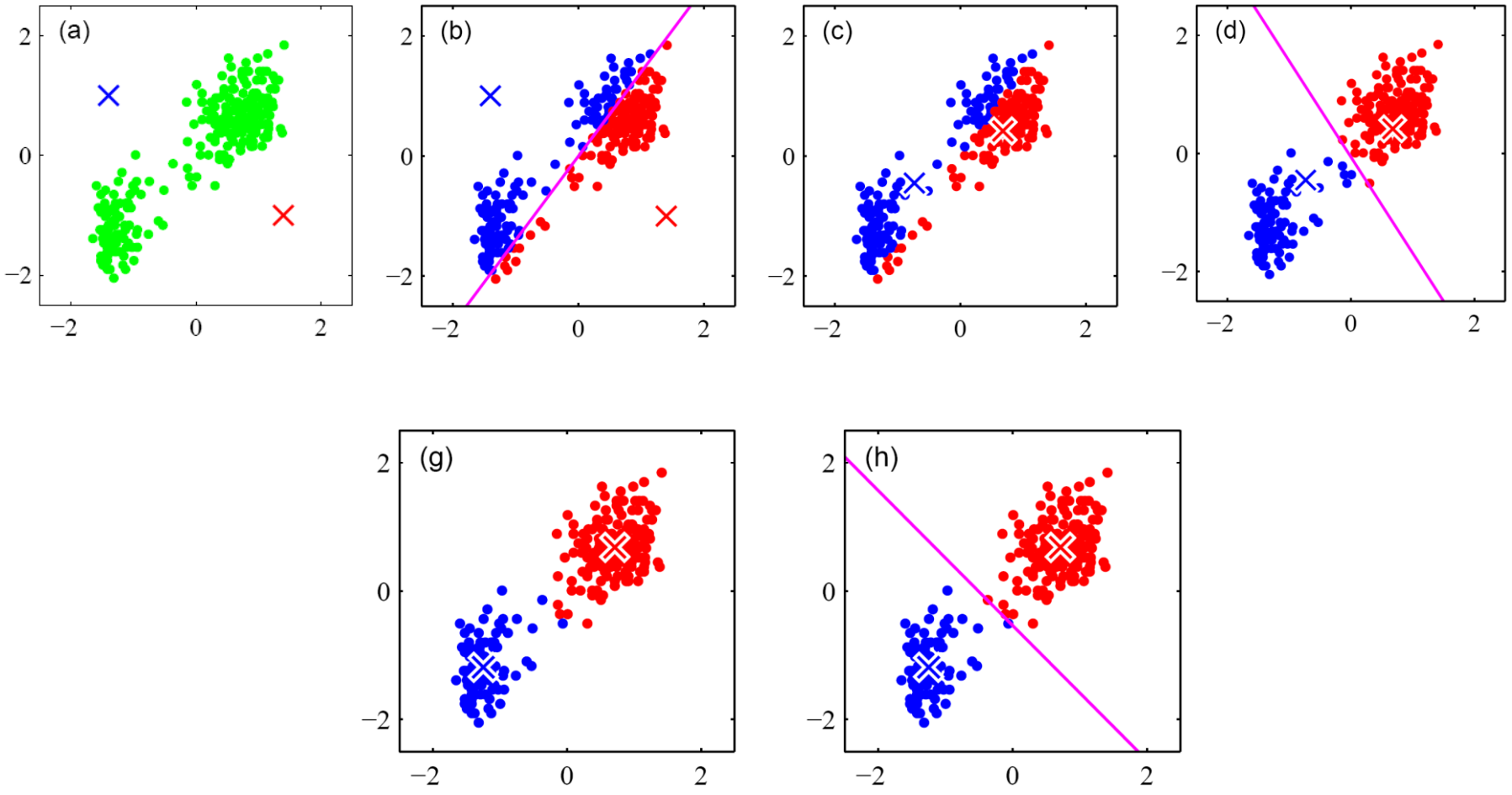
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j || \boldsymbol{x}_n - \mu_j ||^2 \\ 0 & otherwise \end{cases}$$

  - In the 2nd stage: minimizing $\mathcal{J}$ with respect to the μ_k, keeping r_{nk} fixed

$$\mu_k = \frac{\sum_n r_{nk} \boldsymbol{x}_n}{\sum_n r_{nk}}$$

The mean of all of the data points assigned to cluster k

K-means Clustering (3/3)



Outlines

- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

Gaussian Mixture Model (1/4)

- Gaussian mixture distribution can be written as a linear superposition of Gaussian

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\mu_k, \Sigma_k)$$

- **A** binary random variable **z** having a 1-of-K representation

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

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} N(\boldsymbol{x}|\mu_k, \Sigma_k)^{z_k}$$

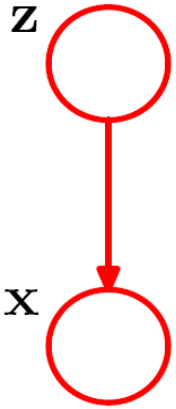
$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$$

$$p(z_k = 1) = \pi_k \qquad \sum_{k=1}^{K} \pi_k = 1$$

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

Gaussian Mixture Model (2/4)

- An equivalent formulation of the Gaussian mixture involving an explicit latent variable
  - Graphical representation of a mixture model
  - The marginal distribution of **x** is a Gaussian mixture of the form (*) (→ for every observed data point $\mathbf{x}_n$, there is a corresponding latent variable $\mathbf{z}_n$)



$$p(\boldsymbol{x}) = \Sigma_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z})$$

Mixtures of Gaussians (3/4)

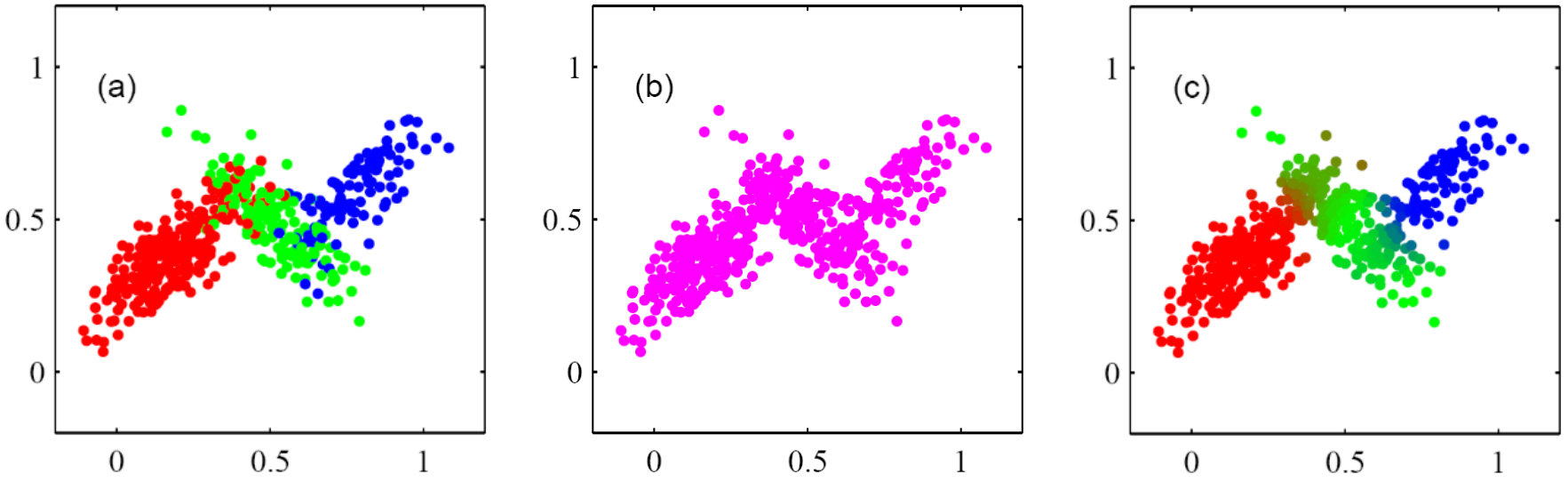
$$\gamma(z_k) \equiv p(z_k = 1 | \boldsymbol{x}) = \frac{p(z_k = 1) p(\boldsymbol{x} | z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1) p(\boldsymbol{x} | z_j = 1)}$$

$$= \frac{\pi_k N(\boldsymbol{x} | \mu_k, \sum_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x} | \mu_j, \sum_j)}$$

- $\gamma(z_k)$ can also be viewed as the responsibility that component *k* takes for explaining the observation **x**

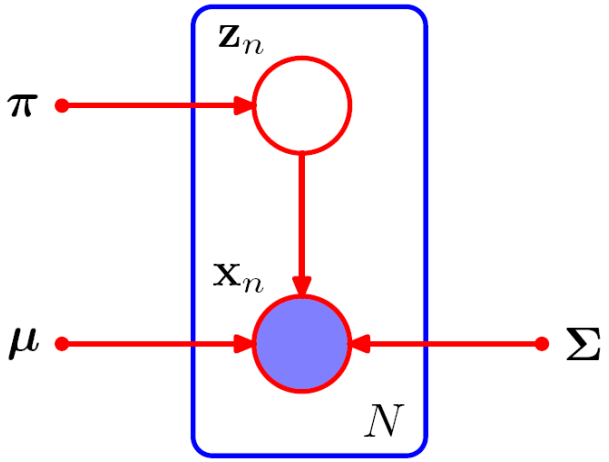
Mixtures of Gaussians (4/4)

- Generating random samples distributed according to the Gaussian mixture model
  - Generating a value for **z**, which denoted as $\hat{z}$ from the marginal distribution p(**z**) and then generate a value for **x** from the conditional distribution $p(\boldsymbol{x}|\hat{z})$



Maximum likelihood (1/3)

Graphical representation of a Gaussian mixture model for a set of N i.i.d. data points $\{x_n\}$, with corresponding latent points $\{z_n\}$



The log of the likelihood function

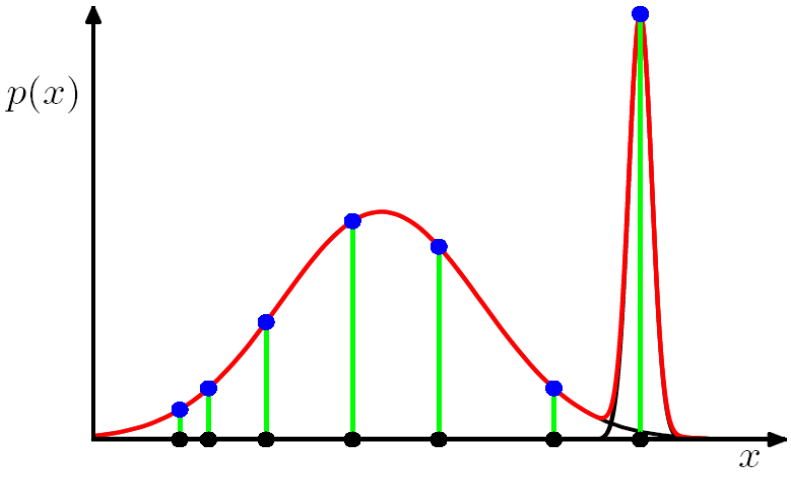
$$\ln p(\boldsymbol{X}|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\mu_k,\Sigma_k) \right\}$$

Maximum likelihood (2/3)

- For simplicity, consider a Gaussian mixture whose components have covariance matrices given by $\Sigma_k = \sigma_k^2 I$
  - Suppose that one of the components of the mixture model has its mean μ_j exactly equal to one of the data points so that $\mu_j = \mathbf{x}_n$
  - This data point will contribute a term in the likelihood function of the form

$$N(\boldsymbol{x}_n | \boldsymbol{x}_n, \sigma_j^2 I) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}$$

  - Once there are at least two components in the mixture, one of the components can have a finite variance and therefore assign finite probability to all of the data points while the other component can shrink onto one specific data point and thereby contribute an ever increasing additive value to the log likelihood → over-fitting problem



Maximum likelihood (3/3)

- Over-fitting problem
  - Example of the over-fitting in a maximum likelihood approach
  - This problem does not occur in the case of Bayesian approach
  - In applying maximum likelihood to a Gaussian mixture models, there should be steps to avoid finding such pathological solutions and instead seek local minima of the likelihood function that are well behaved
- *Identifiability* problem
  - A K-component mixture will have a total of K! equivalent solutions corresponding to the K! ways of assigning K sets of parameters to K components
- Difficulty of maximizing the log likelihood function → the presence of the summation over k that appears inside the logarithm gives no closed form solution as in the single case

Outlines

- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

EM for Gaussian Mixtures (1/4)

I. Assign some initial values for the means, covariances, and mixing coefficients

II. *Expectation* or E step

- Using the current value for the parameters to evaluate the posterior probabilities or *responsibilities*

III. *Maximization* or M step

- Using the result of II to re-estimate the means, covariances, and mixing coefficients

❖ It is common to run the K-means algorithm in order to find a suitable initial values

- The covariance matrices → the sample covariances of the clusters found by the K-means algorithm
- Mixing coefficients → the fractions of data points assigned to the respective clusters

EM for Gaussian mixtures (2/4)

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters

1. Initialize the means μ_k, covariance Σ_k and mixing coefficients π_k
2. E step

$$\gamma(z_{nk}) = \frac{\pi_k N(\boldsymbol{x_n}|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x_n}|\mu_j,\Sigma_j)}$$

3. M step

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

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

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

4. Evaluate the log likelihood

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

EM for Gaussian mixtures (3/4)

$$\ln p(\boldsymbol{X}|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x_n}|\mu_k,\Sigma_k) \right\} \quad \text{...(*2)}$$

Setting the derivatives of (*2) with respect to the means of the Gaussian components to zero →

$$0 = -\sum_{n=1}^{N} \frac{\pi_k n(\boldsymbol{x_n}|\mu_k,\Sigma_k)}{\Sigma_j \pi_j N(\boldsymbol{x_n}|\mu_j,\Sigma_j)} \Sigma_k(\boldsymbol{x_n} - \mu_k) \quad \frac{\pi_k n(\boldsymbol{x_n}|\mu_k,\Sigma_k)}{\Sigma_j \pi_j N(\boldsymbol{x_n}|\mu_j,\Sigma_j)} \leftarrow \text{Responsibility} \gamma(z_{nk})$$

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

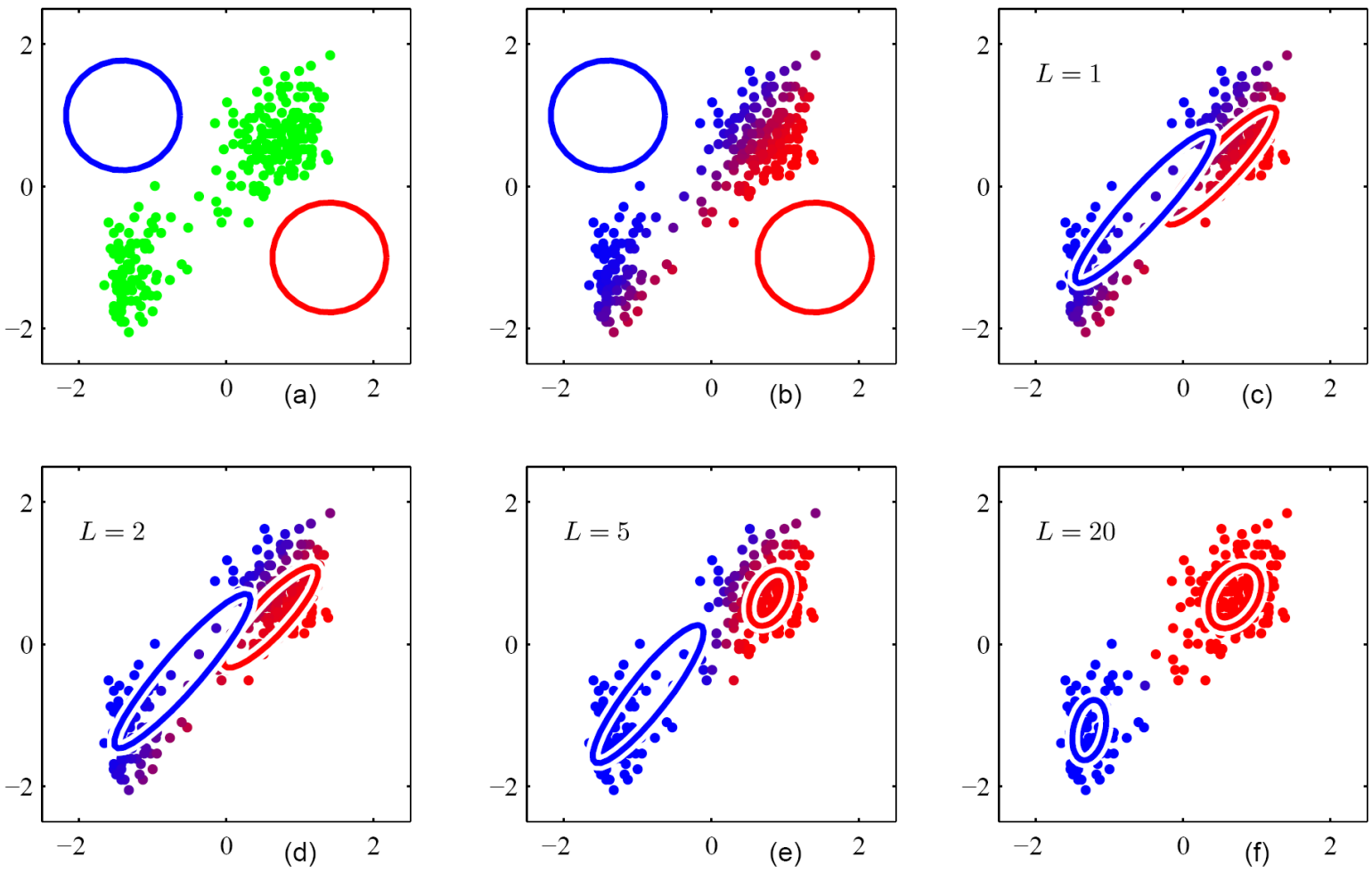
A weighted mean of all of the points in the data set

Setting the derivatives of (*2) with respect to the covariance of the Gaussian components to zero →

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

- Each data point weighted by the corresponding posterior probability
- The denominator given by the effective # of points associated with the corresponding component

EM for Gaussian mixtures (4/4)



Outlines

- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

An Alternative View of EM

- In maximizing the log likelihood function $\ln p(\boldsymbol{X}|\theta) = \ln\{\Sigma_{\boldsymbol{Z}} P(\boldsymbol{X},\boldsymbol{Z}|\theta)\}$

  the summation prevents the logarithm from acting directly on the joint distribution
- Instead, the log likelihood function for the complete data set {X, Z} is straightforward.
- In practice since we are not given the complete data set, we consider instead its expected value Q under the posterior distribution p(**Z**|**X**, Θ) of the latent variable

- General EM

1. Choose an initial setting for the parameters Θ^{old}
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X},\Theta^{old})$
3. **M step** Evaluate Θ^{new} given by
   $\Theta^{new} = \text{argmax}_{\Theta} Q(\Theta, \Theta^{old})$
   $Q(\Theta, \Theta^{old}) = \Sigma_{Z} p(\mathbf{Z}|\mathbf{X}, \Theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\Theta)$
4. It the covariance criterion is not satisfied, then let $\Theta^{old} \leftarrow \Theta^{new}$

Gaussian Mixtures Revisited (1/2)

- Maximizing the likelihood for the complete data {X, Z}

$$p(X, Z \mid \mu, \Sigma, \pi) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k N(x_n \mid \mu_k, \Sigma_k)]^{z_{nk}}$$

$$\ln p(X, Z \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln N(x_n \mid \mu_k, \Sigma_k)\}$$

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

- The logarithm acts directly on the Gaussian distribution → much simpler solution to the maximum likelihood problem
  - the maximization with respect to a mean or a covariance is exactly as for a single Gaussian (closed form)

Gaussian Mixtures Revisited (2/2)

- Unknown latent variables → considering expectation of the complete-data log likelihood with respect to the posterior distribution of the latent variables

- Posterior distribution

$$p(Z \mid X, \mu, \Sigma, \pi) \propto p(X \mid Z, \mu, \Sigma, \pi) p(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k N(x_n \mid \mu_k, \Sigma_k)]^{z_{nk}}$$

$(ref.\ (9.10),\ (9.11))$

- The expected value of the indicator variable under this posterior distribution

$$E(z_{nk}) = p(z_{nk} = 1 \mid x_n) = \frac{p(z_{nk} = 1)\, p(x_n \mid z_{nk} = 1)}{p(x_n)}$$

$$= \frac{\pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)} = \gamma(z_{nk})$$

- The expected value of the complete-data log likelihood function

$$E[\ln p(X, Z \mid \mu, \Sigma, \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln N(x_n \mid \mu_k, \Sigma_k)\} \quad \ldots (*3)$$

Relation to K-means

- K-means performs a hard assignment of data points to the clusters (each data point is associated uniquely with one cluster
- EM makes a soft assignment based on the posterior probabilities
- K-means can be derived as a particular limit of EM for Gaussian mixtures:

As epsilon gets smaller, the terms for which $\|x_n - \mu_j\|^2$ is farthest will go to zero most quickly. Hence the responsibility go all zero except for the term *k* for which the responsibility will go to unity

$$p(x_n \mid \mu_k, \varepsilon I_k) = \frac{1}{\sqrt{2\pi\varepsilon}} \exp\left\{-\frac{1}{2\varepsilon}\|x_n - \mu_k\|^2\right\}, \qquad \gamma(z_{nk}) = \frac{\pi_k \exp\{-\|x_n - \mu_k\|^2 / 2\varepsilon\}}{\sum_j \pi_j \exp\{-\|x_n - \mu_j\|^2 / 2\varepsilon\}}$$

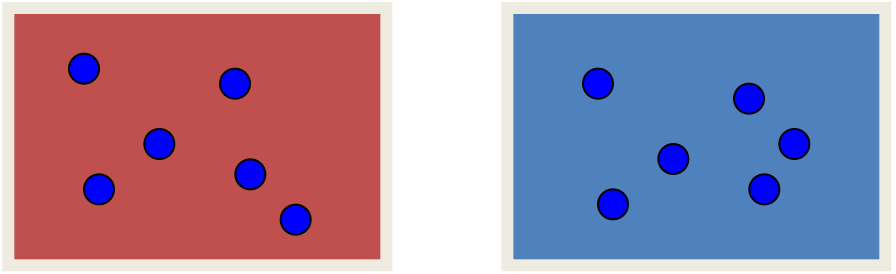
$$E[\ln p(X, Z \mid \mu, \Sigma, \pi)] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|x_n - \mu_k\|^2 + const \qquad J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}||\ \boldsymbol{x}_n - \mu_k||^2$$

- Thus maximizing the expected complete data log-likelihood is equivalent to minimizing the distortion measure J for the K-means
- (In Elliptical K-means, the covariance is estimated also.)

Outlines

- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

Mixtures of Bernoulli Distributions



Mixtures of Bernoulli Distributions (1/3)

* $p(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{(1-x_i)}$ (single component)

  $(\mathbf{x} = (x_1, \cdots, x_D), E(\mathbf{x}) = \boldsymbol{\mu} = (\mu_1, \cdots, \mu_D), \quad Cov(\mathbf{x}) = \boldsymbol{\Sigma} = diag\{\mu_i(1-\mu_i)\})$

* $p(\mathbf{x} \mid \boldsymbol{\mu}, \pi) = \sum_{k=1}^{K} \pi_k p(\mathbf{x} \mid \boldsymbol{\mu}_k), \quad p(\mathbf{x} \mid \boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{(1-x_i)}$ (mixture)

  $(\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}, \quad \boldsymbol{\mu}_k = (\mu_{k1}, \cdots, \mu_{kD}))$

  $E(\mathbf{x}) = \sum_{k=1}^{K} \pi_k E(\mathbf{x} \mid \boldsymbol{\mu}_k) = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k,$

  $Cov(\mathbf{x}) = E(\mathbf{x}\mathbf{x}^{\mathbf{T}}) - E(\mathbf{x})E(\mathbf{x})^T = \sum_{k=1}^{K} \pi_k (E(\mathbf{x}\mathbf{x}^{\mathbf{T}} \mid \boldsymbol{\mu}_k)) - E(\mathbf{x})E(\mathbf{x})^T = \sum_{k=1}^{K} \pi_k \{\Sigma_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathbf{T}}\} - E(\mathbf{x})E(\mathbf{x})^T$

  (Let $E(\mathbf{x}\mathbf{x}^{\mathbf{T}} \mid \boldsymbol{\mu}_k)) \equiv (c_{ij(k)})$, then $c_{ii(k)} = E(x_i^2 \mid \boldsymbol{\mu}_k) = \mu_k, \quad c_{ij(k)} = E(x_i x_j \mid \boldsymbol{\mu}_k) = \mu_k^2 \quad (i \neq j)$

  and note that individual variabes x_i are *independent, given* $\boldsymbol{\mu}$)

Mixtures of Bernoulli Distributions (2/3)

* [*E.g.*] $\mathbf{x} = (1, \quad 0, \quad 0, \quad 1, \quad 1)$

(single component) :

$$p(H) = \mu \; : \; p(\mathbf{x}) = \mu^3(1-\mu)^2, \; E(\mathbf{x}) = (\mu, \cdots, \mu), \; Cov(\mathbf{x}) = \Sigma = \mu(1-\mu)I$$

(mixture):

$$p(H \mid c_1) = \mu_1, \; p(H \mid c_2) = \mu_2 \; : \; p(\mathbf{x}) = \pi_1\mu_1^3(1-\mu_1)^2 + \pi_2\mu_2^3(1-\mu_2)^2,$$

$$E(\mathbf{x}) = (\pi_1\mu_1 + \pi_2\mu_2, \cdots, \pi_1\mu_1 + \pi_2\mu_2)$$

$$Cov(\mathbf{x}) = \pi_1\{\mu_1(1-\mu_1)I + \mu_1^2\mathbf{1}\} + \pi_2\{\mu_2(1-\mu_2)I + \mu_2^2\mathbf{1}\} - (\pi_1\mu_1 + \pi_2\mu_2)^2\mathbf{1}$$

$$= \{\pi_1\mu_1(1-\mu_1) + \pi_2\mu_2(1-\mu_2)\}I + \{\pi_1\mu_1^2 + \pi_1\mu_2^2 - (\pi_1\mu_1 + \pi_2\mu_2)^2\}\mathbf{1}$$

* Because the covariance matrix Cov($\mathbf{x}$) is no longer diagonal, the mixture distribution can capture correlations between the variables, unlike a single Bernulli distribution.

Mixtures of Bernoulli Distributions (3/3)

$$\ln p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k p(\mathbf{x_n} \mid \boldsymbol{\mu}_k) \right\}$$

$$p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mathbf{x} \mid \boldsymbol{\mu_k})^{z_k} \qquad (\mathbf{z} = (z_1, \cdots, z_K)^{\mathrm{T}} \text{ is a binary indicator variables})$$

$$p(\mathbf{z} \mid \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

(complete－data log likelihood function)：

$$\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln (1 - \mu_{ki})] \right\}$$

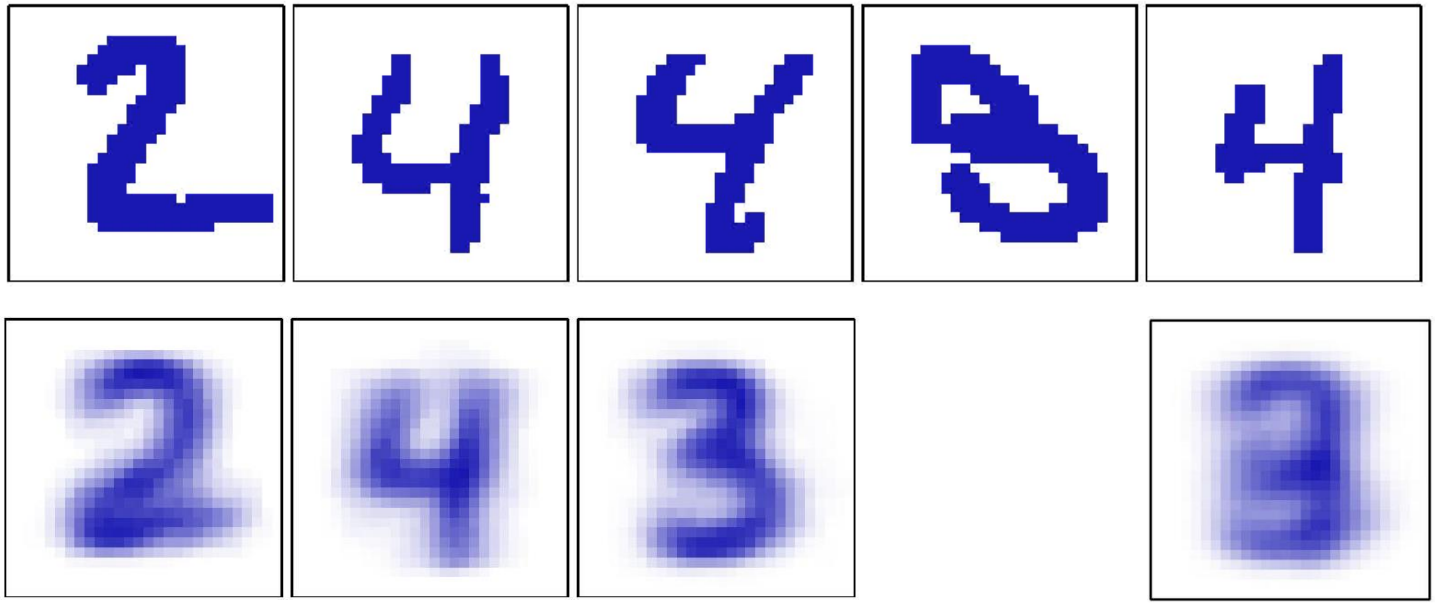
$$\mathrm{E}_Z[\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\mu}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln (1 - \mu_{ki})] \right\}$$

(E－step) $\gamma(z_{nk}) = E[z_{nk}] = \dfrac{\pi_k p(\mathbf{x_n} \mid \boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x_n} \mid \boldsymbol{\mu}_j)}$, $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$, $\bar{\mathbf{x}}_\mathbf{k} = \dfrac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x_n}$

(M－step) $\boldsymbol{\mu}_k = \bar{\mathbf{x}}_\mathbf{k}$, $\pi_\mathrm{k} = \dfrac{N_k}{N}$

* In contrast to the mixture of Gaussians, there are no singularities in which the likelihood goes to infinity

Mixtures of Bernoulli Distributions



- N=600 digit images, 3 mixtures
- A mixture of k=3 Bernoulli distributions by 10 EM iterations
- Parameters for each of the three components/single multivariate Bernoulli
- The analysis of Bernoulli mixtures can be extended to the case of multinomial binary variables having M>2 states (Ex.9.19)

Outlines

- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

The EM Algorithm in General (1/3)

- Direct optimization of $p(X|\theta)$ is difficult while optimization of complete data likelihood $p(X, Z|\theta)$ is significantly easier.
- Decomposition of the likelihood $p(X|\theta)$

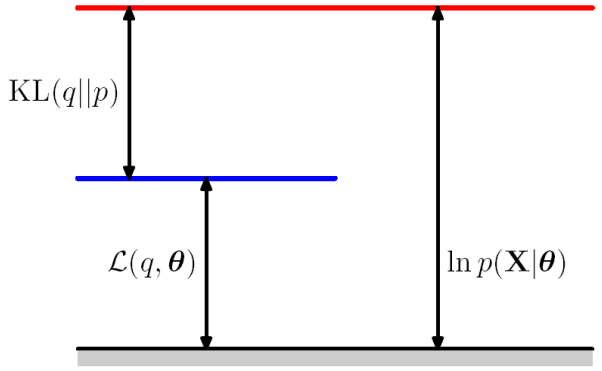
$$p(X \mid \theta) = \sum_Z p(X, Z \mid \theta)$$

$$\boxed{\ln p(X \mid \theta) = \mathcal{L}(q, \theta) + KL(q \,\|\, p)}$$

$$\left(\text{where } \mathcal{L}(q, \theta) = \sum_Z q(Z) \ln \left\{ \frac{p(X, Z \mid \theta)}{q(Z)} \right\},\ KL(q \,\|\, p) = -\sum_Z q(Z) \ln \left\{ \frac{p(Z \mid X, \theta)}{q(Z)} \right\}\right)$$

$$\left(\text{substitute } \ln p(X, Z \mid \theta) = \ln p(Z \mid X, \theta) + \ln p(X \mid \theta) \text{ into the expression for } \mathcal{L}(q, \theta)\right)$$

$\mathcal{L}(q, \theta) \le \ln p(X \mid \theta) \rightarrow \mathcal{L}(q, \theta)$ is a lower bound on $\ln p(X \mid \theta)$

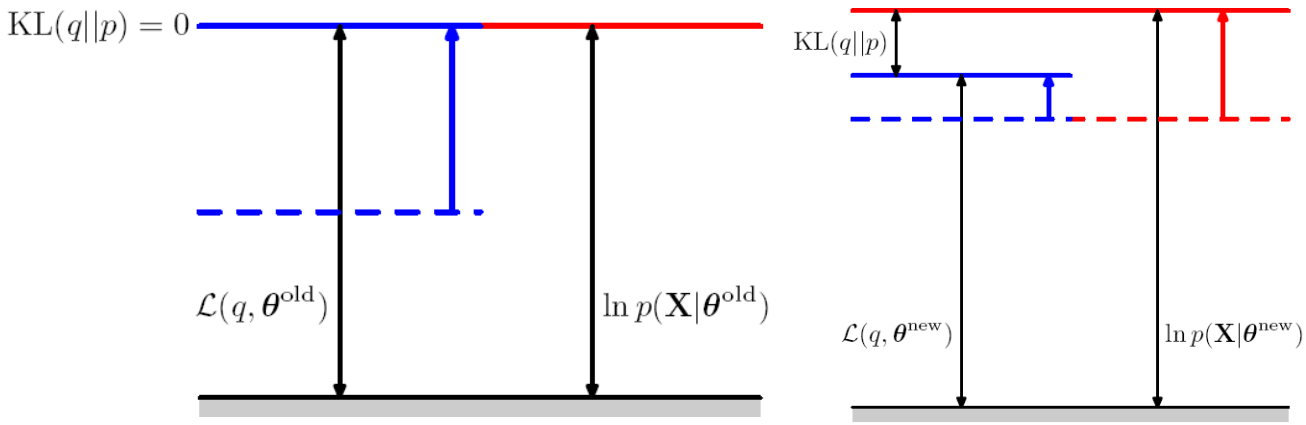


The EM Algorithm in General (2/3)

- (E step) The lower bound $\mathcal{L}(q, \theta_{old})$ is maximized while holding θ_{old} fixed. Since $\ln p(X|\theta)$ does not depend on $q(Z)$, $\mathcal{L}(q, \theta_{old})$ will be the largest when $KL(q||p)$ vanishes (i.e. when $q(Z)$ is equal to the posterior distribution $p(Z|X, \theta_{old})$)
- (M step) $q(Z)$ is fixed and the lower bound $\mathcal{L}(q, \theta_{old})$ is maximized wrt. θ to θ_{new}. When the lower bound is increased, θ is updated making $KL(q||p)$ greater than 0. Thus the increase in the log likelihood function is greater than the increase in the lower bound.
- In the M step, the quantity being maximized is the expectation of the complete-data log-likelihood

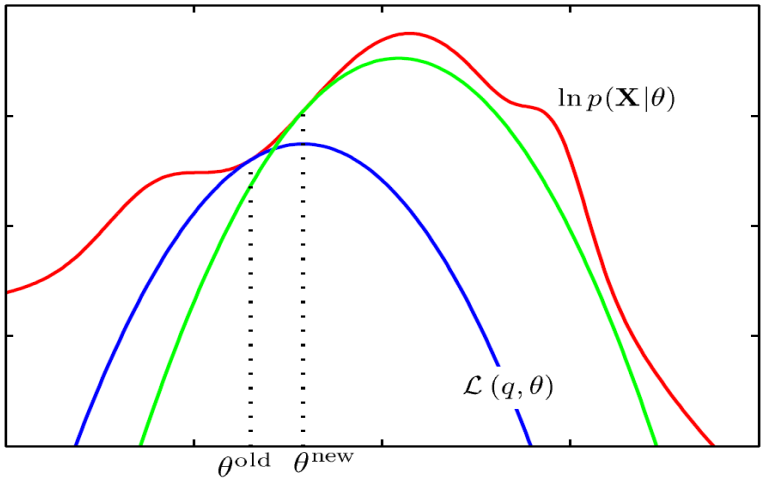
$$\mathcal{L}(q,\theta) = \sum_Z p(Z \mid X, \theta_{old}) \ln p(X, Z \mid \theta) - \sum_Z p(Z \mid X, \theta_{old}) \ln p(Z \mid X, \theta_{old})$$

$$= Q(\theta, \theta_{old}) + const \qquad (where \quad q(Z) \text{ is set to be } p(Z \mid X, \theta_{old}))$$



The EM Algorithm in General (3/3)

- Start with initial parameter value θ_{old}
- In the first E step, evaluation of posterior distribution over latent variables gives rise to a lower bound $\mathcal{L}(q, \theta_{old})$ whose value equals the log likelihood at θ_{old} (blue curve)
- Note that the bound makes a tangential contact with the log-likelihood at θ_{old}, so that both curves have the same gradient
- For mixture components from the exponential family, this bound is a convex function
- In the M step, the bound is maximized giving the value θ_{new} which gives a larger value of log-likelihood than θ_{old}.
- The subsequent E step constructs a bound tangential at θ_{new} (green curve)



The EM Algorithm in General (3a/3)

- EM can be also used to maximize the posterior distribution $p(\theta|X)$ over parameters.
- Optimize the RHS alternatively wrt q and θ
- Optimization wrt q is the same E step
- M step required only a small modification through the introduction of the prior term ln $p(\theta)$

$$p(\theta \mid X) = p(\theta, X) / p(X), \qquad \ln p(\theta \mid X) = \ln p(\theta, X) - \ln p(X)$$

$$\text{by decomposition (9.70)}, \quad \ln p(\theta \mid X) = \ln p(X \mid \theta) + \ln p(\theta) - \ln p(X)$$

$$= \mathcal{L}(q, \theta) + KL(q \| p) + \ln p(\theta) - \ln p(X) \geq \boxed{\mathcal{L}(q, \theta) + \ln p(\theta) + const.}$$

The EM Algorithm in General (3b/3)

For complex problems, either E step or M step or both are intractable:

- Intractable M: Generalized EM (GEM), expectation conditional maximization (ECM)
- Intractable E: Partial E step
- GEM: instead of maximizing $\mathcal{L}(q,\theta)$ wrt θ, it seeks to change the parameters to increase its value.
- ECM: makes several constrained optimization within each M step. For instance, parameters are partitioned into groups and the M step is broken down into multiple steps each of which involves optimizing one of the subset with the remainder held fixed.
- Partial (or incremental) EM: (Note) For any given θ, there is a unique maximum $\mathcal{L}(q^*,\theta)$ wrt q . Since $\mathcal{L}(q^*,\theta) = \ln p(X|\theta)$, there is a θ^* for the global maximum of $\mathcal{L}(q,\theta)$ and $\ln p(X|\theta^*)$ is a global maximum too. Any algorithm that converges to the global maximum of $\mathcal{L}(q,\theta)$ will find a value of θ that is also a global maximum of the log likelihood $\ln p(X|\theta)$
- Each E or M step in partial E step algorithm is increasing the value of $\mathcal{L}(q,\theta)$ and if the algorithm converges to a local (or global) maximum of $\mathcal{L}(q,\theta)$, this will correspond to a local (or global) maximum of the log likelihood function $\ln p(X|\theta)$.

The EM Algorithm in General (3c/3)

- (Incremental EM) For a Gaussian mixture, suppose $\mathbf{x}_m$ is updated with old and new values of responsibilities $\gamma^{old}(z_{mk})$, $\gamma^{new}(z_{mk})$ in the E-step .

  In the M step, the means are updated as,

$$(9.17) \qquad \mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n$$

$$(9.18) \qquad N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

$$\mu_k^{new} = \mu_k^{old} + \left(\frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}}\right)(x_m - \mu_k^{old})$$

$$N_k^{new} = N_k^{old} + \gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})$$

$$N_k^{new}\mu_k^{new} = N_k^{old}\mu_k^{old} + \gamma^{new}(z_{mk})x_m - \gamma^{old}(z_{mk})x_m$$

$$N_k^{new}\mu_k^{new} = [N_k^{new} - \gamma^{new}(z_{mk}) + \gamma^{old}(z_{mk})]\mu_k^{old} + [\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})]x_m$$

$$\mu_k^{new} = \left(1 - \frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}}\right)\mu_k^{old} + \left(\frac{\gamma^{new}(z_{mk}) - \gamma^{old}(z_{mk})}{N_k^{new}}\right)x_m$$

- Both E step and M step take fixed time independent of the total number of data points. Because the parameters are revised after each data point, rather than waiting until after the whole data set is processed, this incremental version can converge faster than the batch version.

HW5

K-means: 9.1

Mixture Model: 9.3 9.7 9.10 9.18 9.19

EM: 9.20 9.21 9.25 9.26

Due Dec. 6