PATTERN RECOGNITION AND MACHINE LEARNING

CHAPTER 9: MIXTURE MODELS AND EM

Learning Objectives

1、What are the differences between supervised and unsupervised learning schemes?

2、What is K-means clustering?

3、What are Gaussian Mixture Models?

4、What are Bernoulli Mixture Models?

5、What is the EM learning scheme?

6、How to understand EM from the perspective of likelihood?

7、How to generalize the EM scheme via decomposition?

Outlines

- Supervised vs Unsupervised Learning
- 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 $\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$
  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(**Z**|**X**,Θ^{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 $\Theta^{old} \leftarrow \Theta^{new}$

Expected Complete-Data Log Likelihood

Expectation of complete-data log likelihood:

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

$$\leq$$

Log Likelihood:

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

Jason Inequality: as ln (x) is a concave function,

E{ln f(x)} ≤ ln(E{f(x)})

An Alternative View of EM for MAP

- In maximizing the log posterior, $\ln p(\Theta \mid \mathbf{X}) \propto \ln p(\mathbf{X} \mid \Theta) + \ln p(\Theta)$, given the prior $p(\Theta)$ 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 $p(\mathbf{Z} \mid \mathbf{X}, \Theta)$ of the latent variable

- **General EM for MAP**

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

Gaussian Mixtures Revisited (I)

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

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

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

- 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 (II)

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

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

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

$$\begin{aligned}
\mathbb{E}[z_{nk}] &= \frac{\sum_{z_{nk}} z_{nk} \left[\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]^{z_{nk}}}{\sum_{z_{nj}} \left[\pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right]^{z_{nj}}} \\
&= \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} = \gamma(z_{nk})
\end{aligned}$$

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

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

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:

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}-\boldsymbol{\mu}_k\|^2\right\}$$

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n-\boldsymbol{\mu}_j\|^2/2\epsilon\right\}}$$

$$\boxed{\begin{array}{l} \epsilon \longrightarrow 0 \qquad \gamma(z_{nk}) \longrightarrow r_{nk} \qquad r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \\ \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] \rightarrow -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2 + \text{const.} \end{array}}$$

Outlines

- Supervised vs Unsupervised Learning
- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

Mixtures of Bernoulli Distributions (I)

Mixtures of Bernoulli Distributions (II)

$$p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k} \qquad p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

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

Mixtures of Bernoulli Distributions (III)

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

$$\Rightarrow \quad \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \Bigg\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \Bigg\}$$

EM for Bernoulli Mixture Models

E-Step:

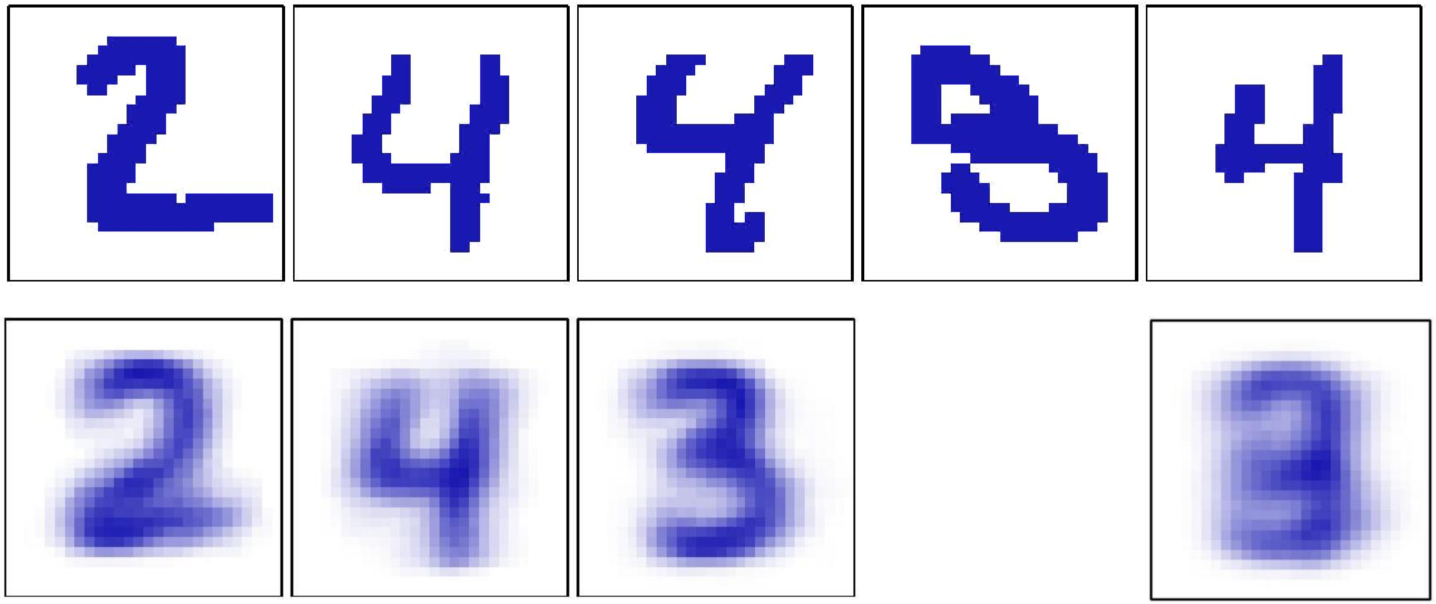
$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk} \left[\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)\right]^{z_{nk}}}{\sum_{z_{nj}} \left[\pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)\right]^{z_{nj}}} = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)}.$$

M-Step:

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

$$\overline{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \boldsymbol{\mu}_k = \overline{\mathbf{x}}_k$$

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

Outlines

- Supervised vs Unsupervised Learning
- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization

The EM Algorithm in General (I)

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

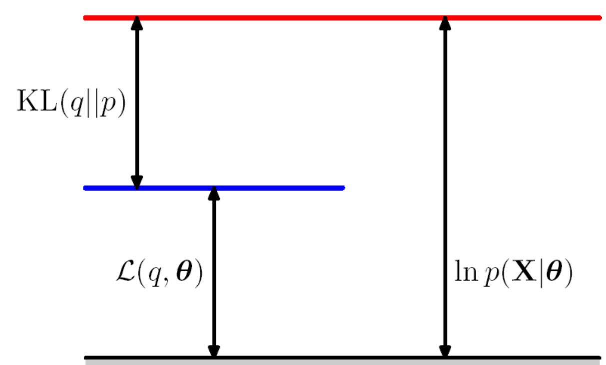
$$\boxed{\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) = \ln p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) + \ln p(\mathbf{X}|\boldsymbol{\theta})}$$

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

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln\left\{\frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})}\right\}$$

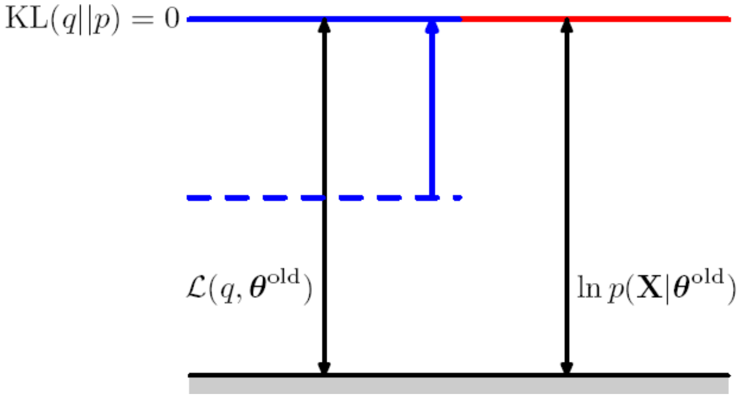
$$\mathrm{KL}(q\|p) \geqslant 0 \quad \Rightarrow \quad \mathcal{L}(q,\boldsymbol{\theta}) \leqslant \ln p(\mathbf{X}|\boldsymbol{\theta})$$

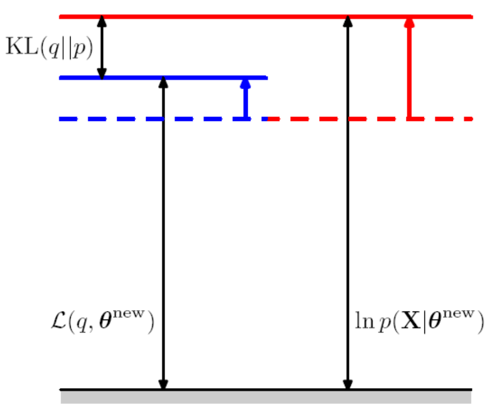


The EM Algorithm in General (II)

- (**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

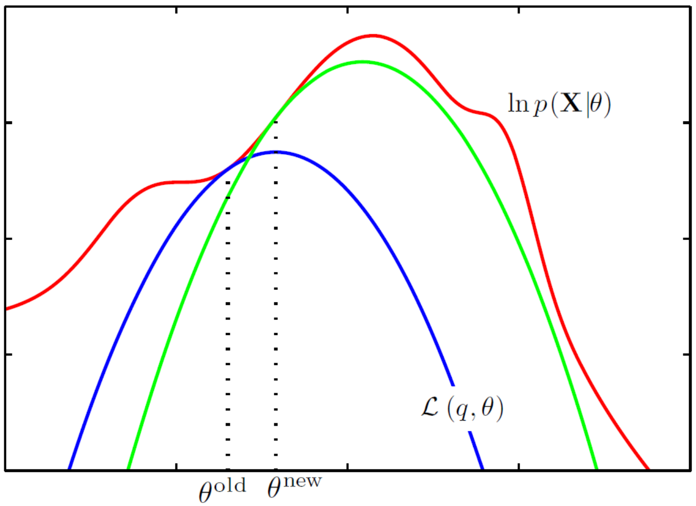
$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\mathrm{old}}) \\ &= \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{const}\end{aligned} \Longleftarrow \boxed{q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\mathrm{old}})}$$





The EM Algorithm in General (III)

- 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 for MAP

- 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 $\ln p(\theta)$

$$\boxed{\ln p(\boldsymbol{\theta}|\mathbf{X}) = \ln p(\boldsymbol{\theta}, \mathbf{X}) - \ln p(\mathbf{X})}$$

$$\begin{aligned} \ln p(\boldsymbol{\theta}|\mathbf{X}) &= \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p) + \ln p(\boldsymbol{\theta}) - \ln p(\mathbf{X}) \\ &\geqslant \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\mathbf{X}). \end{aligned}$$

EM Algorithm Variations

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 $\ln p(X|\theta)$.

Incremental EM Algorithm

- (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 <u>**E-step**</u>
- In the <u>**M step**</u>, the means are updated as,

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \left(\frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \right) \left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right)$$

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

- Both <u>**E step**</u> and <u>**M step**</u> 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.

EM for Linear Regression

- The complete-data log likelihood

$$\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta) = \ln p(\mathbf{t}|\mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha)$$

- E-Step: expectation over **w**

$$\mathbb{E}\left[\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta)\right] = \frac{M}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\mathbb{E}\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right] + \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{n=1}^{N}\mathbb{E}\left[(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)^2\right]$$

- M-Step

$$\alpha = \frac{M}{\mathbb{E}\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right]} = \frac{M}{\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N + \mathrm{Tr}(\mathbf{S}_N)}$$

$$\frac{1}{\beta} = \frac{1}{N}\sum_{i=1}^{N}(t_n - \mathbf{m}_N^T\,\phi_n)^2$$

EM for Sparse Kernel Machines

- The complete-data log likelihood

$$\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta) = \ln p(\mathbf{t}|\mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha)$$

- E-Step: expectation over **w**

$$\mathbb{E}_{\mathbf{w}}\left[\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha})\right]$$

- M-Step

$$\begin{aligned}
\alpha_i^{\text{new}} &= \frac{1}{m_i^2 + \Sigma_{ii}} \\
(\beta^{\text{new}})^{-1} &= \frac{\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\|^2 + \beta^{-1}\sum_i \gamma_i}{N}
\end{aligned}$$

Summary

- Supervised vs Unsupervised Learning
- K-means Clustering
- Gaussian Mixture Model
- Expectation and Maximization
- GMM Revisited
- Bernoulli Mixture Model
- EM Generalization