Machine learning

Parametric Models Part II Expectation-Maximization and Mixture Density Estimation

Mohammad-Reza A. Dehaqani

dehaqani@ut.ac.ir

Missing Features

- Suppose that we have a **Bayesian classifier** that uses the **feature vector** $\mathbf{x}$ but a subset $\mathbf{x}_g$ of $\mathbf{x}$ are **observed** and the values for the remaining features $\mathbf{x}_b$ are **missing**

$$\boxed{g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)}$$

- How can we make a decision?
  - Throw away the observations with missing values.
  - Or, substitute $\mathbf{x}_b$ by their average $\overline{x_b}$ in the training data, and use $x = (\mathbf{x}_g, \overline{x_b})$.
  - Or, **marginalize the posterior over the missing** features, and use the resulting posterior

$$P(w_i|\mathbf{x_g}) = \frac{\int P(w_i|\mathbf{x_g}, \mathbf{x_b})\, p(\mathbf{x_g}, \mathbf{x_b})\, d\mathbf{x_b}}{\int p(\mathbf{x_g}, \mathbf{x_b})\, d\mathbf{x_b}}.$$

Marginal likelihood

- A **likelihood function** in which some **parameter** variables have been marginalized

$$p(\mathbf{X}|\alpha) = \int_\theta p(\mathbf{X}|\theta)\, p(\theta|\alpha)\ \mathrm{d}\theta$$

θ has been **marginalized out** (**integrated** out)

$$\mathcal{L}(\psi; \mathbf{X}) = p(\mathbf{X}|\psi) = \int_\lambda p(\mathbf{X}|\lambda, \psi)\, p(\lambda|\psi)\ \mathrm{d}\lambda$$

Expectation-Maximization

- Expectation–maximization (EM) algorithm is an **iterative** method to find **maximum likelihood** or **maximum a posteriori (MAP) estimates of parameters** in statistical models, from training data.
- The model depends on **unobserved latent** variables.
- It allows **learning of parameters** when some training patterns have **missing features (partial observation)**.

Applications of the EM algorithm

1. **Learning** when the **data is incomplete** or has missing values.
2. **Optimizing** a likelihood (or posterior) function that is **analytically intractable** but can be **simplified** by assuming the existence of and values **for additional but missing** (or hidden) parameters.

- The **second** problem is more **common** in pattern recognition applications

General framework

- Assume that the **observed data X** is generated by some distribution.
- Assume that a complete dataset $\mathbf{Z} = (\mathbf{X},\mathbf{Y})$ exists as a **combination** of the **observed but incomplete data X** and the **missing data Y**.
- The **observations in Z** are assumed to be i.i.d. from the joint density

$$p(\mathbf{z}|\boldsymbol{\Theta}) = p(\mathbf{x}, \mathbf{y}|\boldsymbol{\Theta}) = p(\mathbf{y}|\mathbf{x}, \boldsymbol{\Theta})p(\mathbf{x}|\boldsymbol{\Theta}).$$

New likelihood function

- We can define a new likelihood function

$$\boxed{L(\mathbf{\Theta}|\mathcal{Z}) = L(\mathbf{\Theta}|\mathcal{X},\mathcal{Y}) = p(\mathcal{X},\mathcal{Y}|\mathbf{\Theta})}$$

  called the **complete-data likelihood** where $L(\theta|X)$ is referred to as the **incomplete-data likelihood**.

- The **EM** algorithm:
  - First, finds the **expected value** of the **complete-data log-likelihood** **using the current parameter** estimates (**expectation** **step**).
  - Then, **maximizes this expectation** (**maximization** **step**).

- Applying ML on the $E_{\mathbf{Y}|X,\theta^t}\{\log[L(\theta|Z)]\}$

- Maximum likelihood with **partial observation**

Expected value of the LL(θ|Z) w.r.t. the unknown data Y; **E-step**

- Define

$$Q(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)}) = E\left[\log p(\mathcal{X}, \mathcal{Y}|\mathbf{\Theta}) \,|\, \mathcal{X}, \mathbf{\Theta}^{(i-1)}\right]$$

as the **expected value** of the **complete-data log-likelihood** **w.r.t**. the **unknown data Y** **given** the **observed data X** and the **current** parameter estimates $\Theta^{(i-1)}$; $E_{\mathbf{Y}|X,\theta^t}\{\log[L(\theta|Z)]\}$

- The expected value can be computed as

$$E\left[\log p(\mathcal{X}, \mathcal{Y}|\mathbf{\Theta})|\mathcal{X}, \mathbf{\Theta}^{(i-1)}\right] = \int \log p(\mathcal{X}, \mathbf{y}|\mathbf{\Theta})\, p(\mathbf{y}|\mathcal{X}, \mathbf{\Theta}^{(i-1)})\, d\mathbf{y}.$$

y has been marginalized out

M-step

- Then, the expectation can be **maximized** by finding optimum values for the new parameters $\boldsymbol{\Theta}$ as

$$\boldsymbol{\Theta}^{(i)} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i-1)}).$$

- These two steps are **repeated iteratively** where each iteration is **guaranteed to increase** the log-likelihood.
- The **EM algorithm** is also guaranteed to converge to **a local maximum** of the likelihood function.
- Q and L(θ) **behave similarly**; so we run optimization on Q
- Usually looking for **analytical solution** in M-step
- EM can be considered as **Quasi-static** solution for parameter estimation

Convergence properties of EM

- The solution depends on the initial estimate θ_0
- At each iteration, a value of θ is computed so that the likelihood function does not decrease.
- The algorithm is guaranteed to be stable (i.e., does not **oscillate**).
- There is **no guarantee** that it will convergence to a global maximum.

Generalized Expectation-Maximization

- **Instead** of maximizing $Q(\Theta,\Theta^{(i-1)})$, the **Generalized Expectation-Maximization** algorithm finds some set of parameters $\Theta^{(i)}$ that satisfy

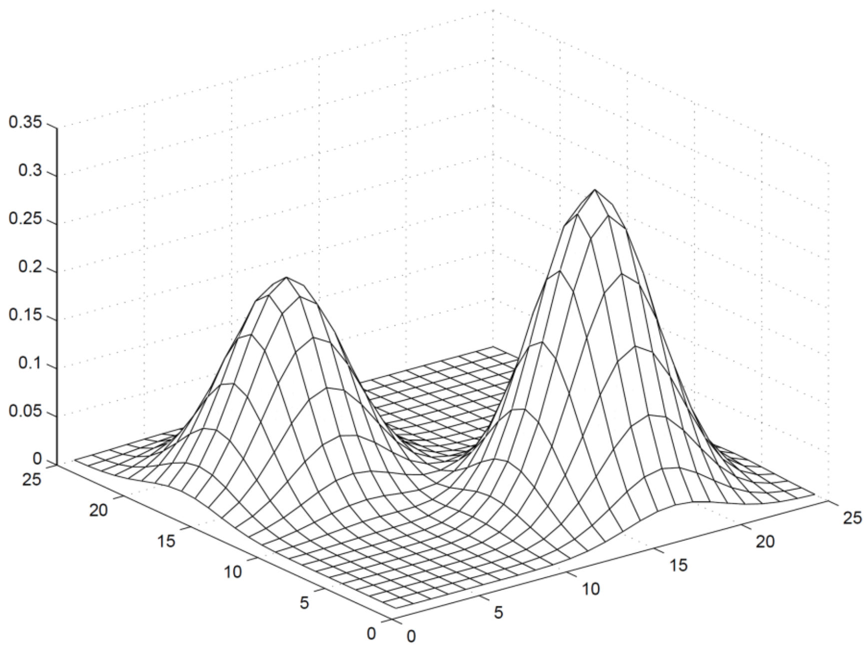
$$Q(\boldsymbol{\Theta}^{(i)}, \boldsymbol{\Theta}^{(i-1)}) > Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i-1)})$$

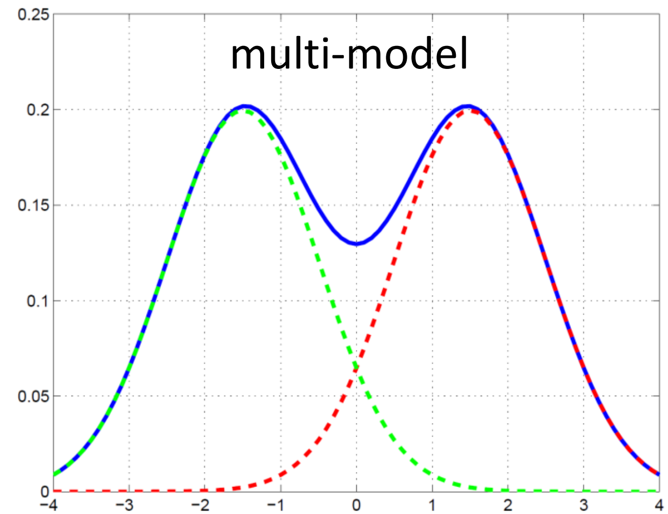
  at **each iteration**.

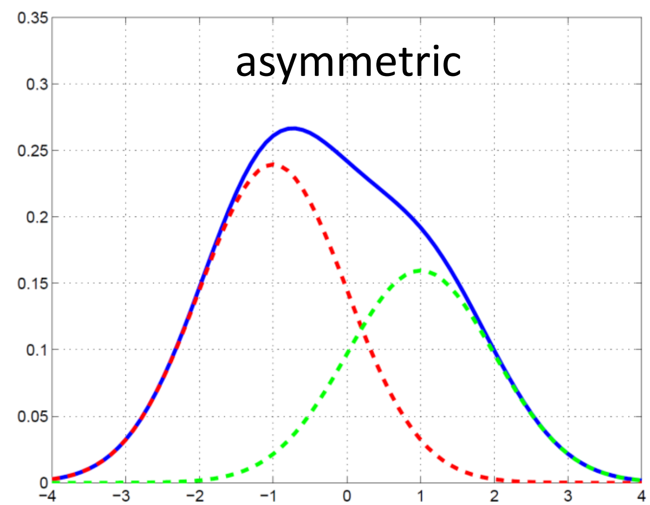
- **Convergence** will not be as rapid as the EM algorithm but it allows **greater flexibility** to choose **computationally simpler steps**.

Gaussians do not well model ! multi-model and asymmetric data

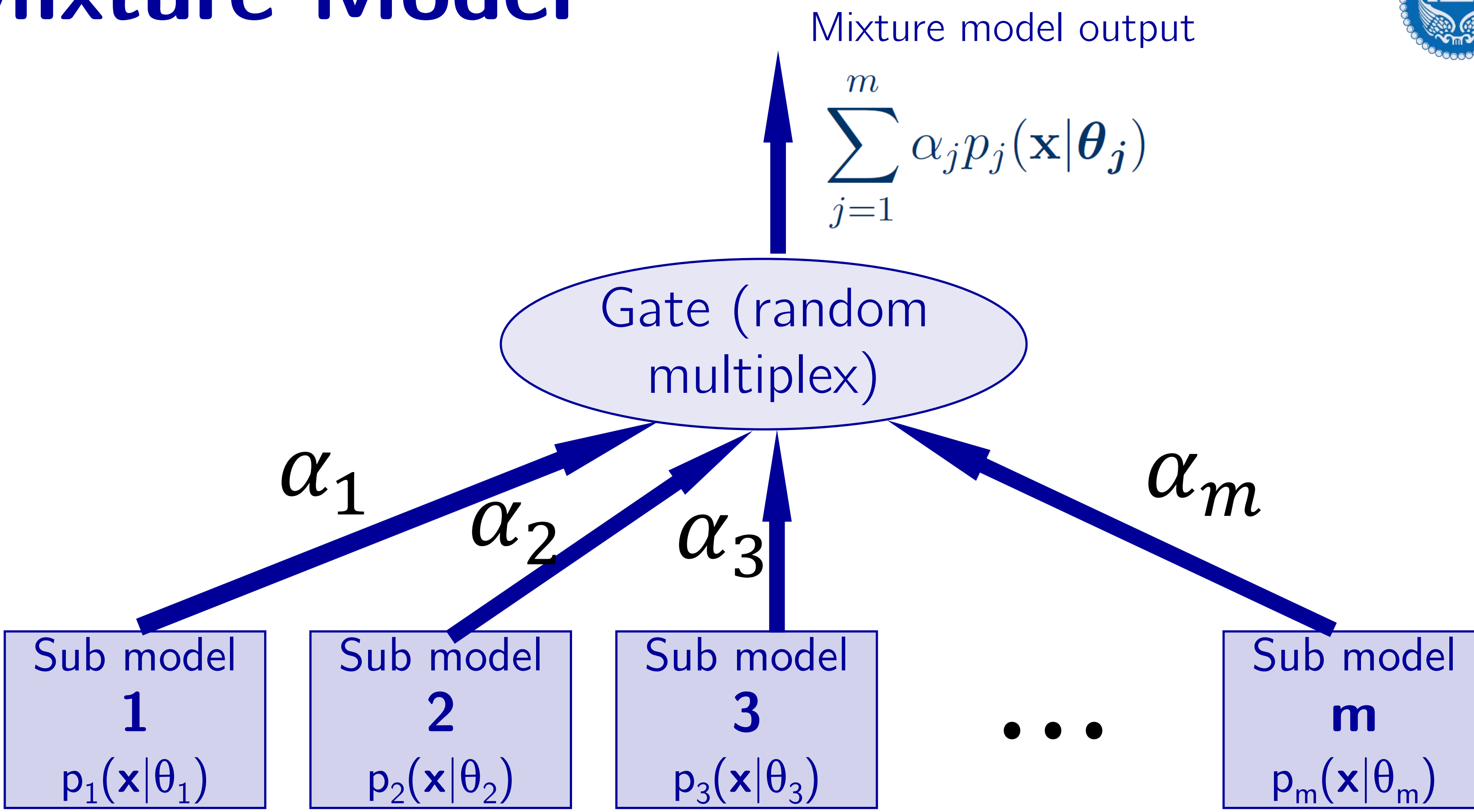
The performance of a **generative model** is highly dependent on the accuracy of the class-conditional PDFs, $p(x|\omega)$







Mixture Model



$$p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{m} \alpha_j p_j(\mathbf{x}|\boldsymbol{\theta_j})$$

Mixture Densities

- A mixture model is a **linear combination** of m densities

$$p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{m} \alpha_j p_j(\mathbf{x}|\boldsymbol{\theta_j})$$

where $\boldsymbol{\Theta} = (\alpha_1, \ldots, \alpha_m, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_m})$ such that $\alpha_j \geq 0$ and $\sum_{j=1}^{m} \alpha_j = 1$.

- $\alpha_1, \ldots, \alpha_m$ are called the **mixing parameters**.
- $p_j(\mathbf{x}|\theta_j)$, $j = 1, \ldots, m$ are called the **component densities**

The log-likelihood of mixture density

- Suppose that $X = \{x_1, \ldots, x_n\}$ is a set of observations i.i.d. with distribution $p(x|\Theta)$.
- The log-likelihood function of Θ becomes

$$\log L(\boldsymbol{\Theta}|\mathcal{X}) = \log \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\Theta}) =$$

- We **cannot** obtain an **analytical** solution for Θ by simply setting the **derivatives** of $\log L(\Theta|X)$ to zero because of the **logarithm of the sum**.

Mixture Density Estimation via EM

- Consider **X** as incomplete and define **hidden variables** Y= $\{y_i\}_{i=1}^{n}$ where $\mathbf{y_i}$ corresponds to **which mixture component** generated the data vector $\mathbf{x_i}$. ($y_i \in \{1,2,\dots,m\}$)
- In other words, $\mathbf{y_i} = \mathbf{j}$ if the i'th **data vector** was generated by the j'th **mixture component**.
- Then, the **log-likelihood** becomes:

$$\log L(\boldsymbol{\Theta}|\mathcal{X},\mathcal{Y}) = \log p(\mathcal{X},\mathcal{Y}|\boldsymbol{\Theta})$$

$$= \sum_{i=1}^{n} \log(p(\mathbf{x}_{\boldsymbol{i}}|y_i,\boldsymbol{\theta_i})p(y_i|\boldsymbol{\theta_i}))$$

$$= \sum_{i=1}^{n} \log(\alpha_{y_i} p_{y_i}(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{y_i}})).$$

Initial parameters and latent variable distribution in EM Mixture Density

- Assume we have the **initial parameter** estimates

$$\mathbf{\Theta}^{(g)} = (\alpha_1^{(g)}, \ldots, \alpha_m^{(g)}, \boldsymbol{\theta}_{\boldsymbol{1}}^{(g)}, \ldots, \boldsymbol{\theta}_{\boldsymbol{m}}^{(g)}).$$

- Compute

$$p(y_i|\mathbf{x}_{\boldsymbol{i}}, \mathbf{\Theta}^{(g)}) = \frac{\alpha_{y_i}^{(g)} p_{y_i}(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{y_i}}^{(g)})}{p(\mathbf{x}_{\boldsymbol{i}}|\mathbf{\Theta}^{(g)})} = \frac{\alpha_{y_i}^{(g)} p_{y_i}(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{y_i}}^{(g)})}{\sum_{j=1}^{m} \alpha_j^{(g)} p_j(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}_{\boldsymbol{j}}^{(g)})}$$

- and

$$p(\mathcal{Y}|\mathcal{X}, \mathbf{\Theta}^{(g)}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_{\boldsymbol{i}}, \mathbf{\Theta}^{(g)}).$$

$Q(\Theta, \Theta^{(g)})$ in EM Mixture Density

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(g)}) = \sum_{\mathbf{y}} \log p(\mathcal{X}, \mathbf{y}|\boldsymbol{\Theta}) p(\mathbf{y}|\mathcal{X}, \boldsymbol{\Theta}^{(g)})$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \log(\alpha_j p_j(\mathbf{x_i}|\boldsymbol{\theta_j})) p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)})$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \log(\alpha_j) p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)})$$

$$+ \sum_{j=1}^{m} \sum_{i=1}^{n} \log(p_j(\mathbf{x_i}|\boldsymbol{\theta_j})) p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)}).$$

Mixture Density Estimation via EM

- We can maximize the two sets of summations for α_j and θ_j **independently** because they are **not related**.
- The estimate for α_j can be computed as

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\log(\alpha_j)p(j|\mathbf{x}_i,\boldsymbol{\Theta}^{(g)})$$

lagrange multiplier method

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} p(j|\mathbf{x}_i,\boldsymbol{\Theta}^{(g)})$$

- where

$$p(j|\mathbf{x}_i,\boldsymbol{\Theta}^{(g)}) = \frac{\alpha_j^{(g)} p_j(\mathbf{x}_i|\boldsymbol{\theta}_j^{(g)})}{\sum_{t=1}^{m}\alpha_t^{(g)} p_t(\mathbf{x}_i|\boldsymbol{\theta}_t^{(g)})}.$$

It is a number and completely describe by $\theta^{(g)}$

Mixture of Gaussians

- We can obtain analytical expressions for θ_j for the special case of a **Gaussian mixture** where $\theta_j = (\mu_j, \Sigma_j)$ and

$$p_j(\mathbf{x}|\boldsymbol{\theta_j}) = p_j(\mathbf{x}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma_j}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_j})^T \boldsymbol{\Sigma_j}^{-1}(\mathbf{x}-\boldsymbol{\mu_j})\right].$$

- Equating the **partial derivative** of $Q(\Theta, \Theta^{(g)})$ with respect to μ_j to zero gives

$$+\sum_{j=1}^{m}\sum_{i=1}^{n}\log(p_j(\mathbf{x_i}|\boldsymbol{\theta_j}))p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)}).$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{j}} = \frac{\sum_{i=1}^{n} p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)})\mathbf{x_i}}{\sum_{i=1}^{n} p(j|\mathbf{x_i}, \boldsymbol{\Theta}^{(g)})}.$$

Mixture of Gaussian; Σ estimation

$$\boldsymbol{\Sigma}_j = \sigma^2 \mathbf{I}$$

$$\hat{\sigma}^2 = \frac{1}{nd} \sum_{j=1}^{m} \sum_{i=1}^{n} p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)}) \|\mathbf{x}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}}\|^2$$

$$\boldsymbol{\Sigma}_{\boldsymbol{j}} = \sigma_j^2 \mathbf{I}$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)}) \|\mathbf{x}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}}\|^2}{d \sum_{i=1}^{n} p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)})}$$

Mixture of Gaussian; Σ estimation

$$\boldsymbol{\Sigma_j} = \mathrm{diag}(\{\sigma^2_{jk}\}_{k=1}^{d})$$

$$\hat{\sigma}^2_{jk} = \frac{\sum_{i=1}^{n} p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)})(\mathbf{x}_{\boldsymbol{i}k} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}k})^2}{\sum_{i=1}^{n} p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)})}$$

$$\boldsymbol{\Sigma_j} = \boldsymbol{\Sigma}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)})(\mathbf{x}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})(\mathbf{x}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})^T$$

Mixture of Gaussian; general case

$\boldsymbol{\Sigma}_{\boldsymbol{j}} =$ arbitrary

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{j}} = \frac{\sum_{i=1}^{n} p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)})(\mathbf{x}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})(\mathbf{x}_{\boldsymbol{i}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{j}})^T}{\sum_{i=1}^{n} p(j|\mathbf{x}_{\boldsymbol{i}}, \boldsymbol{\Theta}^{(g)})}$$

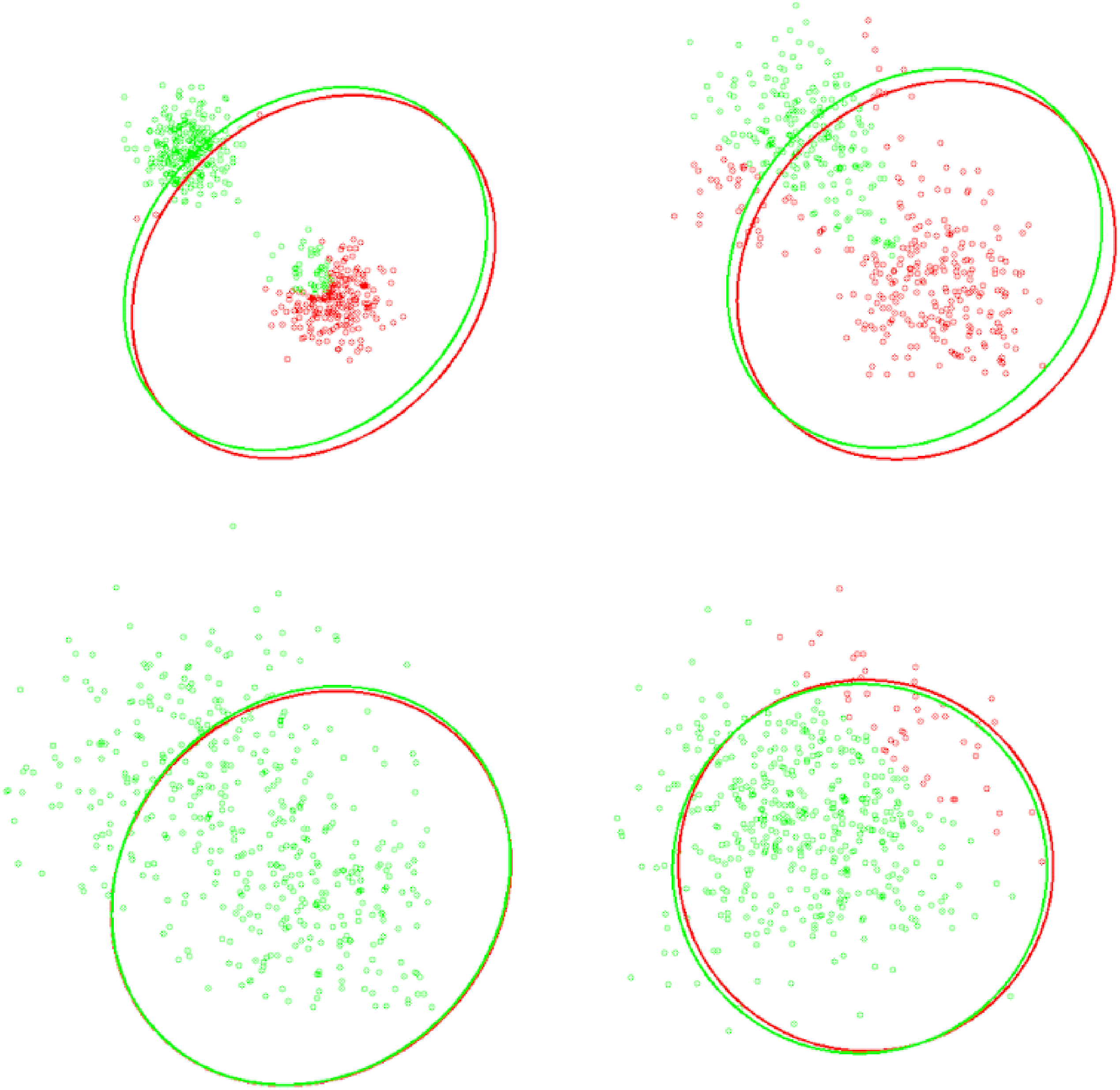
Mixture of Gaussians

- Estimates for α_j , μ_j and Σ_j perform **both expectation and maximization steps simultaneously**.
- EM iterations proceed by using the **current estimates** as the initial estimates for the next iteration.
  - The **priors are** computed from the proportion of examples belonging to each mixture component.
  - The **means are** the component centroids.
  - The **covariance matrices** are calculated as the sample covariance of the points associated with each component

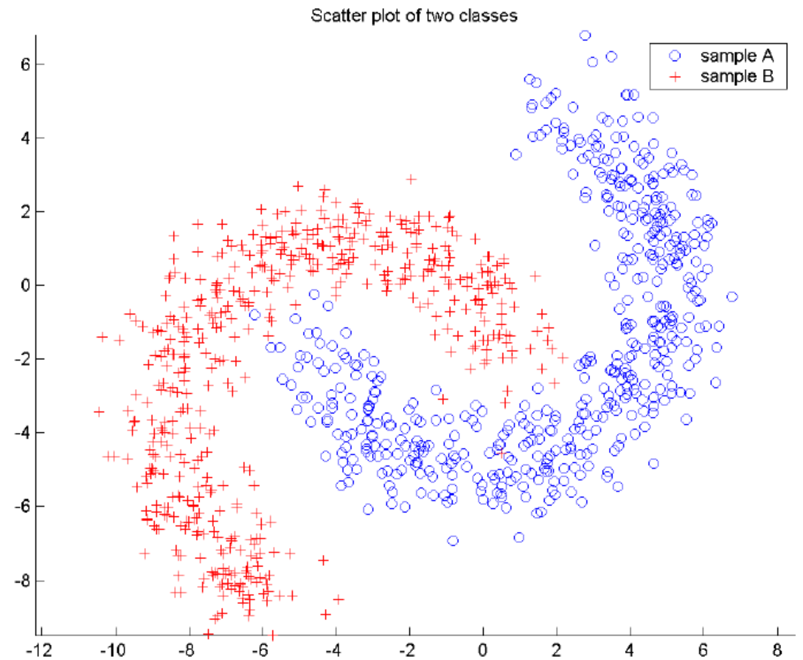
Mixture of Gaussians; an iterative algorithm

- The **number** of components in the mixture?
- The **initial** estimates for Θ?
- When to **stop** the iterations?
    - Stop if the **change** in log-likelihood between two iterations is less **than a threshold**.
    - Or, use a threshold for the **number of iterations**

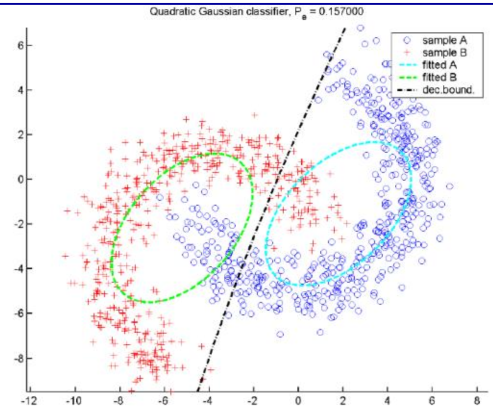
Example



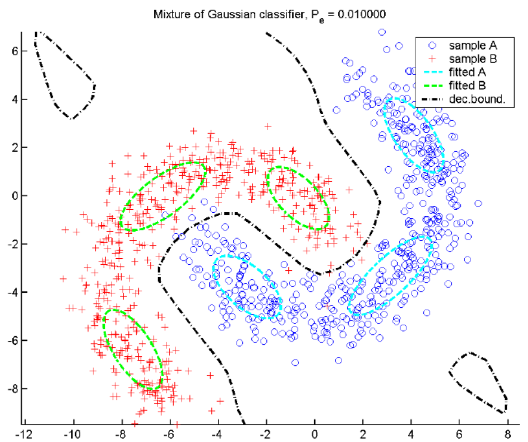
2-D Bayesian classification examples where the data for each class come from a banana shaped distribution



(a) Scatter plot.



(b) Quadratic Gaussian classifier with $P_e = 0.1570$.



(c) Mixture of Gaussian classifier with $P_e = 0.0100$.