

Parametric Models

Part II: Expectation-Maximization and Mixture Density Estimation

Selim Aksoy

Department of Computer Engineering
Bilkent University
saksoy@cs.bilkent.edu.tr

CS 551, Spring 2009



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

$$P(w_i | \mathbf{x}_g) = \frac{\int P(w_i | \mathbf{x}_g, \mathbf{x}_b) p(\mathbf{x}_g, \mathbf{x}_b) d\mathbf{x}_b}{\int p(\mathbf{x}_g, \mathbf{x}_b) d\mathbf{x}_b}.$$



Expectation-Maximization

- ▶ We can also extend maximum likelihood techniques to allow learning of parameters when some training patterns have missing features.
- ▶ The *Expectation-Maximization (EM)* algorithm is a general iterative method of finding the maximum likelihood estimates of the parameters of a distribution from training data.



Expectation-Maximization

- ▶ There are two main applications of the EM algorithm:
 - ▶ Learning when the data is incomplete or has missing values.
 - ▶ Optimizing a likelihood 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.



Expectation-Maximization

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

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



Expectation-Maximization

- ▶ We can define a new likelihood function

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

called the complete-data likelihood where $L(\Theta|\mathcal{X})$ 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).



Expectation-Maximization

- Define

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

as the expected value of the complete-data log-likelihood w.r.t. the unknown data \mathcal{Y} given the observed data \mathcal{X} and the current parameter estimates $\Theta^{(i-1)}$.

- The expected value can be computed as

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

- This is called the *E-step*.



Expectation-Maximization

- ▶ Then, the expectation can be maximized by finding optimum values for the new parameters Θ as

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

- ▶ This is called the *M-step*.
- ▶ 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.



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

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

at each iteration.

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



Mixture Densities

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

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

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

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



Mixture Densities

- ▶ Suppose that $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is a set of observations i.i.d. with distribution $p(\mathbf{x}|\Theta)$.
- ▶ The log-likelihood function of Θ becomes

$$\log L(\Theta|\mathcal{X}) = \log \prod_{i=1}^n p(\mathbf{x}_i|\Theta) = \sum_{i=1}^n \log \left(\sum_{j=1}^m \alpha_j p_j(\mathbf{x}_i|\theta_j) \right).$$

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



Mixture Density Estimation via EM

- ▶ Consider \mathcal{X} as incomplete and define hidden variables $\mathcal{Y} = \{y_i\}_{i=1}^n$ where y_i corresponds to which mixture component generated the data vector \mathbf{x}_i .
- ▶ In other words, $y_i = j$ if the i 'th data vector was generated by the j 'th mixture component.
- ▶ Then, the log-likelihood becomes

$$\begin{aligned}\log L(\boldsymbol{\Theta}|\mathcal{X}, \mathcal{Y}) &= \log p(\mathcal{X}, \mathcal{Y}|\boldsymbol{\Theta}) \\ &= \sum_{i=1}^n \log(p(\mathbf{x}_i|y_i, \boldsymbol{\theta}_i)p(y_i|\boldsymbol{\theta}_i)) \\ &= \sum_{i=1}^n \log(\alpha_{y_i} p_{y_i}(\mathbf{x}_i|\boldsymbol{\theta}_{y_i})).\end{aligned}$$



Mixture Density Estimation via EM

- Assume we have the initial parameter estimates

$$\Theta^{(g)} = (\alpha_1^{(g)}, \dots, \alpha_m^{(g)}, \theta_1^{(g)}, \dots, \theta_m^{(g)}).$$

- Compute

$$p(y_i | \mathbf{x}_i, \Theta^{(g)}) = \frac{\alpha_{y_i}^{(g)} p_{y_i}(\mathbf{x}_i | \theta_{y_i}^{(g)})}{p(\mathbf{x}_i | \Theta^{(g)})} = \frac{\alpha_{y_i}^{(g)} p_{y_i}(\mathbf{x}_i | \theta_{y_i}^{(g)})}{\sum_{j=1}^m \alpha_j^{(g)} p_j(\mathbf{x}_i | \theta_j^{(g)})}$$

and

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



Mixture Density Estimation via EM

- Then, $Q(\Theta, \Theta^{(g)})$ takes the form

$$\begin{aligned} Q(\Theta, \Theta^{(g)}) &= \sum_{\mathbf{y}} \log p(\mathcal{X}, \mathbf{y} | \Theta) p(\mathbf{y} | \mathcal{X}, \Theta^{(g)}) \\ &= \sum_{j=1}^m \sum_{i=1}^n \log(\alpha_j p_j(\mathbf{x}_i | \theta_j)) p(j | \mathbf{x}_i, \Theta^{(g)}) \\ &= \sum_{j=1}^m \sum_{i=1}^n \log(\alpha_j) p(j | \mathbf{x}_i, \Theta^{(g)}) \\ &\quad + \sum_{j=1}^m \sum_{i=1}^n \log(p_j(\mathbf{x}_i | \theta_j)) p(j | \mathbf{x}_i, \Theta^{(g)}). \end{aligned}$$



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

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

where

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



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

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

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

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



Mixture of Gaussians

- ▶ We consider five models for the covariance matrix Σ_j :
 - ▶ $\Sigma_j = \sigma^2 \mathbf{I}$

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

- ▶ $\Sigma_j = \sigma_j^2 \mathbf{I}$

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



Mixture of Gaussians

- Covariance models continued:

- $\Sigma_j = \text{diag}(\{\sigma_{jk}^2\}_{k=1}^d)$

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

- $\Sigma_j = \Sigma$

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

- $\Sigma_j = \text{arbitrary}$

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



Mixture of Gaussians

► Summary:

- 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

► Questions:

- How can we find the number of components in the mixture?
- How can we find the initial estimates for Θ ?
- How do we know 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.



Examples

- ▶ Mixture of Gaussians examples
- ▶ 1-D Bayesian classification examples
- ▶ 2-D Bayesian classification examples



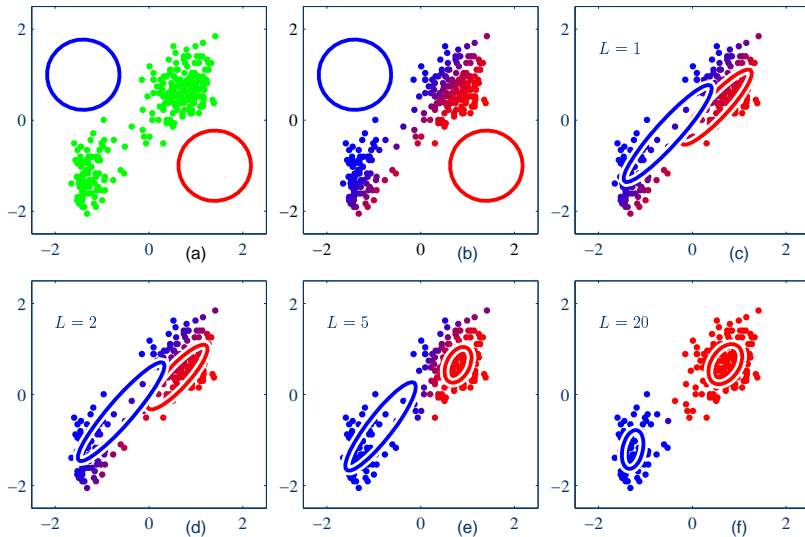
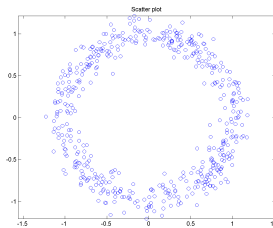
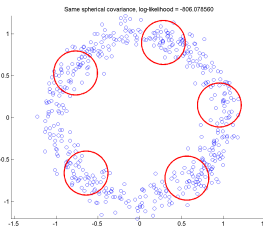


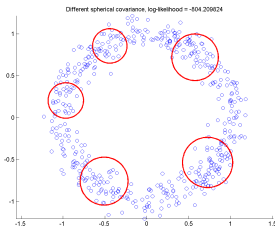
Figure 1: Illustration of the EM algorithm iterations for a mixture of two Gaussians.



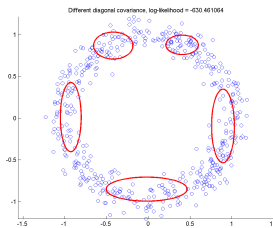
(a) Scatter plot.



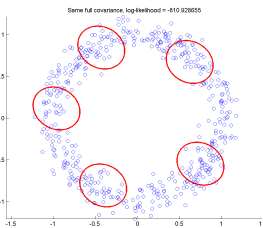
(b) Same spherical covari-
ance, log-likelihood = -806.08.



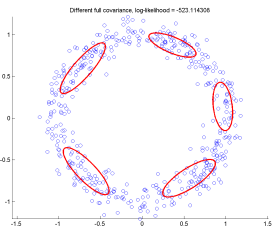
(c) Different spherical covari-
ance, log-likelihood = -804.21.



(d) Different diagonal covari-
ance, log-likelihood = -630.46.

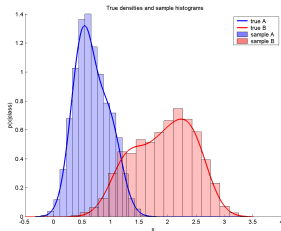


(e) Same arbitrary covariance,
log-likelihood = -810.93.

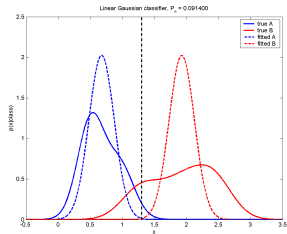


(f) Different arbitrary covari-
ance, log-likelihood = -523.11.

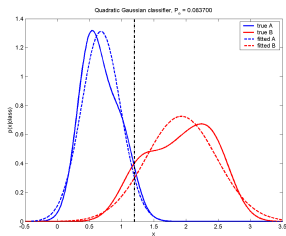
Figure 2: Fitting mixtures of 5 Gaussians to data from a circular distribution



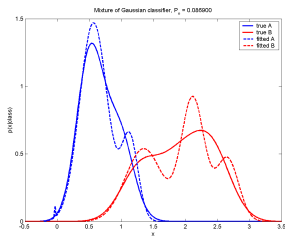
(a) True densities and sample histograms.



(b) Linear Gaussian classifier with $P_e = 0.0914$.



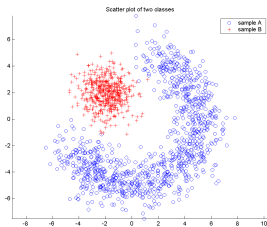
(c) Quadratic Gaussian classifier with $P_e = 0.0837$.



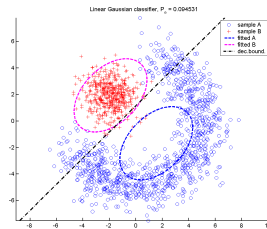
(d) Mixture of Gaussian classifier with $P_e = 0.0869$.

Figure 3: 1-D Bayesian classification examples where the data for each class come from a mixture of three Gaussians. Bayes error is $P_e = 0.0828$.

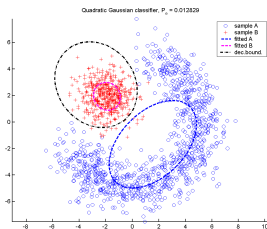




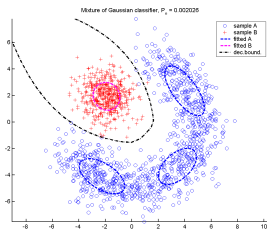
(a) Scatter plot.



(b) Linear Gaussian classifier with $P_e = 0.094531$.



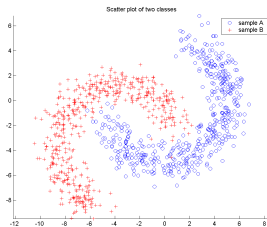
(c) Quadratic Gaussian classifier with $P_e = 0.012829$.



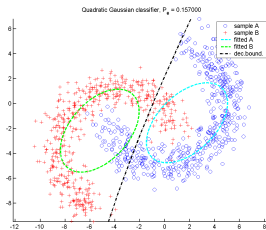
(d) Mixture of Gaussian classifier with $P_e = 0.002026$.

Figure 4: 2-D Bayesian classification examples where the data for the classes come from a banana shaped distribution and a bivariate Gaussian

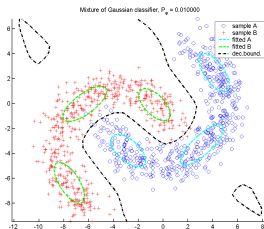




(a) Scatter plot.



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



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

Figure 5: 2-D Bayesian classification examples where the data for each class come from a banana shaped distribution.