Parametric Models
Part II: Expectation-Maximization and Mixture Density Estimation

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

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)p(x|\Theta).$$
Expectation-Maximization

• We can define a new likelihood function

\[ L(\Theta|Z) = L(\Theta|X, Y) = p(X, Y|\Theta) \]

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).
Expectation-Maximization

• Define

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

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 \( \Theta^{(i-1)} \).

• The expected value can be computed as

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

• This is called the E-step.
Expectation-Maximization

• Then, the expectation can be maximized by finding optimum values for the new parameters $\Theta$ 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

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

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

• $\alpha_1, \ldots, \alpha_m$ are called the mixing parameters.

• $p_j(x|\theta_j), \ j = 1, \ldots, m$ are called the component densities.
Mixture Densities

• Suppose that $\mathcal{X} = \{x_1, \ldots, x_n\}$ is a set of observations i.i.d. with distribution $p(x|\Theta)$.

• The log-likelihood function of $\Theta$ becomes

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

• We cannot obtain an analytical solution for $\Theta$ 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 $x_i$.

• 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

$$
\log L(\Theta | \mathcal{X}, \mathcal{Y}) = \log p(\mathcal{X}, \mathcal{Y} | \Theta) \\
= \sum_{i=1}^n \log(p(x_i | y_i, \theta_i)p(y_i | \theta_i)) \\
= \sum_{i=1}^n \log(\alpha_{y_i} p_{y_i}(x_i | \theta_{y_i})).
$$
Mixture Density Estimation via EM

- Assume we have the initial parameter estimates
  \[ \Theta^{(g)} = (\alpha_1^{(g)}, \ldots, \alpha_m^{(g)}, \theta_1^{(g)}, \ldots, \theta_m^{(g)}) \].

- Compute

  \[ p(y_i | x_i, \Theta^{(g)}) = \frac{\alpha_{y_i}^{(g)} p_{y_i}(x_i | \theta^{(g)}_{y_i})}{p(x_i | \Theta^{(g)})} = \frac{\alpha_{y_i}^{(g)} p_{y_i}(x_i | \theta^{(g)}_{y_i})}{\sum_{j=1}^m \alpha_j^{(g)} p_j(x_i | \theta^{(g)}_j)} \]

  and

  \[ p(Y | X, \Theta^{(g)}) = \prod_{i=1}^n p(y_i | x_i, \Theta^{(g)}). \]
Mixture Density Estimation via EM

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

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

$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \log(\alpha_{j}p_{j}(\mathbf{x}_{i}|\theta_{j}))p(j|\mathbf{x}_{i}, \Theta^{(g)})$$

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

$$+ \sum_{j=1}^{m} \sum_{i=1}^{n} \log(p_{j}(\mathbf{x}_{i}|\theta_{j}))p(j|\mathbf{x}_{i}, \Theta^{(g)}).$$
Mixture Density Estimation via EM

- We can maximize the two sets of summations for $\alpha_j$ and $\theta_j$ independently because they are not related.
- The estimate for $\alpha_j$ can be computed as

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

where

$$
p(j | x_i, \Theta^{(g)}) = \frac{\alpha_j^{(g)} p_j(x_i | \theta_j^{(g)})}{\sum_{t=1}^{m} \alpha_t^{(g)} p_t(x_i | \theta_t^{(g)})}.
$$
Mixture of Gaussians

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

$$p_j(x|\theta_j) = p_j(x|\mu_j, \Sigma_j)$$

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

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

$$\hat{\mu}_j = \frac{\sum_{i=1}^n p(j|x_i, \Theta^{(g)}) x_i}{\sum_{i=1}^n p(j|x_i, \Theta^{(g)})}.$$
Mixture of Gaussians

- We consider five models for the covariance matrix $\Sigma_j$:
  
  $\Sigma_j = \sigma^2 I$

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

  $\Sigma_j = \sigma_j^2 I$

  $$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} p(j \mid x_i, \Theta^{(g)}) \| x_i - \hat{\mu}_j \|^2}{d \sum_{i=1}^{n} p(j \mid x_i, \Theta^{(g)})}$$
Mixture of Gaussians

- Covariance models continued:
  - $\Sigma_j = \text{diag}(\{\sigma^2_{jk}\}_{k=1}^d)$
    \[
    \hat{\sigma}^2_{jk} = \frac{\sum_{i=1}^{n} p(j|x_i, \Theta^{(g)}) (x_{ik} - \hat{\mu}_{jk})^2}{\sum_{i=1}^{n} p(j|x_i, \Theta^{(g)})}
    \]
  - $\Sigma_j = \Sigma$
    \[
    \hat{\Sigma} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n} p(j|x_i, \Theta^{(g)}) (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T
    \]
  - $\Sigma_j = \text{arbitrary}$
    \[
    \hat{\Sigma}_j = \frac{\sum_{i=1}^{n} p(j|x_i, \Theta^{(g)}) (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T}{\sum_{i=1}^{n} p(j|x_i, \Theta^{(g)})}
    \]
Mixture of Gaussians

- **Summary:**
  - Estimates for $\alpha_j$, $\mu_j$ and $\Sigma_j$ 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 $\Theta$?
- 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: Fitting mixtures of 5 Gaussians to data from a circular distribution.
Figure 2: 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) 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 $P_e = 0.0869$. 

CS 551, Spring 2007 ©2007, Selim Aksoy
Figure 3: 2-D Bayesian classification examples where the data for the classes come from a banana shaped distribution and a bivariate Gaussian.
Figure 4: 2-D Bayesian classification examples where the data for each class come from a banana shaped distribution.