Lecture 7: EM Algorithm

Department of Statistics & Biostatistics
Rutgers University

Oct 18, 2011
Outline

1. Generative Algorithms

2. EM Algorithm
Generative Method

- Discriminative methods focus on the decision boundary directly (perceptron algorithm), or on the conditional distribution of $G$ given $X$ (logistic regression), while the distribution of $X$ is not considered.
- For generative methods, the joint distribution of $(X, G)$ is of interest.
- Bayes Theorem.

\[
P(G = k | X = x) = \frac{P(X = x | G = k)P(G = k)}{\sum_{l=1}^{K} P(X = x | G = l)P(G = l)}.
\]

- In general, assume the conditional density of $X$ given $G = k$ is given by $f_k(x)$. Set $\pi_k = P(G = k)$. We have

\[
P(G = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.
\]
Generative Method: Linear Discriminant Analysis

- **Gaussian discriminant analysis** assumes
  \[ f_k(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma_k)^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right\}. \]

- **Linear discriminant analysis** also assumes \( \Sigma_k = \Sigma \) for all \( 1 \leq k \leq K \).

- To compare two classes \( k \) and \( l \), it suffices to look at the log odds ratio
  \[ \log \frac{P(G = k|X = x)}{P(G = l|X = x)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2} (\mu_k + \mu_l) \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l) \]

- The **discriminant functions** are given by
  \[ \delta_k(x) = \log(\pi_k) - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \mu_k^T \Sigma^{-1} x, \quad 1 \leq k \leq K. \]

- The parameters \( \{\pi_1, \mu_1, \ldots, \pi_k, \mu_k, \Sigma\} \) are estimated by MLE.
  \[ \hat{\pi}_k = \frac{1}{N} \sum_{i=1}^{N} I\{g_i = k\} \]
  \[ \hat{\mu}_k = \frac{\sum_{i=1}^{N} x_i I\{g_i = j\}}{\sum_{i=1}^{N} I\{g_i = k\}} \]
  \[ \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu}_{g_i})^T (x_i - \hat{\mu}_{g_i}). \]
Comparison: Logistic Regression and LDA

- **Logistic Regression**

\[
\log \frac{P(G = k | X = x)}{P(G = K | X = x)} = \beta_{k,0} + \beta_k^T x
\]

- **LDA**

\[
\log \frac{P(G = k | X = x)}{P(G = K | X = x)} = \log \frac{\pi_k}{\pi_K} - \frac{1}{2} (\mu_k + \mu_K) \Sigma^{-1} (\mu_k - \mu_K) + x^T \Sigma^{-1} (\mu_k - \mu_K) + \beta_{k,0} + \beta_k^T x
\]

When Gaussian assumption is correct, LDA is more efficient.
Logistic regression needs less assumption, and is more robust.
MLE of logistic regression may not exist.
LDA can take unobserved labels (missing data) into account.
**Generative Method: Naive Bayes**

- $X = (X_1, \ldots, X_p)$, where each feature $X_j$ is discrete.
- Consider the simplest case where all $X_j$ are binary, and $G$ is also binary.
- **Naive Bayes** assumes that $X_1, \ldots, X_p$ are conditionally independent given $G$. That is, for $z = (z_1, \ldots, z_p)^T \in \{0, 1\}^p$

\[
P(X_1 = z_1, \ldots, X_p = z_p | G = 0) = \prod_{j=1}^{p} \left[ \pi_{j,0} z_j (1 - \pi_{j,0})^{1-z_j} \right]
\]

\[
P(X_1 = z_1, \ldots, X_p = z_p | G = 1) = \prod_{j=1}^{p} \left[ \pi_{j,1} z_j (1 - \pi_{j,1})^{1-z_j} \right].
\]

- Assuming $P(G = 1) = \pi$, we then have a joint distribution of $(X, G)$.
- The discriminant function is given by

\[
P(G = 1 | X = z) = \frac{\pi \prod_{j=1}^{p} \left[ \pi_{j,1} z_j (1 - \pi_{j,1})^{1-z_j} \right]}{\pi \prod_{j=1}^{p} \left[ \pi_{j,1} z_j (1 - \pi_{j,1})^{1-z_j} \right] + (1 - \pi) \prod_{j=1}^{p} \left[ \pi_{j,0} z_j (1 - \pi_{j,0})^{1-z_j} \right]}
\]
Training set \((x_1, g_1), \ldots, (x_N, g_N)\). MLE

\[
\pi = \frac{\sum_{i=1}^{N} I\{g_i = 1\}}{N}
\]

\[
\pi_{j,0} = \frac{\sum_{i=1}^{N} I\{g_i = 0, x_{ij} = 1\}}{\sum_{i=1}^{N} I\{g_i = 0\}}
\]

\[
\pi_{j,1} = \frac{\sum_{i=1}^{N} I\{g_i = 1, x_{ij} = 1\}}{\sum_{i=1}^{N} I\{g_i = 1\}}
\]

Laplace smoothing.

\[
\pi_{j,0} = \frac{\sum_{i=1}^{N} I\{g_i = 0, x_{ij} = 1\} + 1}{\sum_{i=1}^{N} I\{g_i = 0\} + 2}
\]

\[
\pi_{j,1} = \frac{\sum_{i=1}^{N} I\{g_i = 1, x_{ij} = 1\} + 1}{\sum_{i=1}^{N} I\{g_i = 1\} + 2}
\]

“Round” probabilities smaller than 0.001 and larger than 0.999 to 0.001 and 0.999, respectively.
LDA and NB make stringent assumptions. To get more flexible models, we could work in the enlarged feature space.

Another popular approach is to model the distribution of each $X|G = k$ as a mixture.

**Gaussian Mixtures.**

$$f(x) = \sum_{m=1}^{M} \pi_m f_m(x), \text{ where } f_m(\cdot) \sim N(\mu_m, \Sigma_m)$$

**Bernoulli Mixtures.** Let $z = (z_1, \ldots, z_p)^T \in \{0, 1\}^d$.

$$P(X = z) = \sum_{m=1}^{M} \phi_m \prod_{j=1}^{p} \left[ \pi_{j,m}^{z_j} (1 - \pi_{j,m})^{1-z_j} \right].$$
Outline

1. Generative Algorithms

2. EM Algorithm
Suppose there is a latent variable $Z$ which takes values in $\{1, \ldots, M\}$. Assume the joint distribution of $(X, Z)$ is given by

- Given $Z = m$, the conditional distribution of $X$ is $N(\mu_m, \Sigma_m)$.
- $Z$ has the multinomial distribution $P(Z = m) = \pi_m$.

The marginal density of $X$ is then given by

$$f(x) = \sum_{m=1}^{M} \pi_m f_m(x), \quad \text{where} \quad f_m(\cdot) \sim N(\mu_m, \Sigma_m).$$

Alternatively, think $Z = (Z_1, \ldots, Z_M)^T$ as a $M$-dimensional random vector. Define $e_m \in \mathbb{R}^M$ be the vector whose $m$-th entry is one and all other entries are zero. The distribution of $Z$ is given by

$$P(Z = e_m) = \pi_m.$$ 

The joint density of $(X, Z)$ is given by

$$f(x, z) = \prod_{m=1}^{M} \pi_m^{z_m} [f_m(x)]^{z_m}.$$
EM Algorithm

- Training set: \((x_1, z_1), \ldots, (x_N, z_n)\), where \(z_i = (z_{i1}, \ldots, z_{iM})^T\).

- The log likelihood of the complete data is given by

\[
\ell(X, Z; \theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} z_{im} \left[ \log(\pi_m) + \log f_m(x_i) \right]
\]

- **E-Step.** Let \(\gamma(z_{im}) = \mathbb{E}(z_{im} = 1|x_i; \theta^{\text{old}})\), and compute

\[
\ell_{Z|X;\theta}(X; \theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} \gamma(z_{im}) \left[ \log(\pi_m) + \log f_m(x_i) \right]
\]

- **M-Step.** View \(\gamma(z_{im})\) as fixed, maximize \(\ell_{Z|X;\theta}(X; \theta)\) over \(\theta\) to get the updated estimate \(\theta^{\text{new}}\).
EM Algorithm for Gaussian Mixtures

- E-Step.
  \[ \gamma(z_{im}) = \frac{\pi_{m}^{old} f_{m}(x; \mu_{m}^{old}, \Sigma_{m}^{old})}{\sum_{m'=1}^{M} \pi_{m'}^{old} f_{m'}(x; \mu_{m'}^{old}, \Sigma_{m'}^{old})}. \]

- M-Step.
  \[ \pi_{m}^{new} = \frac{\sum_{i=1}^{N} \gamma(z_{im})}{N} \]
  \[ \mu_{m}^{new} = \frac{\sum_{i=1}^{N} \gamma(z_{im}) x_{i}}{\sum_{i=1}^{N} \gamma(z_{im})} \]
  \[ \Sigma_{m}^{new} = \frac{\sum_{i=1}^{N} \gamma(z_{im}) (x_{i} - \mu_{m}^{new}) (x_{i} - \mu_{m}^{new})^{T}}{\sum_{i=1}^{N} \gamma(z_{im})} \]
Bernoulli Mixtures

\[ f(x; \theta) = \sum_{m=1}^{M} \phi_m f_m(x; \theta_m), \]

\[ f_m(x; \theta_m) = \prod_{j=1}^{p} \pi_{j,m}^{x_j} (1 - \pi_{j,m})^{1-x_j}, \]

where \( \theta = (\phi_1, \ldots, \phi_M, \theta_1, \ldots, \theta_M) \) and \( \theta_m = (\pi_{1,m}, \ldots, \pi_{p,m})^T \).