EE 6885 Statistical Pattern Recognition

Fall 2005
Prof. Shih-Fu Chang
http://www.ee.columbia.edu/~sfchang

Lecture 5 (9/21/05)

Reading
- Model Parameter Estimation
  - ML Estimation, Chap. 3.2
- Mixture of Gaussian and EM
  - Reference Book, HTF Chap. 8.5
  - Textbook, DHS 3.9
- Homework #2 due 2005-09-28, Wed
- No class/office hours next Monday, 2005-09-26
Multi-variate Gaussian

\[ p(x) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)} \]

Bayesian Classifiers

- Decision Boundaries for Gaussians

\[ w'(x - x_\omega) = 0 \quad w = \Sigma^{-1}(\mu_j - \mu_i) \quad x_\omega = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln[P(\omega_i) / P(\omega_o)]}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)} (\mu_i - \mu_j) \]

Missing Features by Marginalization

- \( x = [x_g, x_b] \), \( x_g \): good features, \( x_b \): bad features

compute \( P(w_i | x_g) = \frac{p(w_i, x_g)}{p(x_g)} = \frac{\int p(w_i, x_g, x_b) dx_b}{p(x_g)} = \frac{\int p(x_g, x_b | w_i) p(w_i) dx_b}{\int p(x_g, x_b) dx_b} \)

Joint prob. of \((\omega, x_g, x_b)\) marginalized over \(x_b\)

Parameter Estimation

- Parametric form of distribution
e.g., \( p(x | w_j) \sim N(\mu, \Sigma_j) \quad p(x | w_j) = p(x | w_j, \theta) \)

- How to estimate \( \theta \)?

\[ \text{learn from data samples} \quad D = \{x_1, x_2, ..., x_n\} \]

Likelihood

\[ l(\theta) = p(D | \theta) = \prod_{k=1}^{n} p(x_k | \theta) \quad \text{assume } x_1, ..., x_n \text{ independent} \]

Find \( \hat{\theta} = \arg \max_{\theta} p(D | \theta) = \arg \max_{\theta} \prod_{k=1}^{n} p(x_k | \theta) \)

\[ = \arg \max_{\theta} \sum_{k=1}^{n} \ln p(x_k | \theta) \]

- Use gradient operation

\[ \nabla_{\theta} l(\theta) = 0 \]
Case I: Gaussian: $\mu$ unknown

\[
\ln P(x_k | \mu) = -\frac{1}{2} \ln [(2\pi)^d |\Sigma|] - \frac{1}{2} (x_k - \mu)^T \Sigma^{-1} (x_k - \mu)
\]

and

\[
\nabla_{\theta_k} \ln P(x_k | \mu) = \Sigma^{-1} (x_k - \mu)
\]

\[
\nabla_{\phi_k} \sum_k \ln P(x_k | \mu) = 0
\]

\[
\sum_k \Sigma^{-1} (x_k - \mu) = 0
\]

\[
\hat{\mu} = \frac{\sum_k x_k}{n}
\]

ML estimator of the Gaussian mean is the sample mean

Case II: Gaussian Case: unknown $\mu$ and $\sigma$ (1D)

- $\theta = (\mu, \sigma^2)$

\[
\ln P(x_k | \theta) = -\frac{1}{2} \ln (2\pi \sigma^2) - \frac{1}{2\sigma^2} (x_k - \mu)^2
\]

\[
\nabla_{\theta} \ln P(x_k | \theta) = \left[ \frac{1}{\theta_2} (x_k - \mu) \right] - \frac{1}{2\sigma^2} \left( \frac{1}{\sigma^2} (x_k - \mu)^2 \right)
\]

\[
\sum_k \nabla_{\theta_k} \ln P(x_k | \theta) = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{\sum_k x_k}{n}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \hat{\mu})^2
\]

\[
\text{Exp}(\hat{\sigma}^2) = \frac{n-1}{n} \sigma^2 \neq \sigma^2 \quad \text{if } n \to \infty \text{ then } \text{Exp}(\hat{\sigma}^2) \to \sigma^2 \quad \text{Asymptotically unbiased}
\]

- Multi-Dimensional $\theta = (\mu, \Sigma)$

\[
\hat{\mu} = (1/n) \sum_k \hat{x}_k, \quad \hat{\Sigma} = (1/n) \sum_k (\hat{x}_k - \mu)(\hat{x}_k - \mu)^T
\]

ML estimator: mean -> sample mean, variance -> biased sample variance
**Mixture Of Gaussians**

- Real distributions seldom follow a single Gaussian \( \rightarrow \) mixture of Gaussians

\[
p(x) = \sum_{z} \pi_i p(x, z) = \sum_{z} \pi_i p(x | z)
\]

\[
= \sum_{z} \pi_i N(x | \mu_i, \Sigma_i) = \sum_{z} \pi_i \frac{1}{(2\pi)^{d/2} \sqrt{\mid\Sigma_i\mid}} e^{-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)}
\]

- Given data \( x_1, ..., x_N \), define log-likelihood:

\[
l = \sum_{n=1}^{N} \log(\pi_0 N(x_n | \mu_0, \Sigma_0) + \pi_i N(x_n | \mu_i, \Sigma_i))
\]

- Posterior probability of \( x \) being generated by a specific component

\[
posterior = \tau_i = p(z_i = 1 | x, \theta), \theta = \{\mu_0, \Sigma_0, \mu_i, \Sigma_i\}
\]

---

**Derivation of the E-M solution**

- Maximization of \( l(\theta) \) directly is hard due to log of sum
- Instead, look at

\[
\Delta l(\theta) = l(\theta) - l(\theta), \quad \theta: current \ estimation \ of \ \theta
\]

- **Jensen’s Inequality**

  If \( f \) is concave, \( f(E\{x\}) \geq E(f\{x\}) \)

  \[
f(E\{g(x)\}) \geq E(f\{g(x)\})
\]

  e.g., \( f(x) = \log(x) \)

\[
\log \left( \sum_i p_i x_i \right) \geq \sum_i p_i \log(x_i), \text{ where } \sum_i p_i = 1
\]

If \( f \) is convex, \( f(E\{x\}) \leq E(f\{x\}) \)
Auxiliary Function in E-M

\[ \Delta l(\theta) = l(\theta) - l(\theta_i) = \sum_{n=1}^{N} \log p(x_n | \theta) - \sum_{n=1}^{N} \log p(x_n | \theta_i) \]

\[ = \sum_{n=1}^{N} \log \frac{p(x_n | \theta)}{p(x_n | \theta_i)} = \sum_{n=1}^{N} \log \sum_{z} \frac{p(x_n, z | \theta)}{p(x_n | \theta_i)} \]

\[ = \sum_{n=1}^{N} \log \sum_{z} \frac{p(x_n, z | \theta) p(x_n | \theta)}{p(x_n | \theta_i) p(x_n, z | \theta_i)} \]

\[ = \sum_{n=1}^{N} \log \sum_{z} p(z | x_n, \theta_i) \frac{p(x_n, z | \theta)}{p(x_n, z | \theta_i)} \]

\[ \geq \sum_{n=1}^{N} \sum_{z} p(z | x_n, \theta_i) \log \frac{p(x_n, z | \theta)}{p(x_n, z | \theta_i)} \]

\[ = Q(\theta | \theta_i) \]

Note there is no log_of_sum. So taking derivative is easier.

E-M improves likelihood

- Auxiliary function derived based on Jensen’s Inequality,

\[ Q(\theta | \theta_i) = \sum_{n=1}^{N} \sum_{z} \frac{p(z | x_n, \theta_i) \log p(x_n, z | \theta) + \text{const}}{\text{expectation over } \tau_z \text{ with current } \theta_i} + \text{joint likelihood of observed & hidden} \]

- Now estimate \( \theta_{t+1} \) by maximizing \( Q \)

\[ \theta_{t+1} = \arg \max_{\theta} Q(\theta | \theta_i) \]

- So in the expectation step, compute \( r_n^z \), the ‘responsibility’ of component \( z \) for sample \( x_n \).

- In the maximization step, take derivative of \( Q \) to \( \theta \), and find the new estimate for \( \theta \) (Note only sum_of_log is involved)
EM Always Improves Likelihood

- Why does EM always improve $l(\theta)$?
  
  \[
  \Delta l(\theta_{t+1}) = l(\theta_{t+1}) - l(\theta_t) \geq Q(\theta_{t+1} | \theta_t) - Q(\theta_t | \theta_t) = 0 \quad \therefore \quad \Delta l(\theta_{t+1}) \geq 0
  \]

  \[
  Q(\theta | \theta_t) = \sum_{n=1}^N \sum_{i} p(z | x_n, \theta_t) \log p(x_n, z | \theta_t) + \text{const}
  \]

  - General steps of EM:
    - Define likelihood model with parameters $\theta$
    - Identify hidden variables $z$
    - Derive the auxiliary function and the E and M equations
    - In each iteration, estimate the posteriors of hidden variables
    - Re-estimate the model parameters. Repeat until stop

Expectation-Maximization (E-M) Solution of GMM

- EM for estimating $\theta$ and $\tau_i$.
  - Follow ‘divide and conquer’ principle. In iteration step $t$:

  **Expectation**: 
  \[
  \tau_n^{(t)} = \frac{\pi_i^{(t)} N(x_n | \mu_i^{(t)}, \Sigma_i^{(t)})}{\sum_j \pi_j^{(t)} N(x_n | \mu_j^{(t)}, \Sigma_j^{(t)})}
  \]
  Weight from component $j$

  **Maximation**: 
  \[
  \pi_i^{(t+1)} = \frac{\sum_n \tau_n^{(t)} x_n}{\sum_n \tau_n^{(t)}}
  \]
  Divide data to each group, Compute mean and variance from each group

  \[
  \mu_i^{(t+1)} = \frac{\sum_n \tau_n^{(t)} (x_n - \mu_i^{(t)})(x_n - \mu_i^{(t)})^T}{\sum_n \tau_n^{(t)}}
  \]

EE6887-Chang 5-11
GMM for Clustering

- Given the estimated GMM model, compute \( p(x) \), \( \theta = \{ \mu_0, \Sigma_0, \mu_i, \Sigma_i \} \)

\[
p(z_i \mid x) = \sum_{\theta} \pi_i \mathcal{N}(x \mid \mu_i, \Sigma_i)
\]

- Estimate the probability that \( x \) is generated by cluster \( i \)

\[
\pi_i^{(i)} = \frac{\pi_i^{(i)} N(x_n \mid \mu_i, \Sigma_i)}{\sum_j \pi_j^{(i)} N(x_n \mid \mu_j, \Sigma_j)}
\]

- Each sample is assigned to every cluster with a ‘soft’ decision.

Comparison: K-Mean Clustering

- Training data \( \{x_i\} + \{\text{label}(i) ? \} \)

- Unsupervised learning

- K-mean clustering
  - Fix K values
  - Initialize the representative of each cluster
  - Map samples to closest cluster (hard decision)
  - Re-compute the centers

\[
\text{for } i = 1, 2, ..., N, \\
x_i \rightarrow C_k, \text{if } \text{Dist}(x_i, C_k) < \text{Dist}(x_i, C_{k'})
\]

end

EE6887-Chang

5-13

5-14