EE 6885 Statistical Pattern Recognition

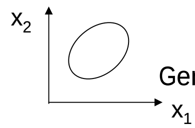
Fall 2005
Prof. Shih-Fu Chang
http://www.ee.columbia.edu/~sfchang

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

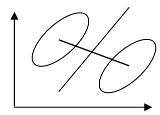
Multi-variate Gaussian

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\sqrt{|\mathbf{\Sigma}|}} e^{\left(\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}$$



General $\mathbf{\Sigma}$

Bayesian Classifiers



- Decision Boundaries for Gaussians

$$w^t(x - x_0) = 0 \qquad w = \Sigma^{-1}(\mu_i - \mu_j) \qquad x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

Missing Features by Marginalization

- $x=[x_g, x_b]$, x_g: good features, x_b : bad features

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

Joint prob. of (ω_i, x_g, x_b) marginalized over x_b

EE6887-Chang

5-3

Parameter Estimation

- Parametric form of distribution
  e.g., $p(x \mid w_i) \sim N(\mu_i, \Sigma_i) \qquad p(x \mid w_i) = p(x \mid w_i, \theta_i)$
- How to estimate θ_i ?
  →learn from data samples $D = \{x_1, x_2 ... x_n\}$

Likelihood $l(\theta) = p(D \mid \theta) = \prod_{k=1}^{n} p(x_k \mid \theta)$ assume $x_1, ..., x_n$ independent

$$\text{Find } \hat{\theta} = \arg\max_{\theta} p(D \mid \theta) = \arg\max \prod_{k=1}^{n} p(x_k \mid \theta)$$
$$= \arg\max \sum_{1}^{n} \ln p(x_k \mid \theta)$$

- Use gradient operation

$$\nabla_\theta = \begin{bmatrix} \partial / \partial\theta_1 \\ ... \\ \partial / \partial\theta_p \end{bmatrix} \qquad \nabla_\theta l(\theta) = 0$$

EE6887-Chang

5-4

Case I: Gaussian: μ unknown

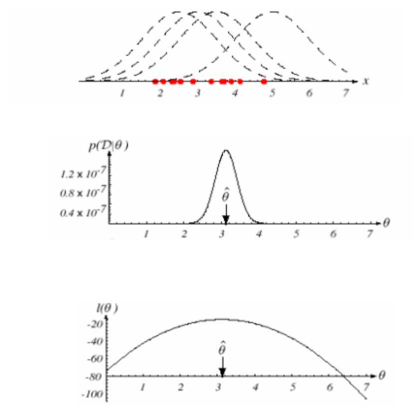
$$\ln P(x_k \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d |\Sigma|\right] - \frac{1}{2}(x_k-\mu)^t \Sigma^{-1}(x_k-\mu)$$

$$and \quad \nabla_{\theta\mu} \ln P(x_k \mid \mu) = \Sigma^{-1}(x_k-\mu)$$

$$\nabla_{\theta\mu} \sum_k \ln P(x_k \mid \mu) = 0$$

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

$$\hat{\mu} = \frac{\sum_k x_k}{n}$$



ML estimator of the Gaussian mean is the sample mean

Case II: Gaussian Case: *unknown* μ *and* σ (1D)

- $\theta = (\mu, \sigma^2)$ $\quad l = \ln P(x_k \mid \theta) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_k-\mu)^2$

$$\nabla_\theta l = \begin{pmatrix} \frac{\partial}{\partial\mu}(\ln P(x_k \mid \theta)) \\ \frac{\partial}{\partial\theta_2}(\ln P(x_k \mid \theta)) \end{pmatrix} = \begin{bmatrix} \frac{1}{\theta_2}(x_k-\mu) \\ -\frac{1}{2\sigma^2} - \frac{(x_k-\mu)^2}{2(\sigma^2)^2} \end{bmatrix}$$

$$\sum_k \nabla_\theta \ln P(x_k \mid \theta) = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{\sum_k x_k}{n} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})^2$$

$$Exp(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

if $n \to \infty$ then $Exp(\hat{\sigma}^2) \to \sigma^2$

Asymptotically unbiased

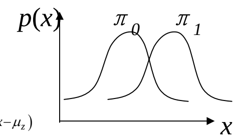
- Multi-Dimensional $\theta = (\mu, \Sigma)$

$$\Rightarrow \quad \hat{\mu} = (1/n)\sum_k \vec{x}_k \qquad \hat{\Sigma} = (1/n)\sum_k (\vec{x}_k-\mu)(\vec{x}_k-\mu)^t$$

ML estimator: mean -> sample mean, variance ->biased sample variance

Mixture Of Gaussians

- Real distributions seldom follow a single Gaussian
  → mixture of Gaussians



$$p(x)=\sum_z p(x,z) \quad =\sum_z p(z)p(x\,|\,z)$$

$$=\sum_z \pi_z N\left(x\middle|\mu_z,\Sigma_z\right) \quad =\sum_{z=1}^{Z}\pi_z \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma_z|}} e^{-\frac{1}{2}(x-\mu_z)^T\Sigma_z^{-1}(x-\mu_z)}$$

- Given data $x_1, \ldots, x_N$, define log-likelihood:

$$l=\sum_{n=1}^{N}\log\left(\pi_0 N(x_n|\mu_0,\Sigma_0)+\pi_1 N(x_n|\mu_1,\Sigma_1)\right)$$

- Posterior probability of x being generated by a specific component

$$posteriers = \tau^i = p\left(z_i = 1\middle|x,\theta\right), \quad \theta=\{\mu_0,\Sigma_0,\mu_1,\Sigma_1\}$$

(*responsiblity* of component *i*)

EE6887-Chang

5-7

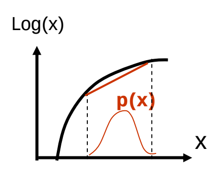
Derivation of the E-M solution

log likelihood $\quad l(\theta)=\sum_{n=1}^{N}\log\sum_z p(x_n,z\,|\,\theta)$

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

$$\Delta l(\theta)=l(\theta)-l(\theta_t), \quad \theta_t : current \text{ estimation of } \theta$$

- Jensen's Inequality



$$If \text{ f is concave, } \mathrm{f}\left(\mathrm{E}\{\mathrm{x}\}\right)\geq \mathrm{E}\left(\mathrm{f}\{\mathrm{x}\}\right)$$

$$\mathrm{f}\left(\mathrm{E}\{\mathrm{g}(\mathrm{x})\}\right)\geq \mathrm{E}\left(\mathrm{f}\{\mathrm{g}(\mathrm{x})\}\right)$$

$$e.g., \mathrm{f}(\mathrm{x})=\log(x)$$

$$\log\left(\sum_i p_i x_i\right)\geq\sum_i p_i\log(x_i), where \sum_i p_i = 1$$

$$If \text{ f is convex, } \mathrm{f}\left(\mathrm{E}\{\mathrm{x}\}\right)\leq \mathrm{E}\left(\mathrm{f}\{\mathrm{x}\}\right)$$

EE6887-Chang

5-8

Auxiliary Function in E-M

$$\Delta l(\theta) = l(\theta) - l(\theta_t) = \sum_{n=1}^{N} \log p(x_n \mid \theta) - \sum_{n=1}^{N} \log p(x_n \mid \theta_t)$$

$$= \sum_{n=1}^{N} \log \frac{p(x_n \mid \theta)}{p(x_n \mid \theta_t)} = \sum_{n=1}^{N} \log \sum_{z} \frac{p(x_n, z \mid \theta)}{p(x_n \mid \theta_t)}$$

marginalization

$$= \sum_{n=1}^{N} \log \sum_{z} \frac{p(x_n, z \mid \theta)}{p(x_n \mid \theta_t)} \frac{p(x_n, z \mid \theta_t)}{p(x_n, z \mid \theta_t)}$$

$$= \sum_{n=1}^{N} \log \sum_{z} p(z \mid x_n, \theta_t) \frac{p(x_n, z \mid \theta)}{p(x_n, z \mid \theta_t)}$$

$$\geq \sum_{n=1}^{N} \sum_{z} p(z \mid x_n, \theta_t) \log \frac{p(x_n, z \mid \theta)}{p(x_n, z \mid \theta_t)}$$

Jensen's inequality

$$= Q(\theta \mid \theta_t)$$

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

E-M improves likelihood

- Auxiliary function derived based on Jensen's Inequality,

$$Q(\theta \mid \theta_t) = \sum_{n=1}^{N} \sum_{z} \underbrace{p(z \mid x_n, \theta_t)}_{\substack{\text{expectation over} \\ \text{z with current } \theta_t}} \underbrace{\log p(x_n, z \mid \theta)}_{\substack{\text{joint likelihood of} \\ \text{observed \& hidden}}} + const$$

- Now estimate θ_{t+1} by maximizing Q

$$\theta_{t+1} = \arg\max_{\theta} Q(\theta \mid \theta_t)$$

- So in the expectation step, compute τ_n^z , the 'responsibility ' of component z for sample x_n
- In the maximization step, take derivative of Q to θ, and find the new estimate for θ *(Note only sum_of_log is involved)*

EM Always Improves Likelihood

- Why does EM always improve $l(\theta)$?

$$\Delta l(\theta_{t+1}) = l(\theta_{t+1}) - l(\theta_t) \geq Q(\theta_{t+1} | \theta_t)$$

$$Q(\theta_{t+1} | \theta_t) = \max_\theta Q(\theta | \theta_t) \geq Q(\theta_t | \theta_t) = 0 \quad \therefore \quad \Delta l(\theta_{t+1}) \geq 0$$

$$Q(\theta | \theta_t) = \sum_{n=1}^{N} \sum_z \underbrace{p(z | x_n, \theta_t)}_{\text{expectation over z with current } \theta_t} \underbrace{\log p(x_n, z | \theta)}_{\text{joint likelihood of observed \& hidden}} + const$$

- General steps of EM:
  - Define likelihood model with parameters θ
  - 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 θ and τ_i.
  - Follow 'divide and conquer' principle. In iteration step t:

$$Expectation: \tau_n^{i(t)} = \frac{\pi_i^{(t)} N(x_n | \mu_i^{(t)}, \Sigma_i^{(t)})}{\sum_j \pi_i^{(t)} N(x_n | \mu_j^{(t)}, \Sigma_j^{(t)})}$$

Weight from component i

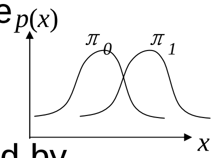
$$Maximation: \mu_i^{(t+1)} = \frac{\sum_n \tau_n^{i(t)} x_n}{\sum_n \tau_n^{i(t)}} \qquad \Sigma_i^{(t+1)} = \frac{\sum_n \tau_n^{i(t)} (x_n - \mu_i^{(t)})(x_n - \mu_i^{(t)})^T}{\sum_n \tau_n^{i(t)}}$$

$$\pi_i^{(t+1)} = \frac{\sum_n \tau_n^t}{N}$$

Divide data to each group,
Compute mean and variance
from each group

GMM for Clustering

- Given the estimated GMM model, compute



$$posteriers = \tau^i = p(z_i = 1 | x, \theta), \quad \theta = \{\mu_0, \Sigma_0, \mu_1, \Sigma_1\}$$

- Estimate the probability that *x* is generated by cluster *i*

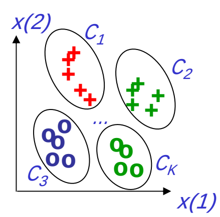
$$Expectation: \tau_n^{i(t)} = \frac{\pi_i^{(t)} N(x_n | \mu_i^{(t)}, \Sigma_i^{(t)})}{\sum_j \pi_i^{(t)} N(x_n | \mu_j^{(t)}, \Sigma_j^{(t)})}$$

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

Comparison: K-Mean Clustering

- Training data

$$\{x_i\} + \{label(i)\ ?\}$$



- 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

$$x_1, x_2, ..., x_N \ \ samples$$
$$for\ i=1,2,...,N,$$
$$x_i \rightarrow C_k, if\ \ Dist(x_i, C_k) < Dist(x_i, C_{k'}), k \neq k'$$
$$end$$