EE 6885 Statistical Pattern Recognition

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

Lecture 6 (9/28/05)

- Reading
  - EM for Missing Features
    - Textbook, DHS 3.9
  - Bayesian Parameter Estimation and Sufficient Statistics
    - Textbook, DHS 3.4-3.6
  - Application: Bayesian Face Detection
    - Reference paper (see course web site)

Parameter Estimation: ML and EM

- Given Data D

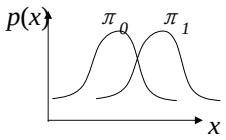
  Find $\hat{\theta} = \arg\max_{\theta} p(D \mid \theta)$

- Gaussian

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

- Mixture of Gaussian

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



- Derive Auxiliary Function

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

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

EM for Missing Feature

$D = \{D_g, D_b\}$ D_g : good feature , D_b : missing feature

$Q(\theta; \theta^i) = E_{D_b}[\ln p(D_g, D_b; \theta) \mid D_g; \theta^i]$ Marginalize over the missing feature

- Example, 2-D Gaussian

  $D = \{x_1, x_2, x_3, x_4\} = \{\begin{pmatrix}0\\2\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}*\\4\end{pmatrix}\}$ $\theta = \begin{bmatrix}\mu_1\\ \mu_2\\ \sigma_1^2\\ \sigma_2^2\end{bmatrix}$ $\theta^0 = \begin{bmatrix}0\\0\\1\\1\end{bmatrix}$

  $$Q(\theta; \theta^0) = E_{x_{41}}[\ln p(x_g, x_{41}; \theta) \mid x_g; \theta^0]$$

  $$= \int_{-\infty}^{\infty}\left[\sum_{k=1}^{3}[\ln p(x_k \mid \theta)] + \ln p(x_4 \mid \theta)\right] p(x_{41} \mid \theta^0; x_{42} = 4) dx_{41}$$

  $$= \sum_{k=1}^{3}[\ln p(x_k \mid \theta)] + \int_{-\infty}^{\infty} \ln p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix} \mid \theta\right) \frac{p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix} \mid \theta^0\right)}{\int_{-\infty}^{\infty} p\left(\begin{pmatrix}x_{41}'\\4\end{pmatrix} \mid \theta^0\right) dx_{41}'} dx_{41}$$

  $\theta^1 = \begin{bmatrix}0.75\\2.0\\0.938\\2.0\end{bmatrix}$ $\theta = \begin{bmatrix}1.0\\2.0\\0.5\\2.0\end{bmatrix}$

- E step and M step?

See figure in Sec. 3.9

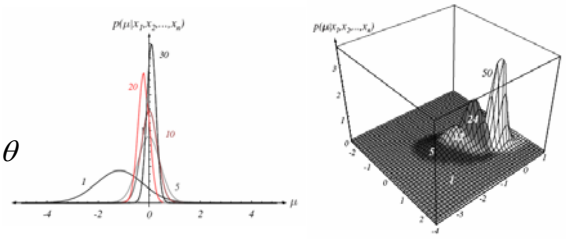
Bayesian Parameter Estimation

- Instead of fixed unknown, we assume θ as random variable with distribution $p(\theta)$
- Given samples D, find the maximal posterior

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int p(D \mid \theta)p(\theta)d\theta} \qquad p(D \mid \theta) = \sum_{k=1}^{n} p(x_k \mid \theta)$$

- Posterior gives the probability distribution of θ .
  When the sample size increases, the distribution becomes peaky.
- Difference from ML estimation?

$$p(x \mid D) = \int p(x \mid \theta)p(\theta \mid D)d\theta$$



EE6887-Chang

6-5

Example: Gaussian

- Univariate case: μ is the only unknowr

$$p(x \mid \mu) \sim N(\mu, \sigma^2) \qquad p(\mu) \sim N(\mu_0, \sigma_0^2)$$

$$\text{posterior} \quad p(\mu \mid D) \propto p(D \mid \mu)p(\mu)$$

$$= \prod_{k=1}^{n} p(x_k \mid \mu)p(\mu)$$

$$= \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right]$$

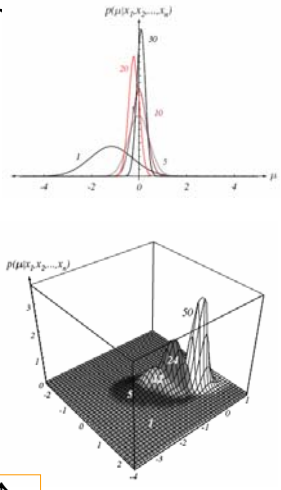
$$= \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right] \sim N(\mu_n, \sigma_n^2)$$

Reproducing density

$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \cdot \mu_0 \quad \text{where} \quad \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_n$$

$$\text{and} \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

what if $n \uparrow, \sigma_0 \uparrow, \sigma \uparrow$?



EE6887-Chang

6-6

Example: Gaussian Case (Cont.)

- Class conditional density

$$p(x \mid D) = \int p(x \mid \mu)\, p(\mu \mid D) d\theta$$

$$p(x \mid D) \sim N(\mu_n, \sigma^2 + \sigma_n^{\,2}) \quad \text{vs.} \quad p(x \mid \mu) \sim N(\mu, \sigma^2)$$

- Mean is replaced, variance is increased.
- Bayesian Classifier: compute posterior $p(x \mid D_i, \omega_i)\, p(\omega_i)$
- Multi-variate case

posterior $p(\mu \mid D) \sim N(\mu_n, \Sigma_n)$

$$\mu_n = \Sigma_0 (\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\, \hat{\mu}_n + \frac{1}{n}\Sigma (\Sigma_0 + \frac{1}{n}\Sigma)^{-1} \mu_0 \quad \text{where} \quad \hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_n$$

$$\Sigma_n = \Sigma_0 (\Sigma_0 + \frac{1}{n}\Sigma)^{-1} \frac{1}{n}\Sigma$$

- Class conditional density $p(x \mid D) \sim N(\mu_n, \Sigma + \Sigma_n)$

Sufficient Statistics

- sufficient statistics **s**: function of samples that contains all of the information relevant to estimation of parameters θ
- Example: Gaussian $N(\theta, \Sigma)$ θ is unknown

$$P(D \mid \theta) = \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp(\frac{-1}{2}(\mathbf{x}_k - \theta)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \theta))$$

$$= \frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp(\frac{-1}{2}\sum_{k=1}^{n}(\theta^t\boldsymbol{\Sigma}^{-1}\theta - 2\theta^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k + \mathbf{x}^t{}_k\boldsymbol{\Sigma}^{-1}\mathbf{x}_k)$$

$$= \underbrace{\exp\left[\frac{-n}{2}\theta^t\boldsymbol{\Sigma}^{-1}\theta + \theta^t\boldsymbol{\Sigma}^{-1}(\sum_{k=1}^{n}\mathbf{x}_k)\right]}_{g(\hat{\mu}_n, \theta) = \exp\left[\frac{-n}{2}(\theta^t\boldsymbol{\Sigma}^{-1}\theta - 2\theta^t\boldsymbol{\Sigma}^{-1}\hat{\mu}_n)\right]} \times \underbrace{\frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp(\frac{-1}{2}\sum_{k=1}^{n}(\mathbf{x}^t{}_k\boldsymbol{\Sigma}^{-1}\mathbf{x}_k)}_{h(D)\ \text{independent of}\ \theta}$$

where $\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$ Sufficient statistics

Sufficient Statistics for Exponential Family

- Gaussian, exponential, Rayleigh, Poisson, etc.

$$p(x|\theta) = \alpha(x)\exp[a(\theta) + b(\theta)^t c(x)]$$

$$p(D|\theta) = \prod_{k=1}^{n} \alpha(x_k)\exp[na(\theta) + b(\theta)^t \sum_{k=1}^{n} c(x_k)] = g(s,\theta)h(D)$$

$$where \quad s = \frac{1}{n}\sum_{k=1}^{n} c(x_k)$$

$$g(s,\theta) = \exp[n(a(\theta) + b(\theta)^t s)]$$

$$h(D) = \prod_{k=1}^{n} \alpha(x_k)$$

- See Table 3.1 for the S.S. for different distributions.

MAP detectors for face images

- H. Schneiderman & T. Kanade, "Probabilistic modeling of local appearance and spatial relationships for object recognition," CVPR '98.

$$\frac{P(region|\text{object})}{P(region|\overline{\text{object}})} \underset{\overline{\text{object}}}{\overset{\text{object}}{>\atop <}} \lambda = \frac{P(\overline{\text{object}})}{P(\text{object})}$$

- Capture face information from 64x64 regions. Enough details for characterizing faces. Decompose into 16x16 subregions to simplify models and capture spatial locations.

$$P(region|\text{object}) = P(\{subregion\}|\text{object})$$

$$P(\{subregion\}|\text{object}) \approx \prod_{k=1}^{n_{subs}} P(subregion_k|\text{object})$$

- Does not consider subregion dependency. Less penalty for face detection due to consistent appearance and location of salient features, e.g., noises and eyes.
- Also capture the distinctive local appearances in the likelihood ratio

$$\prod_{k=1}^{n_{subs}} \frac{P(subregion_k|\text{object})}{P(subregion_k|\overline{\text{object}})}$$

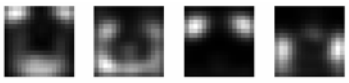
Bayesian face detector (Cont.)

- Simplify the subregion patterns by PCA projection, sparse coding, and quantization → 3.8M distinctive patterns of q_1

$$P(pattern, pos|\text{object}) \approx P(q1, pos|\text{object})$$

- Estimate the location distributions from training data. $P(pos|q1, \text{object})$
- Reduce the complexity of the above distributions by pattern clustering q_2=VQ(q_1)

$$P(pos'|q1, \text{object}) \approx P(pos'|q2, \text{object})$$



- Clustering helps discover salient locations
- Simple model for background and infrequent patterns

$$p(pos|q1, \overline{\text{object}}) \approx \frac{1}{n_{subs}}$$

- Multi-resolution scaling to capture the face features at different scales

$$P(region|\text{object}) \approx \prod_{j=1}^{n_{magn}} P(region^j|\text{object})$$

Bayesian face detector (Cont. 2)

- Final Bayesian Estimation

$$\prod_{j=1}^{n_{magn}} \prod_{i=1}^{n_{subs}} \frac{P(q1_i^j|\text{object})P(pos_i^{j'}|q2_i, \text{object})}{\frac{P(q1_i^j|\overline{\text{object}})}{n_{subs}}} \underset{\overline{\text{object}}}{\overset{\text{object}}{>\!<}} \lambda = \frac{P(\overline{\text{object}})}{P(\text{object})}$$

- Estimate class specific distribution using training frontal face and non-face images → simple frequency count
- For test images, search faces at multiple scales, e.g., 17 scales over 18x18 to 338x338.

Schneiderman & Kanade(125 images)		Rowley, Baluja, and Kanade [5] (130 images)	
Detection rate	False alarms	Detection rate	False alarms
93.0%	88	92.5%	862
90.5%	33	86.6%	79
77.0%	1	77.9%	2

- Key features:
  - joint local appearance-location statistical modeling
  - background modeling
  - discrete non-parameter distribution modeling