EE 6885 Statistical Pattern Recognition

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

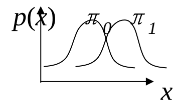
Lecture 7 (10/3/05)

- Reading
  - Problem with Dimensionality
    - Bellman, R.E. 1961. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.
    - G.V. Trunk, "A Problem of Dimensionality: a Simple Example," IEEE Trans-PAMI, July 1979.
  - Nonparametric Estimation
    - DHS Chap. 4.1-4.3
- Homework #3, due Oct. 12th 2005
- Midterm Exam
  - Oct. 24th 2005 Monday

Parameter Estimation

- ML Estimator, Given Data D Find $\hat{\theta} = \arg\max_{\theta} p(D \mid \theta)$
- Gaussian $\Rightarrow$ $\hat{\mu} = (1/n)\sum_k \vec{x}_k$ $\hat{\Sigma} = (1/n)\sum_k (\vec{x}_k - \mu)(\vec{x}_k - \mu)^t$
- Mixture of Gaussian

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



- EM for missing features

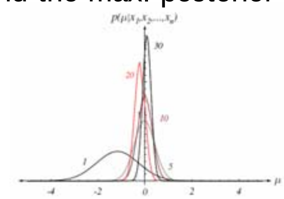
$$Q(\theta; \theta^i) = E_{D_b}[\ln p(D_g, D_b; \theta) \mid D_g; \theta^i]$$

Marginalize over the missing feature

- Bayesian Estimation: Treat θ as R.V., find the max. posterior

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

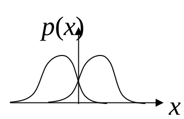
$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \cdot \mu_0$$



- Application in Face Detection:
  joint spatio-appearance features, likelihood ratio, discretization

Problem with High Dimensionality

- A Simple Example (Turk 1978)



$$p(x \mid \omega_1) = N(\mu_1, I)$$

$$p(x \mid \omega_2) = N(\mu_2, I)$$

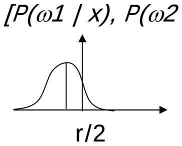
where $\mu_1 = -\mu_2 = \mu = \{(1/i)^{1/2}, i = 1 \ldots n\}$

MAP classifier

assume equal prior $P(\omega_1) = P(\omega_2) = 1/2$ $\Rightarrow$ decide ω_1 if $z = x^t\mu > 0$

- Prob. Of Error *P(error | x) = min [P(ω1 | x), P(ω2 | x)]*

$$P_e = \int_{r/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$



$$r^2 = \|\mu_1 - \mu_2\|^2 = 4\sum_{i=1}^{n}(1/i) \to \infty \quad \text{when } n \to \infty$$

$$\therefore\ P_e \to 0 \quad \text{when } n \to \infty$$

- If true parameters are known, high dimensionality helps

Problem with Finite Sample Estimation

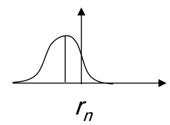
- If true parameters are unknown, need to estimate from data samples $x_1, x_2, \ldots, x_m$

$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x_i$ $\qquad -x_i$ is used if sample comes from ω_2

- Prob. of error $P_e = \Pr(z = x^t\hat{\mu} > 0 \mid \omega_2)$

$$E(z \mid \omega_2) = E\left(x^t\left(\frac{-x_1 - x_2 - \ldots - x_m}{m}\right)\right) = -\sum_{i=1}^{n}(1/i)$$

$$\operatorname{var}(z) = \left(1 + \frac{1}{m}\right)\sum_{i=1}^{n}(1/i) + n/m$$



$\Rightarrow (z - E(z))/(Var(z))^{1/2}$ becomes a normal dist. when $n \to \infty$

$$P_e = \int_{\gamma_n}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

where $\gamma_n = [\sum_{i=1}^{n}(1/i)]/\operatorname{var}(z)$

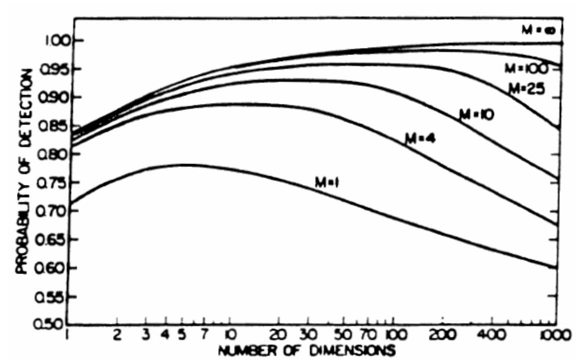
$\to 0$ when $n \to \infty$ and m finite

$\therefore \lim_{n\to\infty} P_e = 0.5$

Curve of dimensionality (R.E. Bellman '61): convergence of any estimator to the true value of a smooth function defined on a space of high dimension is very slow.

Problem of High Dimensionality (Cont.)

- Prob. of error → 0.5 when $n \to \infty$ and m finite



- Compare with random guess?
- If for 1-D unit interval, we need n_1 samples to estimate distribution, then we need n_1^K for the K-D unit hypercube

Property of High-Dimensional Space

- If we want to estimate pdf $p(x)$ over the hypercube R^d in d-dimensional space with n samples
- Interpoint distances are all large and roughly equal

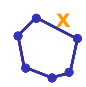
volume of hyper-rectangle containing a point and its nearest point

$$\Delta_1 \Delta_2 \ldots \Delta_d = \delta$$

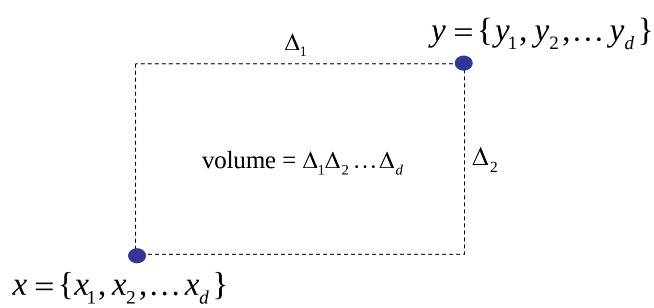
note $0 \le \Delta_i \le 1$ and most likely some Δ_i are large

Therefore, $L_2 = \left[\sum_{i=1}^{n} (\Delta_i)^2 \right]^{1/2}$ will be large for any pair of points

- Similarly, every point is close to at least one face of the hypercube. Why?
- Most samples are on the convex hull of the training set, i.e., most points can be considered as outliers for the rest.

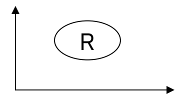


- Predicting a new point: extrapolation or interpolation?



Nonparametric Techniques

- Assumptions about the underlying distributions may be incorrect.
- General approach: estimate the density directly.



$p(x) \simeq \frac{k/n}{V}$, where k: # points falling in R, V: volume of R

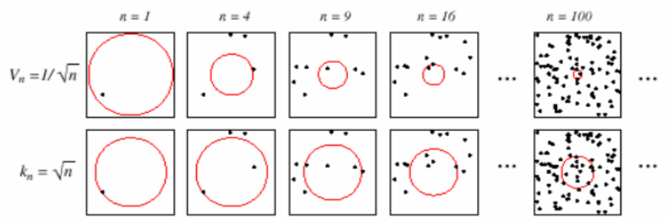
form a sequence of R_n: $p_n(x) \simeq \frac{k_n/n}{V_n}$

For $p_n(x) \to p(x)$, required conditions: $\lim_{n\to\infty} V_n = 0$; $\lim_{n\to\infty} k_n = \infty$; $\lim_{n\to\infty} k_n/n = 0$

- Two approaches:

1: control and shrink the volumn V_n, e.g., $1/\sqrt{n}$ $\to$ Parzen window

2: control k_n, e.g., $\sqrt{n}$ $\to$ k_n nearest-neighbor method



EE6887-Chang

7-9