Reading

Problem with Dimensionality

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 $\hat{\theta} = \arg \max p(D | \theta)$
- Gaussian $\Rightarrow \hat{\mu} = (1/n)\sum \bar{x}_k$ $\hat{\Sigma} = (1/n)\sum (\bar{x}_k - \mu)(\bar{x}_k - \mu)'$
- Mixture of Gaussian

$$I = \sum_{i=1}^{N} \log(p(x|\mu_i,\Sigma_i) + \pi_i p(x|\mu,\Sigma))$$

- EM for missing features

$Q(\theta; \theta') = E_{D_h}[\ln p(D_h|D_b; \theta)|D_h; \theta']$ Marginalize over the missing feature

- Bayesian Estimation: Treat $\theta$ as R.V., find the max. posterior $p(x|\mu) \sim N(\mu,\sigma^2)$ $p(\mu) \sim N(\mu_0,\sigma_0^2)$

$$\mu_n = \frac{n_0 \sigma_0^2 + \sigma^2}{n_0 \sigma_0^2 + \sigma^2} \mu_0 + \frac{n \sigma^2}{n_0 \sigma_0^2 + \sigma^2} \hat{\mu}$$

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

Problem with High Dimensionality

- A Simple Example (Turk 1978)

\[ p(x | \omega_i) = N(\mu_i, I) \]
\[ p(x | \omega_2) = N(\mu_2, I) \]

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

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

- Prob. Of Error

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

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

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

- If true parameters are known, high dimensionality helps
Problem with Finite Sample Estimation

- If true parameters are unknown, need to estimate from data samples $x_1, x_2, \ldots, x_m$
  $$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} x_i$$  
  $\hat{x}_i$ is used if sample comes from $\omega_2$

- Prob. of error $P_e = Pr(z = x' \hat{\mu} > 0 | \omega_2)$
  $$E(z | \omega_2) = E(x'(-x_1 - x_2 - \ldots - x_m)) = -\sum_{i=1}^{m}(1/i)$$

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

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

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

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

  $$\Rightarrow 0 \text{ when } n \to \infty \text{ and } m \text{ finite}$$

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

  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 $\to 0.5$ 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 \( 0 \leq \Delta_i \leq 1 \) and most likely some \( \Delta_i \) are large.

  Therefore, \( L_2 = \left[ \sum_{i=1}^{d} (\Delta_i)^2 \right]^{1/2} \) 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?

\[ x = \{ x_1, x_2, \ldots, x_d \} \]
\[ y = \{ y_1, y_2, \ldots, y_d \} \]

\[ \Delta_1 \Delta_2 \ldots \Delta_d \]

\[ \text{volume} = \Delta_1 \Delta_2 \ldots \Delta_d \]
Nonparametric Techniques

- Assumptions about the underlying distributions may be incorrect.
- General approach: estimate the density directly.
  \[ p(x) = \frac{k}{n/V} \] where \( k \): # points falling in \( R \), \( V \): volume of \( R \)
  form a sequence of \( R_n \): \[ p_n(x) = \frac{k_n}{V_n} \]
  For \( p_n(x) \to p(x) \), required conditions:
  \[ \lim_{n \to \infty} V_n = 0; \quad \lim_{n \to \infty} k_n = \infty; \quad \lim_{n \to \infty} k_n/n = 0 \]
- Two approaches:
  1: control and shrink the volume \( V_n \), e.g., \( 1/\sqrt{n} \) \to Parzen window
  2: control \( k_n \), e.g., \( \sqrt{n} \) \to \( k_n \) nearest-neighbor method