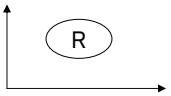
EE 6885 Statistical Pattern Recognition

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

Lecture 9 (10/10/05)

- Reading
  - Nonparametric Estimation
    - DHS Chap. 4.1-4.3
  - Nearest Neighbor Estimation, Distance Metrics
    - DHS Chap. 4.4-4.5, 4.6
  - Paper: Bayesian Classifier with VQ and Parzen Window
    - A. Vailaya, M. Figueiredo, A. Jain, and HJ Zhang, "A Bayesian Framework for Semantic Classification of Outdoor Vacation Images," IEEE Trans. Image Processing, Vol. 10, No. 1, pp. 157-172, Jan. 2001.
- Homework #3, due Oct. 12th 2005
- Midterm Exam
  - Oct. 24th 2005 Monday 1pm-2:30pm (90mins)
    - Open books/notes, no computer

Review: Nonparametric Techniques



- General approach: estimate the density directly.

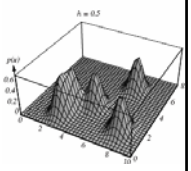
  form local region sequence R_n : $p_n(x) \simeq \dfrac{k_n / n}{V_n}$

  $p_n(x) \to p(x)$ $\lim_{n\to\infty} V_n = 0;$ $\lim_{n\to\infty} k_n / n = 0$ $\lim_{n\to\infty} k_n = \infty$

- Parzen Window

  $$\Rightarrow p_n(x) = \frac{1}{n}\sum_{i=1}^{i=n} \frac{1}{V_n}\,\varphi\left(\frac{x - x_i}{h_n}\right)$$

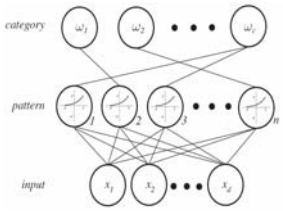
  Superposition of n estimate functions



  $$E[p_n(x)] = \int \delta_n(x - v)\, p(v)\, dv \qquad \sigma_n^{\,2}(x) \le \frac{\sup(\varphi(.))\,\overline{p}_n(x)}{nV_n}$$

- Parallel implementation

  $$\varphi\left(\frac{x - x_k}{h_n}\right) \propto \exp(-(x - x_k)^t (x - x_k)/2\sigma^2)$$
  $$= \exp(x^t x_k - 1)/\sigma^2)$$

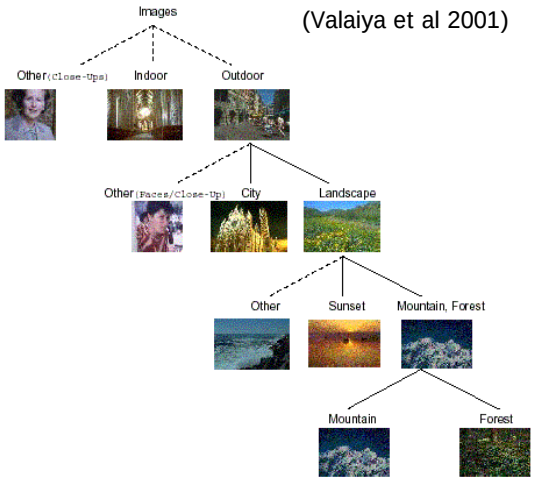


- Complexity

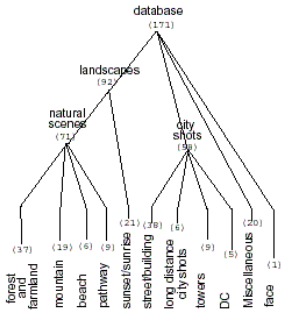
  space: $O((n+1)d)$, time: $O(nd)$

Application: Bayesian Image Classification

Exclusive classes –
no rejections nor overlap



(Valaiya et al 2001)



Bayesian Image Classifiers with Parzen Window

- Features: color/edge histogram, coherence histogram

Feature independence $f_{\mathbf{X}}(\boldsymbol{x} \mid \omega) \equiv f_{\mathbf{Y}}(\boldsymbol{y} \mid \omega) = \prod_{i=1}^{M} f_{\mathbf{Y}^{(i)}}(\boldsymbol{y}^{(i)} \mid \omega).$

MAP Classification $\hat{\omega} = \delta(\boldsymbol{x}) = \arg\max_{\omega \in \Omega} \{p(\omega \mid \boldsymbol{y})\} = \arg\max_{\omega \in \Omega} \{f_{\mathbf{Y}}(\boldsymbol{y} \mid \omega)\, p(\omega)\}.$

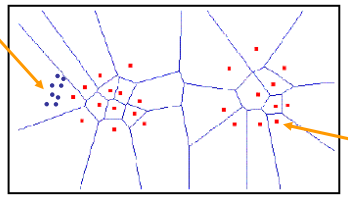
- Probability density estimation through Vector Quantization

$$VQ: \; y \to \hat{y}_i \text{ if } D(y, \hat{y}_i) \le D(y, \hat{y}_j), \forall j \ne i, \; j = 1, \ldots, q$$

$\hat{y}_i$: codebook vectors

- Design codebook by distortion minimization (Euclidean or Mahala. Dist.)

Multiple samples in a cell



Voronoi cells and points from VQ

codeword

VQ density estimation & MDL

Parzen Window: Approximate density with proportion of sample data in each cell

$$f_{\mathbf{Y}^{(i)}}(\boldsymbol{y}^{(i)} \mid \omega) \approx \frac{m_j^{(i)}}{Vol(S_j^{(i)})},$$ Piecewise constant

$$f_{\mathbf{Y}^{(i)}}(\boldsymbol{y}^{(i)} \mid \omega) \propto \sum_{j=1}^{q} m_j^{(i)} * \exp(-\|\boldsymbol{y}^{(i)} - \boldsymbol{v}_j^{(i)}\|^2/2).$$ GMM

- Difference from the standard Parzen Window?
- What happens if codebook size q increases?
  - Likelihood increases; Model size increases (overfitting)
- Consider the total data length for describing the data and model
  - Minimal Description Length (MDL)

MDL optimization principle $\hat{q} = \arg\min \{L(\mathcal{Y}|\boldsymbol{\theta}_{(q)}) + L(\boldsymbol{\theta}_{(q)})\}$

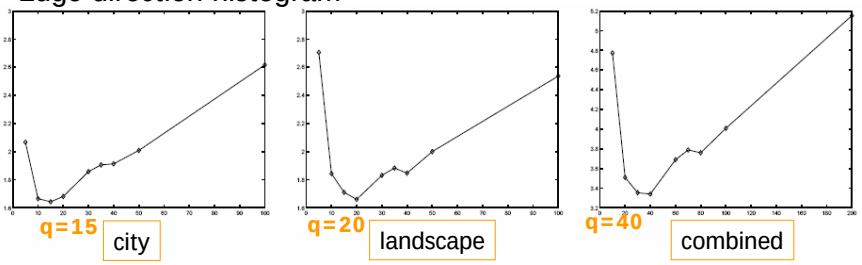
Optimal data length given model $L(\mathcal{Y}|\boldsymbol{\theta}_{(q)}) = -\sum_{j=1}^{n} \log f(\boldsymbol{y}^{(i)}(j)|\omega, \boldsymbol{\theta}_{(q)})$

Model description length

$$\begin{cases} L(\boldsymbol{\theta}_{(q)}) = (\zeta(q)/2) \log n \\ \zeta(q) = \{q + q \dim(\boldsymbol{y}^{(i)})\} \end{cases}$$

Experiment: Optimal model size

- Edge direction histogram

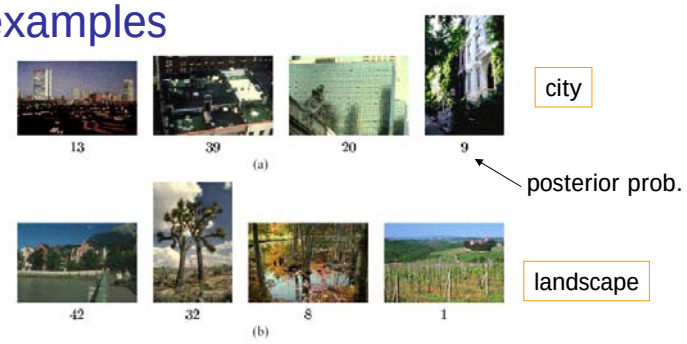


- Why the optimal model size increases when combining data sets?
- Edge information important for detecting “city” images

Test Data	EDH	EDCV	CH	CCV	EDH & CH	EDH & CCV	EDCV & CH	EDCV & CCV
Training Set	94.7	96.7	83.7	83.5	94.8	95.4	96.4	96.9
Test Set	92.0	92.7	75.4	76.0	92.5	92.8	93.4	93.8
Entire Database	93.4	94.7	79.6	79.8	93.7	94.1	94.9	95.3

- Color information important for discriminating landscape subclasses

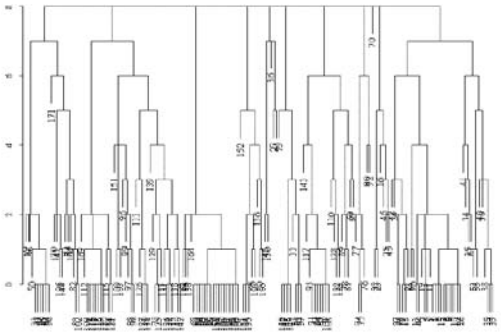
Failure examples

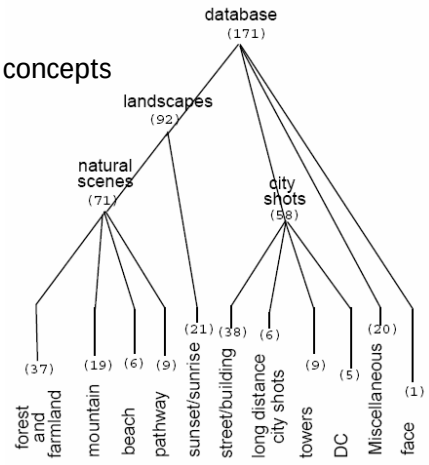


- Key features:
  - Reduce the number of estimate functions by VQ
  - Reduce memory size and computational cost
  - Estimate local density by GMM Parzen Window
- Possible improvements:
  - Did not use cross validation to assess performance and choose q
  - Add classes; non-hierarchical classes
  - Multiple binary classifiers, e.g., “city” vs. “no-city”

Interesting Issue:
Discovering image classes through sorting task (Valaiya et al 98)

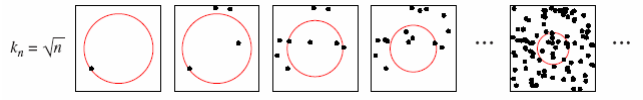
- What categories to classify? How to organize?
- Human subjects to sort images to groups (unconstrained)
- Define pair-wise distance matrix
- Perform hierarchical clustering
- Cut at different levels to obtain salient concepts





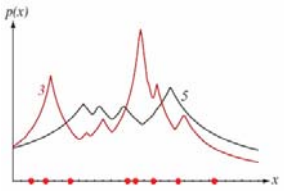
k_n-Nearest-Neighbor

$$p_n(x) \simeq \frac{k_n / n}{V_n}$$



- Necessary and sufficient conditions for $p_n(x) \to p(x)$

$$\lim_{n\to\infty} k_n / n = 0 \qquad \lim_{n\to\infty} k_n = \infty$$



- Example:
  - Peak of $p_n(x)$ often are away from sample points.

- For classification, estimate $p(x)$ for each class ω_i

$$p_n(x, \omega_i) = \frac{k_i / n}{V} \qquad p_n(\omega_i \mid x) = \frac{p_n(x, \omega_i)}{\sum_{j=1}^{c} p_n(x, \omega_j)} = \frac{k_i}{k}$$

- Use labels of neighbors to est. posteriors

Error Rate of Nearest Neighbor Classifier

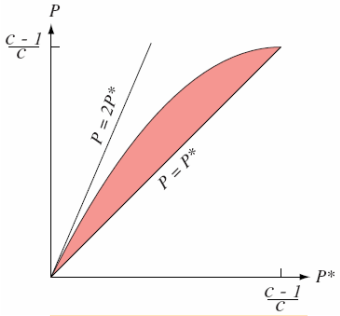
- When k=1, nearest neighbor

$$P^* \le \lim_{n\to\infty} P_n(e) \le P^*\left(2 - \frac{c}{c-1}P^*\right)$$

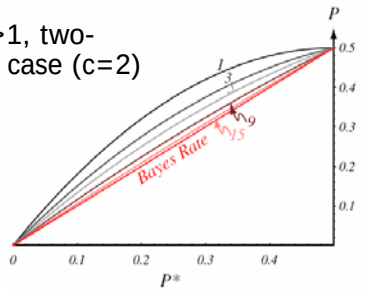
where c: # of classes, P^*: Bayesian Error Prob.

$$P^* = \int P^*(e \mid x)\, p(x)\, dx$$

$$P^*(e \mid x) = 1 - \max_i P(\omega_i \mid x)$$



- When k>1, two-category case (c=2)



Compared to random guess?

Deriving the error bound ...

Assume n samples: $(x_1, \theta_1), (x_2, \theta_2), \ldots, (x_n, \theta_n)$

Assume x'_n is the nearest neighbor to x **Assume i.i.d.**

$$P_n(e \mid x, x'_n) = 1 - \sum_{i=1}^{c} P(\theta = \omega_i, \theta'_n = \omega_i \mid x, x'_n) = 1 - \sum_{i=1}^{c} P(\omega_i \mid x) P(\omega_i \mid x'_n)$$

assume $p(x'_n)$ peaks at x

$$\lim_{n\to\infty} P_n(e \mid x) = \lim \int P_n(e \mid x, x'_n) p(x'_n) dx'_n = \lim \int P_n(e \mid x, x'_n) \delta(x'_n - x) dx'_n$$

$$= \int \left[1 - \sum_{i=1}^{c} P(\omega_i \mid x) P(\omega_i \mid x'_n)\right] \delta(x'_n - x) dx'_n = 1 - \sum_{i=1}^{c} P^2(\omega_i \mid x)$$

$$P = \lim_{n\to\infty} P_n(e) = \lim_{n\to\infty} \int P_n(e \mid x) p(x) dx = \int [1 - \sum_{i=1}^{c} P^2(\omega_i \mid x)] p(x) dx$$

maximized when $P(\omega_i \mid x)$ are equal except the largest

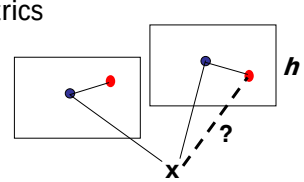
$$\Rightarrow \quad P^* \le \lim_{n\to\infty} P_n(e) \le P^*\left(2 - \frac{c}{c-1}P^*\right)$$

Distance Metrics

- Nearest neighbor rules need distance metrics
- Required properties of a metric

  1. non-negativity: $D(a,b) \geq 0$
  2. reflexivity: $D(a,b) = 0$ iff $a = b$
  3. symmetry: $D(a,b) = D(b,a)$
  4. trangular inequality: $D(a,b) + D(b,c) \geq D(c,a)$

     $D(a,b) \geq D(c,a) - D(b,c)$



useful in indexing

- Minkowski Metric
  - Euclidean
  - Manhattan
  - L_∞

$$L_k(a,b) = \left(\sum_{i=1}^{d} | a_i - b_i |^k\right)^{1/k}$$

- Tanimono Metric
  - sets of elements
  - Point-point distance not useful

$$D_{\text{tanimono}}(S_1, S_2) = \frac{n_1 + n_2 - 2n_{12}}{n_1 + n_2 - n_{12}} = \frac{(n_1 - n_{12}) + (n_2 - n_{12})}{n_1 + n_2 - n_{12}}$$