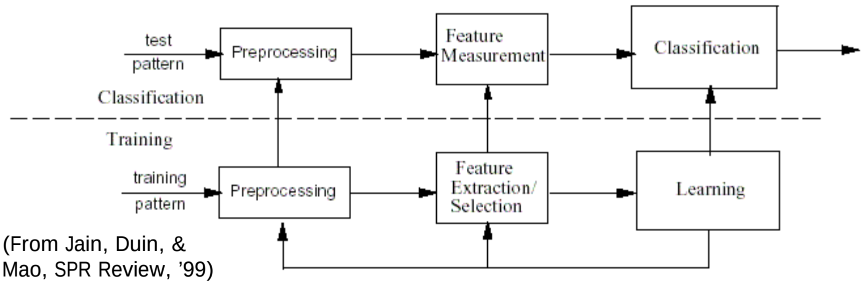
EE 6885 Statistical Pattern Recognition

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

Lecture 2 (9/12/05)

- Reading
  - Bayesian Decision Theory
    - Textbook DHS Chapter 2.1 – 2.5
  - Paper about image features:
    - Y Rui, TS Huang, SF Chang, “Image retrieval: Current techniques, promising directions and open issues,” - Journal of Visual Communication and Image Representation, 1999
    - Other papers (such as MPEG-7 features) on the course web site
- Homework #1 due 2005-09-21
- Basic Matlab sample code (course web site)

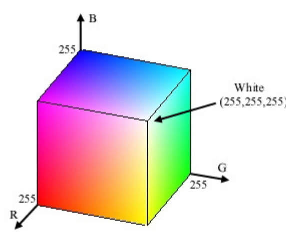
A Very High-Level Stat. Pattern Recog. Architecture



(From Jain, Duin, & Mao, SPR Review, '99)

Example image features: color histogram

- Color Spaces
  - RGB – cube in Euclidean space



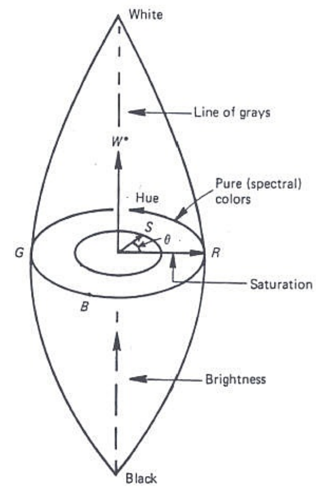
$$r = \frac{R}{R+G+B} \quad g = \frac{G}{R+G+B} \quad b = \frac{B}{R+G+B}$$

$$h_{RGB}[r,g,b] = \sum_m \sum_n \begin{cases} 1 & \text{if } I_R[m,n] = r, I_G[m,n] = g, I_B[m,n] = b \\ 0 & \text{otherwise} \end{cases}$$

A color histogram represents the distribution of colors where each histogram bin corresponds to a color in the quantized color space

Choose the right color space

- HSI space is closer to human perception than RGB
  - hue (color tone): the circumference
  - saturation : the radius
  - Brightness (intensity) : the vertical axis
- HSI color histogram
  - quantize the HSI space and count the number of pixels in each bin



EE6887-Chang

2-5

Color Histogram (cont.)

- Advantages of color histograms
  - Compact representation of color information
    - Issue of dimension, 64-D, 166-D, 1024-D etc
  - Global color distribution
    - Does not require segmentation
    - Issue of invariance
  - Histogram distance metrics
    - Issue of correspondence to human perception

EE6887-Chang

2-6

Histogram Distance Metrics

- L1 distance $$D_1(i,i+1) = \sum_j \left| H_i(j) - H_{i+1}(j) \right|$$
- L2 distance $$D_2(i,i+1) = \sum_j \left| H_i(j) - H_{i+1}(j) \right|^2$$
- Histogram Intersection $$D_I = 1 - \frac{\sum_j \min\left(H_i(j), H_{i+1}(j)\right)}{\min\left(\sum_j H_{i+1}(j), \sum_j H_i(j)\right)}$$
- Quadratic Distance $$D_Q = \sum_{j_1} \sum_{j_2} \left(H_i(j_1) - H_{i+1}(j_1)\right) \alpha(j_1, j_2) \left(H_i(j_2) - H_{i+1}(j_2)\right)$$
  $\alpha(j_1, j_2)$: *correlation between colors* j_1, j_2. *e.g. 1-*d_{j_1, j_2}
- Other histogram features
  - Edge histogram
  - Issue: affected by quality of edge extraction, lighting, noise etc

EE6887-Chang

2-7

Bayesian Decision Theory

- Class-dependent feature probability distribution $P(x/\omega_j)$
  - *E.g.,* ω_1*: indoor,* ω_2*: outdoor*
- Given x, which class ω_i is more likely?
- Use probabilities
- Posterior = (Likelihood x Prior) / Evidence

  $P(\omega_j \mid x) = P(x \mid \omega_j) \cdot P(\omega_j) / P(x)$
- In case of two categories $$P(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j) P(\omega_j)$$

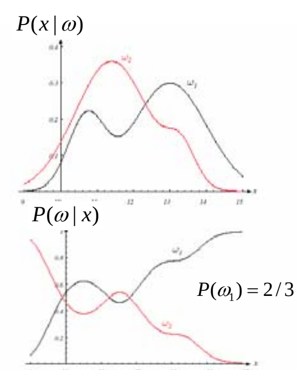


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

EE6887-Chang

2-8

Review: Probability

- A, B are events
  - E.g., sequence of coin tossing outcome
- Independent

$$P(A \cap B) = P(A)P(B)$$

$$A_i, i = 1,2,\ldots,N, \textit{are independent} \Leftrightarrow P(\bigcap_{j=1}^{N} A_j) = \prod_{j=1}^{N} P(A_j)$$

- Conditional Probability of A given B

$$P(A\,|\,B) = \frac{P(A,B)}{P(B)} \quad \text{Bayes Theorem: } P(A\,|\,B) = \frac{P(B\,|\,A)P(A)}{P(B)}$$

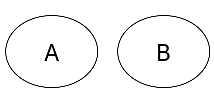
$$\text{Total prob. theorm: } P(B) = P\left(\bigcup_{i=1}^{\infty}(B \cap A_i)\right) = \sum_{i=1}^{\infty} P(B\,|\,A_i)P(A_i),$$

$$\textit{if } A_i \text{ are disjoint, i.e., } A_i \cap A_j = \Phi, \forall i, j, j \neq j$$

Probability

- Independence & mutual exclusion (uncorrelated) are different

$$Cov(A,B) = 0 \Rightarrow \textit{uncorrelated}$$



$$P(A \cap B) = P(A)P(B) \Rightarrow \textit{independent}$$

$$\textit{independent} \Rightarrow \textit{uncorrelated}; \text{ but converse not true}$$

- Probability of continuous random variable
  - Cumulative distribution function (cdf) $F_X(x) = Prob\{X \in (-\infty, x]\}$
  - Probability density function (pdf) $p_X(x) = \text{Prob}\{X = x\}$

$$= \left.\frac{dF_X(x')}{dx'}\right|_{x'=x}$$

$x_1, \ldots, x_N$ are independent iff

$$\text{P}(x_{i_1} \leq \text{x}_1; \ldots; x_{i_n} \leq \text{x}_\text{n}) = \prod_{k=1}^{n} P(\text{x}_{\text{i}_\text{k}} \leq \text{x}_\text{k}), \text{i}_\text{k}, n \leq N,$$

Probability

- Joint distribution $F_{X,Y}(x,y) = Prob\{X \le x, Y \le y\}$
- Joint density $f_{X,Y}(x,y) = \frac{\partial^2}{\partial x' \partial y'} F_{X,Y}(x',y')\Big|_{x'=x, y'=y}$
- Marginal distribution

$$f_X(x) = \int_R f_{X,Y}(x,y)dy$$

- Conditional probability of x given y

$$f_X(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad \text{note: a function of both x \& y}$$

$$Bayes \text{ theorem}: f_X(x|y) = \frac{f_Y(y|x) f_X(x)}{f_Y(y)}$$

$$\text{Total prob. theorem}: \mathrm{f_X}(x) = \int f_X(x \mid y) f_Y(y)dy$$

Entropy: measure the degree of uncertainty

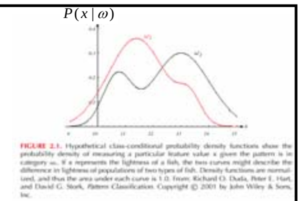
- Entropy (bits) $H = -\sum_{i=1}^{m} P_i \log_2 P_i$ $\quad H = -\int_{-\infty}^{\infty} p(x) \log_2 p(x) dx$
- Given same mean and variance,
  - Which has the max entropy?

$$Gaussian:$$
$$H_{gau} = 0.5 + \log_2(\sqrt{2\pi}\sigma)$$

- For discrete x and arbitrary function f(.)
  - Processing never increases entropy for discrete variables
  - Because prob. cannot be split to two different values after processing

$$H(f(x)) \le H(x)$$

Minimal Decision Risk



- Given x, which class ω_i is more likely? Which action?
  - Formulate as a minimal risk decision problem
- Let $\{\omega_1, \omega_2, \dots, \omega_c\}$ be the set of c states of nature (or "categories", "states")
- Let $\{\alpha_1, \alpha_2, \dots, \alpha_a\}$ be the set of possible actions
  - Example action: "decide class ω_i"
- Let $\lambda(\alpha_i \,/\, \omega_j)$ be the loss incurred for taking action α_i when the true state of nature is ω_j
- Risk for taking action α_i, given x

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

Two-Category case

α_1 : *deciding* ω_1

α_2 : *deciding* ω_2

$\lambda_{ij} = \lambda(\alpha_i \,/\, \omega_j)$

$\lambda(\alpha_i \,/\, \omega_j)$ represents loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

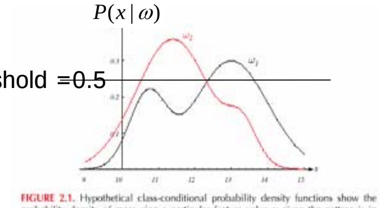
$$R(\alpha_1 \,/\, x) = \lambda_{11} P(\omega_1 \,/\, x) + \lambda_{12} P(\omega_2 \,/\, x)$$
$$R(\alpha_2 \,/\, x) = \lambda_{21} P(\omega_1 \,/\, x) + \lambda_{22} P(\omega_2 \,/\, x)$$

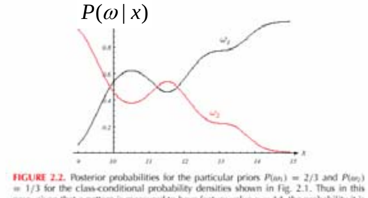
Minimal Risk Decision Rule

- Decide ω_1, if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$
- *i.e.* Decide ω_1, if
  $(\lambda_{21} - \lambda_{11}) P(x \mid \omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) P(x \mid \omega_2) P(\omega_2)$
- *i.e.* Decide ω_1, $\quad if \ \dfrac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \dfrac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \dfrac{P(\omega_2)}{P(\omega_1)}$
- If $\lambda_{12} = \lambda_{21} = 1$, $\lambda_{11} = \lambda_{22} = 0$, *i.e., loss = 'classification error'*

$$Decide\ \omega_1,\ if\ \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \frac{P(\omega_2)}{P(\omega_1)},\ \text{i.e. } P(\omega_1 \mid x) > P(\omega_2 \mid x)$$

Max A Posterioi (MAP) classifier





Bayesian Classification Rule (Contd.)

- Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2
  - Bayesian Classification Error

  $P(error \mid x) =$

  $\min\ [P(\omega_1 \mid x), P(\omega_2 \mid x)]$

- If $P(\omega_1) = P(\omega_2)$,

  ML classifier

$$Decide\ \omega_1,\ if\ \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > 1,\ \text{i.e. } P(x \mid \omega_1) > P(x \mid \omega_2)$$

(Likelihood Ratio)

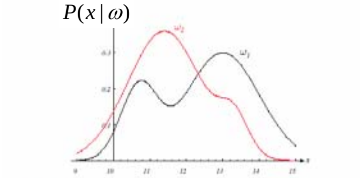


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.