Introduction to Pattern Recognition and Data Mining

Lecture 2: Bayesian Decision Theory

Instructor: Dr. Giovanni Seni

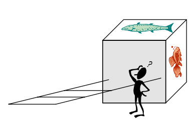
Department of Computer Engineering
Santa Clara University

Overview

- Basic statistical concepts
  - Apriori probability, class-conditional density
  - Bayes formula & decision rule
  - Loss function & minimum-risk classifier
- Discriminant functions
- Decision regions/boundaries
- The Normal density
  - Discriminant functions (LDA)

Introduction
Statistical Approach

- A formalization of common-sense procedures…
- Quantify tradeoffs between various classification decisions using probability
- Initially assume all relevant probability values are known
- State of nature
  - What fish type (ω) will come out next?
    - ω_1 = *salmon*, ω_2 = *sea bass*
  - ω is unpredictable – i.e., a random variable

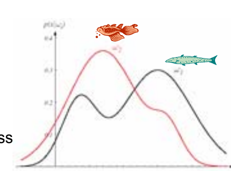


- A priori probability -- prior knowledge of how likely each fish type is -- $P(\omega_1) + P(\omega_2) = 1$

Introduction
Statistical Approach (2)

- Best decision rule about next fish type before it actually appears?
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2
  - How well it works?
    - $P(error) = min [P(\omega_1), P(\omega_2)]$
- Incorporating lightness/length info
  - Class-conditional probability density

    $p(x|\omega 1)$ and $p(x|\omega 2)$ describe the difference in lightness between populations of sea bass and salmon



Introduction
Statistical Approach (3)

- *p(x|ω_j)* also called the likelihood of ω_j with respect to x
  - Other things being equal, ω_j for which $p(x|\omega_j)$ is largest is more "likely" to be true class
- Combining prior & likelihood into *posterior* – Bayes formula

$$p(w_j, x) = P(w_j \mid x)p(x) = p(x|w_j)P(w_j)$$

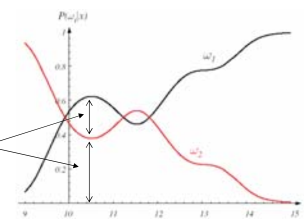
$$P(w_j \mid x) = \frac{p(x|w_j)P(w_j)}{p(x)}$$

where $$p(x) = \sum_j p(x|w_j)P(w_j)$$

Introduction
Statistical Approach (4)

- Bayes Decision Rule
  - Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2

    or
  - Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2



Introduction
Statistical Approach (5)

- Is Bayes rule optimal?
  - i.e., will rule minimize average probability of error?
- For a particular x,

$$p(error|x) = \begin{cases} P(w_1|x) & decide\ \ w_2 \\ P(w_2|x) & decide\ \ w_1 \end{cases}$$

  - This is as small as it can be
- Average probability of error

$$p(error) = \int_{-\infty}^{\infty} p(error|x)p(x)dx$$

Bayesian Decision Theory
Loss Function

- $\lambda(\alpha_i | \omega_j)$: cost incurred for taking action α_i (i.e., classification or rejection) when the state of nature is ω_j
- Example
  - ***x***: financial characteristics of firms applying for a bank loan
  - ω_0 – company did not go bankrupt

    ω_1 – company failed
  - $P(\omega_i|\mathbf{x})$ – predicted probability of bankruptcy
  - Confusion matrix:

	Algorithm: ω_0	Algorithm: ω_1
Truth: ω_0	TN	FP
Truth: ω_1	FN	TP

  - FN are 10 times as costly as FP

    $\Rightarrow \lambda(\alpha_0 | \omega_1) = \lambda_{01} = 10 \times \lambda(\alpha_1 | \omega_0) = 10 \times \lambda_{10}$

Bayesian Decision Theory
Minimum Risk Classifier

- Expected loss (or risk) associated with taking action α_i

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

- Overall risk

$$R = \int R(\alpha(x)|x)p(x)dx$$

- Decision function $\alpha(\boldsymbol{x})$ chosen so that $R(\alpha_i|\mathbf{x})$ is as small as possible for every x
- Decision rule: compute $R(\alpha_i|\mathbf{x})$ for $i = 1,\ldots a$ and select α_i for which $R(\alpha_i|\mathbf{x})$ is minimum

Bayesian Decision Theory
Minimum Risk Classifier (2)

- Two-category case

$$R(\alpha_0|x) = \lambda_{00}P(\omega_0|x) + \lambda_{01}P(\omega_1|x)$$
$$R(\alpha_1|x) = \lambda_{10}P(\omega_0|x) + \lambda_{11}P(\omega_1|x)$$

- Expressing minimum-risk rule: pick ω_o if $R(\alpha_0|\mathbf{x}) < R(\alpha_1|\mathbf{x})$, or

$$(\lambda_{10} - \lambda_{00})P(\omega_0|x) > (\lambda_{01} - \lambda_{11})P(\omega_1|x)$$

- In our loan example: $\lambda_{00} = \lambda_{11} = 0$

$$\frac{P(\omega_0|x)}{P(\omega_1|x)} > \frac{\lambda_{01}}{\lambda_{10}} \implies P(\omega_0|x) > 10 \times P(\omega_1|x)$$

Bayesian Decision Theory
Minimum Risk Classifier (3)

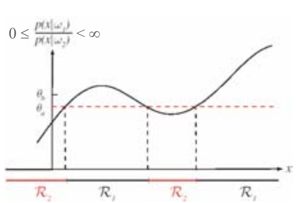
- Likelihood ratio: pick ω_1 if

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \underbrace{\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \times \frac{P(\omega_2)}{P(\omega_1)}}_{\theta}$$

- Zero-one loss

$$\lambda = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\Rightarrow \theta = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$



Bayesian Decision Theory
Minimum Error Rate Classifier

- Zero-one loss function leads to:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|x)$$
$$= \sum_{j \neq i} P(\omega_j|x)$$
$$= 1 - P(\omega_i|x)$$

- i.e., choose ω_i for which $P(\omega_i|\mathbf{x})$ is maximum
  - same rule as in Slide 6 as expected

Bayesian Decision Theory
Discriminant Function

- A useful way of representing a classifier
  - One function $g_i(\mathbf{x})$ for each class
  - Assign $\mathbf{x}$ to ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$
- Minimum risk: $g_i(\mathbf{x}) = - R(\alpha_i|\mathbf{x})$
- Minimum error: $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$
  - Monotonic increasing transformations are equivalent

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

Bayesian Decision Theory
Discriminant Function (2)

- Two-category case – dichotomizer
  - A single function suffices:

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

  - Decision rule:

Choose ω_1 if $g(\mathbf{x}) > 0$; *otherwise choose* ω_2

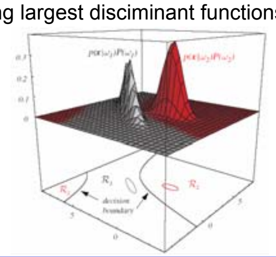
  - Convenient forms

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

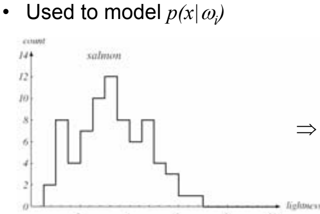
Bayesian Decision Theory
Decision Regions & Boundaries

- R_i region in feature space where $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$
  - Might not be simply connected
- Decision boundary: surfaces in feature space where ties occur among largest disciminant functions

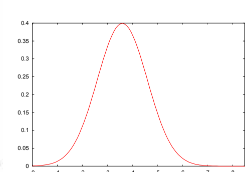


Normal Density
Introduction

- Used to model $p(x|\omega_i)$



⇒



- Special attention due to:
  - Analytically tractable
  - A continuous-valued feature x can be seen as randomly corrupted version of a single typical μ (asymptotically Gaussian)

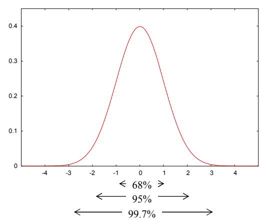
Normal Density
Univariate Case

- $x \sim N(0, 1)$ -- x is normally distributed with zero *mean* and unit *variance*

$$p_x(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

$$0 = \mu = \varepsilon[x]$$

$$1 = \sigma^2 = \varepsilon[(x-\mu)^2]$$



- Location-scale shift

$z = \sigma x + \mu$

$\sim N(\mu, \sigma)$

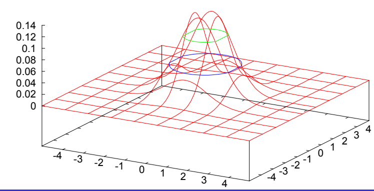
$$p_z(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} = \frac{1}{\sigma} p_x\left(\frac{z-\mu}{\sigma}\right)$$

Normal Density
Bivariate Case

- If $x \sim N(0, 1)$ and $y \sim N(0, 1)$ are independent

$$p(x,y) = p(x) \times p(y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}$$

- Contours: $p(x,y) = c_1 \;\Rightarrow x^2 + y^2 = c_2$

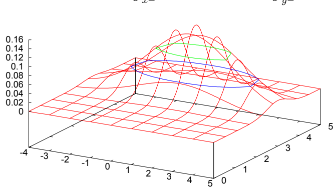


Normal Density
Bivariate Case (2)

- If $x \sim N(\mu_x, \sigma_x)$ and $y \sim N(\mu_y, \sigma_y)$ are independent

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - \frac{1}{2}\left(\frac{x-\mu_y}{\sigma_y}\right)^2}$$

- Contours: $\frac{1}{\sigma_x 2}(x-\mu_x)^2 + \frac{1}{\sigma_y 2}(y-\mu_y)^2 = c$



$$p(x,y) = \mathrm{N}\left(\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}, \begin{bmatrix}\sigma_x^2 & 0 \\ 0 & \sigma_y^2\end{bmatrix}\right)$$

$$= \mathrm{N}\left(\begin{bmatrix}1 \\ 3\end{bmatrix}, \begin{bmatrix}2^2 & 0 \\ 0 & \frac{1}{2}^2\end{bmatrix}\right)$$

variance-covariance matrix

Normal Density
Multivariate Case

- We say $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

where,

$\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^t$ (t stands for the transpose vector form)

$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma = d \times d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

$(\boldsymbol{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ is (square) *Mahalanobis* distance

Bayesian Decision Theory
Discriminant Function – Normal Density

- $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
- We had $g_i(x) = \ln p(x \mid \omega_i) + \ln P(\omega_i)$

$$\Rightarrow \quad g_i(x) = -\tfrac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \tfrac{d}{2}\ln 2\pi$$
$$-\frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- Case 1: $\Sigma_i = \sigma^2 I$
- Case 2: $\Sigma_i = \Sigma$

} linear discriminant function

- Case 3: $\Sigma_i = arbitrary$

Bayesian Decision Theory
Discriminant Function – Normal Density (2)

- Case 1: features are statistically independent ($\sigma_{ij} = 0$) and share same variance σ^2

$$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$
$$= -\frac{1}{2\sigma^2}[x^t x - 2\mu_i^t x + \mu_i^t \mu_i] + \ln P(\omega_i)$$
$$= \boxed{w_i^t x + w_{i0}}$$

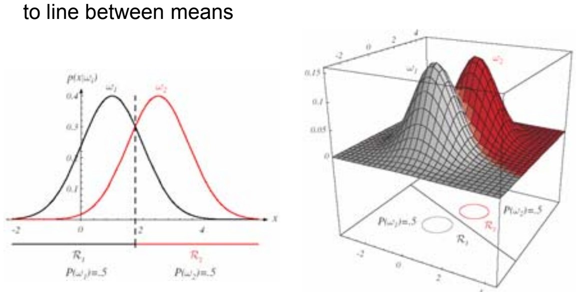
where $w_i = \frac{1}{\sigma^2}\mu_i$

$w_{i0} = -\frac{1}{2\sigma^2}[\mu_i^t \mu_i] + \ln P(\omega_i)$

- All priors equal $\Rightarrow$ Minimum (Euclidean) distance classifier

Bayesian Decision Theory
Discriminant Function – Normal Density (3)

- Case 1: distributions are "spherical" in d dimensions; boundary is a *hyperplane* in d-1 dimensions perpendicular to line between means



Bayesian Decision Theory
Discriminant Function – Normal Density (4)

- Case 2: samples fall in hyperellipsoidal clusters of equal size and shape

$$g_i(x) = -\tfrac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) + \ln P(\omega)$$
$$= w_i^t x + w_{i0} \quad as\ \ x^t \Sigma^{-1} x\ \ can\ \ be\ \ dropped$$

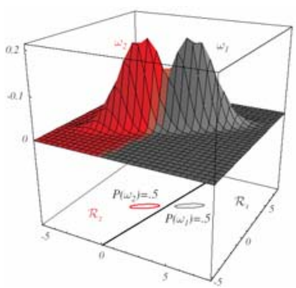
where $w_i = \Sigma^{-1}\mu_i$

$w_{i0} = -\frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i + \ln P(\omega_i)$

- All priors equal $\Rightarrow$ Minimum (Mahalanobis) distance classifier

Bayesian Decision Theory
Discriminant Function – Normal Density (5)

- Case 2: hyperplane separating class regions is generally not perpendicular to line between the means



Bayesian Decision Theory
Discriminant Function – Normal Density (6)

- Case 3: decision surfaces are hyperquadratics (i.e., hyperplanes, pairs of hyperplanes, hypershpheres, hyperellipsoids, hyperhyperboloids)