



Pattern classification

Biometrics
CSE 190
Lecture 4

CSE190a, Winter 2011

Announcements

- Readings on E-reserves
- HW1 will be posted shortly
- Project description on web page

CSE190a, Winter 2011

Key Probabilities

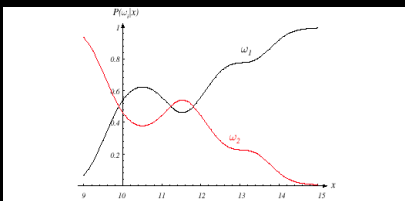
- ω_j – class label
- X – feature vector
- $P(\omega_j)$ – Prior probability of class j
- $P(x)$ – Probability distribution function of feature values.
- $P(x | \omega_j)$ – Class conditional density function or likelihood of feature x given ω_j
- $P(\omega_j | x)$ – posterior probability density function of the class given the feature.

Pattern Classification, Chapter 1

- Posterior, likelihood, evidence
 - $P(\omega_j | x) = (P(x | \omega_j) * P(\omega_j)) / P(x)$ **(BAYES RULE)**
 - In words, this can be said as:
Posterior = (Likelihood * Prior) / Evidence
 - Where in case of two categories

$$P(x) = \sum_{j=1}^{j=2} P(x | \omega_j) P(\omega_j)$$

Pattern Classification, Chapter 1



- Intuitive decision rule given the posterior probabilities:
Given x :
if $P(\omega_1 | x) > P(\omega_2 | x)$ $\Rightarrow$ True state of nature = ω_1
if $P(\omega_1 | x) < P(\omega_2 | x)$ $\Rightarrow$ True state of nature = ω_2

Why do this?: Whenever we observe a particular x , the probability of error is :

$$P(\text{error} | x) = P(\omega_1 | x) \text{ if we decide } \omega_2$$

$$P(\text{error} | x) = P(\omega_2 | x) \text{ if we decide } \omega_1$$

Pattern Classification, Chapter 1

Bayesian Decision Theory – Continuous Features

- Let $\mathbf{X}$ be a vector of features.
- Let $\{\omega_1, \omega_2, \dots, \omega_c\}$ be the set of c states of nature (or "classes")
- Let $\{\alpha_1, \alpha_2, \dots, \alpha_d\}$ be the set of possible actions
- Let $\lambda(\alpha_i | \omega_j)$ be the loss for action α_i when the state of nature is ω_j

Pattern Classification, Chapter 1

What is the Expected Loss for action α_i ?

For any given x the expected loss is

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

$R(\alpha_i | x)$ is called the **Conditional Risk (or Expected Loss)**

Pattern Classification, Chapter 1

Given a measured feature vector x , which action should we take?

Select the action α_i for which $R(\alpha_i | x)$ is minimum

$$\alpha = \arg \min_{\alpha_i} R(\alpha_i | x)$$

Pattern Classification, Chapter 1

Likelihood ratio:

The preceding rule is equivalent to the following rule:

$$\text{if } \frac{P(x | \omega_1)}{P(x | \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

Then take action α_1 (decide ω_1)
Otherwise take action α_2 (decide ω_2)

Pattern Classification, Chapter 1

Classifiers, Discriminant Functions and Decision Surfaces

- Discriminant Functions: A generalization
- The multi-category case
 - Consider a set of c discriminant functions $g_i(x)$, $i = 1, \dots, c$
 - The classifier assigns a feature vector x to class ω_i if:

$$g_i(x) > g_j(x) \quad \forall j \neq i$$
 - Designing a classifier amounts to specifying the $g_i(x)$

Pattern Classification, Chapter 1

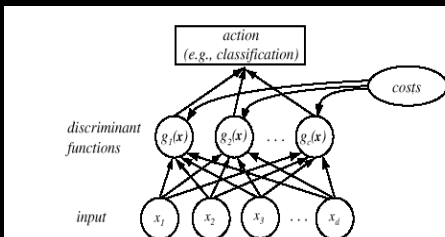


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(x)$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Pattern Classification, Chapter 1

Decision Regions

- Feature space divided into c decision region
 - if $g_i(x) > g_j(x) \quad \forall j \neq i$ then x is in $\mathcal{R}_i$
 - ($\mathcal{R}_i$ means assign x to ω_i)

Decision surfaces

- Boundary between decision regions.

$$\{x: \exists i, j \quad g_i(x) = g_j(x)\}$$

Pattern Classification, Chapter 1

- Bayes Risk as discriminant function.
 - Let $g_i(x) = -R(\alpha_i | x)$
(max. discriminant corresponds to min. risk!)
- For the minimum error rate, discriminant function is:

$$g_i(x) = P(\omega_i | x)$$
 (max. discrimination corresponds to max. posterior!)

$$g_i(x) = P(x | \omega_i) P(\omega_i)$$
- Any function $F(r)$ which is monotonic over $r > 0$ when applied to a set of discriminant functions, yields a new discriminant function with the same decision regions/boundaries.

$$g_i(x) = \ln P(x | \omega_i) + \ln P(\omega_i)$$

(ln: natural logarithm!)

We'll see this form with Normal distributions

Pattern Classification, Chapter 1

The Normal Density

Univariate density

- Density which is analytically tractable
- Continuous density
- A lot of processes are asymptotically Gaussian
- Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

$$P(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right],$$

Where:

μ = mean (or expected value) of x

σ^2 = expected squared deviation or variance

Pattern Classification, Chapter 1

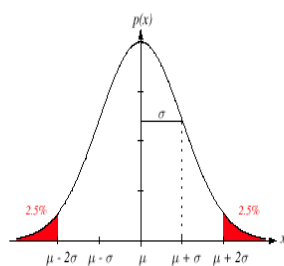
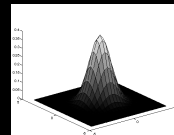


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Pattern Classification, Chapter 1

Multivariate density



- Multivariate normal density in d dimensions is:

$$P(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right]$$

where:

$x = (x_1, x_2, \dots, x_d)'$ ($'$ stands for the transpose vector form)

$\mu = (\mu_1, \mu_2, \dots, \mu_d)'$ mean vector

$\Sigma = d$ by d covariance matrix

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

Pattern Classification, Chapter 1

Discriminant Functions for the Normal Density

- We saw that the minimum error-rate classification can be achieved by the discriminant function

$$g_i(x) = \ln P(x | \omega_i) + \ln P(\omega_i)$$

- Case of multivariate normal for class condition density (likelihood) function

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

$$= -\frac{1}{2} (x - \mu_i)' \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

Pattern Classification, Chapter 1

CASE I

Case $\Sigma_i = \sigma^2 I$ (I stands for the identity matrix)

$$g_i(x) = w_i' x + w_{i0} \text{ (linear discriminant function)}$$

where:

$$w_i = \frac{\mu_i}{\sigma^2}; \quad w_{i0} = -\frac{1}{2\sigma^2} \mu_i' \mu_i + \ln P(\omega_i)$$

(w_{i0} is called the threshold for the i th category!)

Pattern Classification, Chapter 2 (Part 3)

- A classifier that uses linear discriminant functions is called "a linear machine"
- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

$$g_i(x) = g_j(x)$$

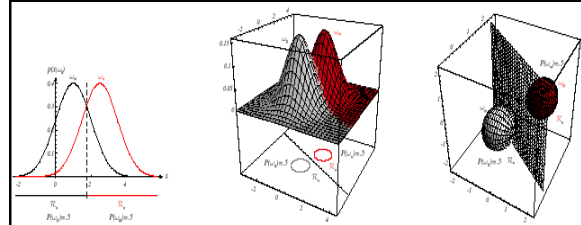


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d - 1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(x|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates R_1 from R_2 . From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The hyperplane separating R_1 and R_2 are given by

$$\mathbf{w}'(\mathbf{x} - \mathbf{x}_0) = 0$$

$$\mathbf{w} = \mu_i - \mu_j$$

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

always orthogonal to the line linking the means!

$$\text{if } P(\omega_i) = P(\omega_j) \text{ then } \mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j)$$

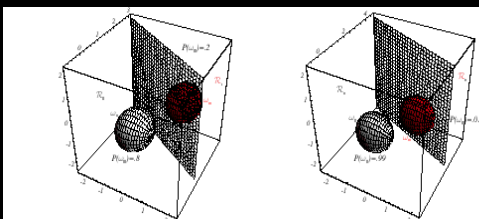
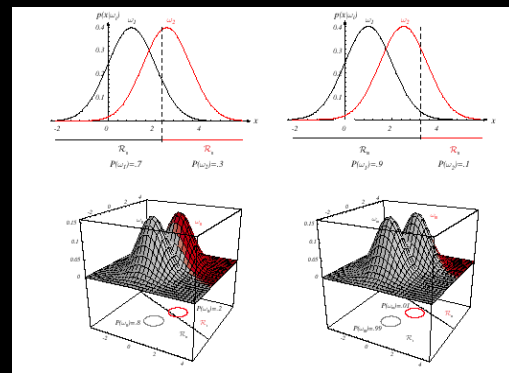


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Case II

Case $\Sigma_i = \Sigma$

(covariance of all classes are identical but arbitrary!)

$$\begin{aligned} g_i(x) &= -\frac{1}{2}(x - \mu_i)' \Sigma_i^{-1}(x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i) \\ &= -\frac{1}{2}(x - \mu_i)' \Sigma^{-1}(x - \mu_i) - \frac{1}{2} \ln |\Sigma| + \ln P(\omega_i) \\ &= -\frac{1}{2}(x - \mu_i)' \Sigma^{-1}(x - \mu_i) + \ln P(\omega_i) \end{aligned}$$

Now consider case where $P(\omega_i) = P(\omega_j)$,

$$g_i(x) = -(x - \mu_i)' \Sigma^{-1}(x - \mu_i)$$

$(x - \mu_i)' \Sigma^{-1}(x - \mu_i)$ is sometimes called the Mahalanobis distance

Case II

27

Case $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)

Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

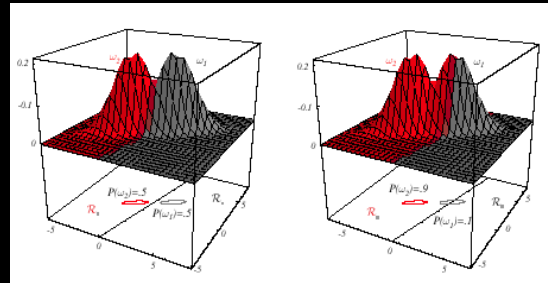
$$\mathbf{w}'(\mathbf{x} - \mathbf{x}_0) = 0$$

$$\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$$

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_j)' \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

Here the hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ is generally not orthogonal to the line between the means!

Pattern Classification, Chapter 2 (Part 3)



Pattern Classification, Chapter 1

28

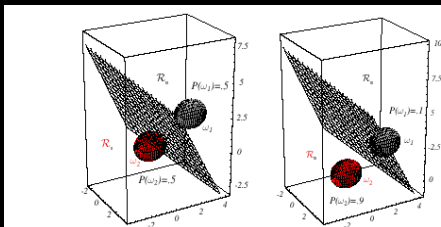


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Pattern Classification, Chapter 1

CASE III

30

Case Σ_i is arbitrary

The covariance matrices are different for each category

$$g_i(\mathbf{x}) = \mathbf{x}'\mathbf{W}_i\mathbf{x} + \mathbf{w}_i'\mathbf{x} + w_{i0}$$

where:

$$\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$$

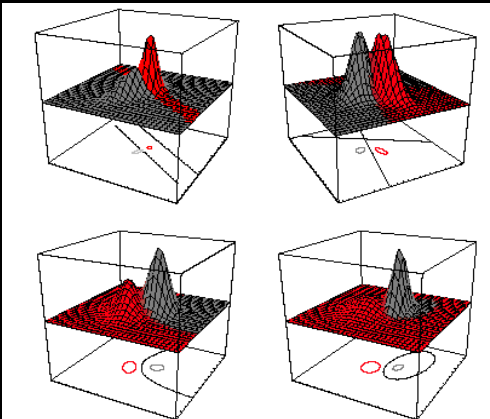
$$\mathbf{w}_i = \Sigma_i^{-1}\mu_i$$

$$w_{i0} = -\frac{1}{2}\mu_i'\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Here the separating surfaces are **Hyperquadrics** which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids)

Pattern Classification, Chapter 2 (Part 3)

31



Pattern Classification, Chapter 1

32

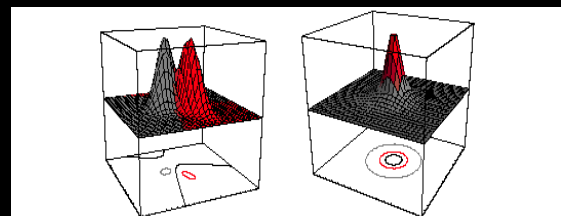


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Pattern Classification, Chapter 1