Pattern classification

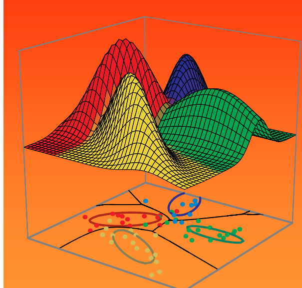
Biometrics
CSE 190-a
Lecture 4



## Announcements

- Readings on E-reserves
- Project description on web page





**Pattern Classification**

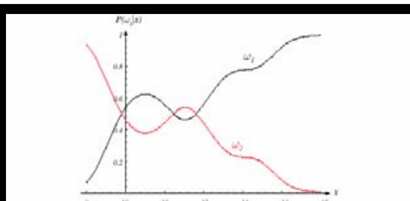


- Posterior, likelihood, evidence
  - $P(\omega_j \mid x) = (P(x \mid \omega_j) * P(\omega_j)) / P(x)$ **(BAYES RULE)**
  - In words, this can be said as:
    Posterior = (Likelihood * Prior) / Evidence
  - Where in case of two categories

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







- Intuitive decision rule given the posterior probabilities:
  Given *x*:
  if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ➡ True state of nature = $\omega_1$
  if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ➡ True state of nature = $\omega_2$

  Why do this?: Whenever we observe a particular *x*, the probability of error is :
  $P(error \mid x) = P(\omega_1 \mid x)$ if we decide $\omega_2$
  $P(error \mid x) = P(\omega_2 \mid x)$ if we decide $\omega_1$





## Bayesian Decision Theory – Continuous Features

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

7

## What is the Expected Loss for action $\alpha_i$?

For any given x the expected loss is

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

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



8

Overall risk

*R = Sum of all* $\underbrace{R(\alpha_i \mid x)}_{\textbf{Conditional risk}}$ *for i = 1,…,a*

Minimizing R ⟺ Minimizing $R(\alpha_i \mid x)$ *for i = 1,…, a*

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

for i = 1,…,a



9

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

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

⇒ R is minimum and R in this case is called the Bayes risk = best performance that can be achieved!



10

## Two-Category Classification

$\alpha_1$ : *deciding* $\omega_1$

$\alpha_2$ : *deciding* $\omega_2$

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

loss incurred for deciding $\omega_i$ when the true state of nature is $\omega_j$

Conditional risk:

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



11

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

$\lambda_{11} P(\omega_1 \mid x) + \lambda_{12} P(\omega_2 \mid x) < \lambda_{21} P(\omega_1 \mid x) + \lambda_{22} P(\omega_2 \mid x)$

action $\alpha_1$: "decide $\omega_1$" is taken

This results in the equivalent rule :

decide $\omega_1$ if:

$$(\lambda_{21} - \lambda_{11}) P(x \mid \omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) P(x \mid \omega_2) P(\omega_2)$$

and decide $\omega_2$ otherwise



12

Likelihood ratio:

The preceding rule is equivalent to the following rule:

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

Then take action $\alpha_1$ (decide $\omega_1$)

Otherwise take action $\alpha_2$ (decide $\omega_2$)

13

## Classifiers, Discriminant Functions and Decision Surfaces

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

    $g_i(x) > g_j(x) \; \forall j \neq I$
  - *Designing a classifer amounts to specifying the $g_i(x)$*



14

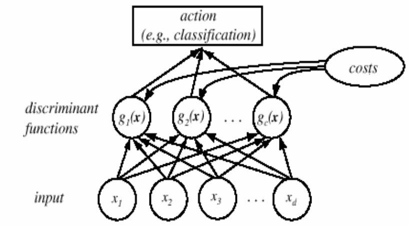



FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes $d$ inputs and $c$ discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the 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.



15

## Decision Regions

- Feature space divided into c decision region

  *if $g_i(x) > g_j(x) \; \forall j \neq i$ then x is in $\mathcal{R}_i$*

  ($\mathcal{R}_i$ means assign $x$ to $\omega_i$)

## Decision surfaces

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



16

- *Bayes Risk as discriminant function.*
  - *Let $g_i(x) = - R(\alpha_i \mid x)$*

    (max. discriminant corresponds to min. risk!)
- For the minimum error rate, discriminant function is:

  $g_i(x) = P(\omega_i \mid x)$

  (max. discrimination corresponds to max. posterior!)

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

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

  (ln: natural logarithm!)

We'll see this form with Normal distributions



19

## 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:

$\mu$ = mean (or expected value) of $x$

$\sigma^2$ = expected squared deviation or variance



20

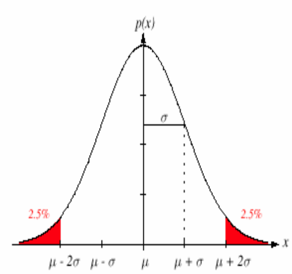
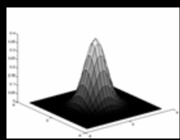
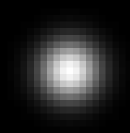



FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John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

## Multivariate density







- Multivariate normal density in d dimensions is:

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

where:
$x = (x_1, x_2, \ldots, x_d)^t$ (t stands for the transpose vector form)
$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector
$\Sigma = d$ *by* $d$ covariance matrix
$|\Sigma|$ and $\Sigma^{-1}$ are determinant and inverse respectively



## Discriminant Functions for the Normal Density



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

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

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

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





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

$$g_i(x) = w_i^t x + w_{i0} \text{ (linear discriminant function)}$$
*where :*
$$w_i = \frac{\mu_i}{\sigma^2};\ w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t\mu_i + \ln P(\omega_i)$$
$(\omega_{i0}$ *is called the threshold for the ith category!* $)$





- 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:

$g_i(x) = g_j(x)$





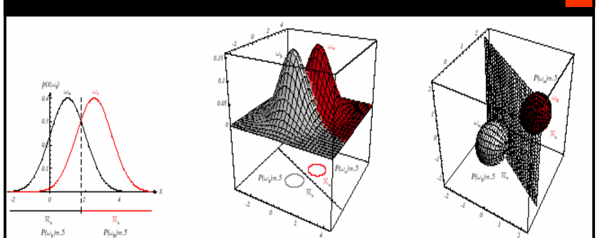


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in $d$ dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



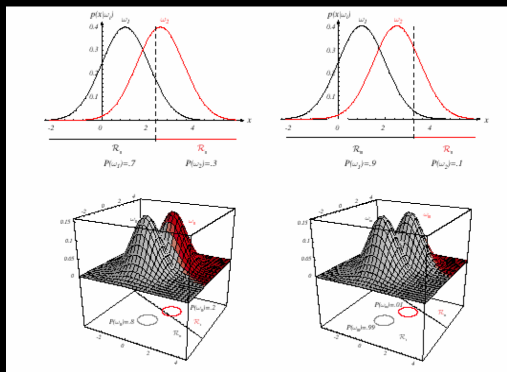


The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ are given by

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0$$
$$\mathbf{w}=\mu_i-\mu_j$$
$$\mathbf{x}_0=\frac{1}{2}(\mu_i+\mu_j)-\frac{\sigma^2}{\|\mu_i-\mu_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\mu_i-\mu_j)$$

always orthogonal to the line linking the means!

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







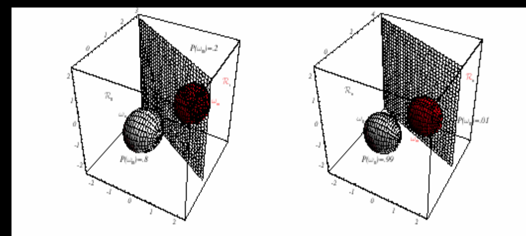


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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





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

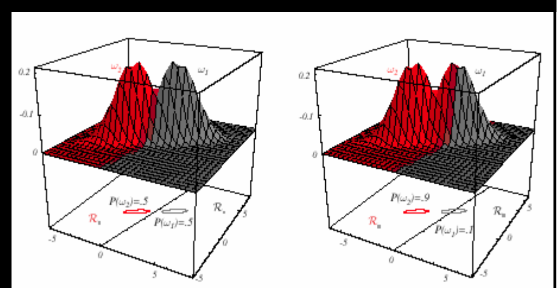
Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0$$
$$\mathbf{w}=\Sigma^{-1}(\mu_i-\mu_j)$$
$$x_0=\frac{1}{2}(\mu_i+\mu_j)-\frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i-\mu_j)^t\Sigma^{-1}(\mu_i-\mu_j)}(\mu_i-\mu_j)$$

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











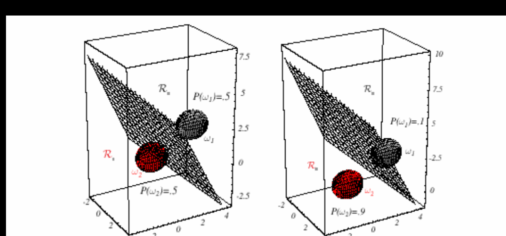


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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





Case $\Sigma_i$ = arbitrary

The covariance matrices are different for each category

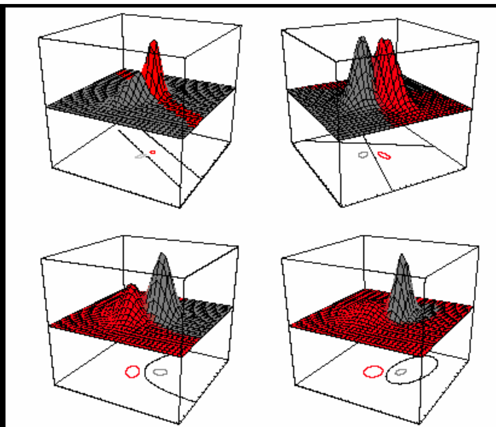
$$g_i(x)=x^tW_ix+w_i^tx=w_{i0}$$

*where :*

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

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

33



Pattern Classification, Chapter 1

34

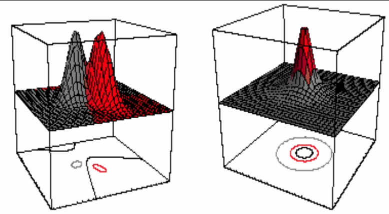


**FIGURE 2.14.** Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant 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