

## Part II Bayesian Decision Theory

NASA Space Program

0

## Outline

- Introduction
- Bayesian Decision Theory–Continuous Features
- Minimum Error Rate Classification
- Classifiers, Discriminant Functions and Decision Surfaces
- The Normal Density
- Discriminant Functions for the Normal Density
- Bayes Decision Theory – Discrete Features

1

## Introduction

### • The sea bass/salmon example

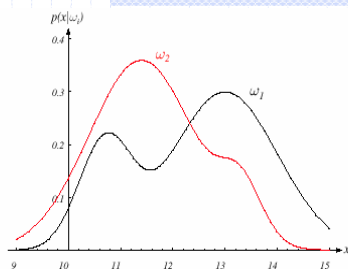
- State of nature, prior
  - State of nature is a random variable
  - The catch of salmon and sea bass is equiprobable
- $P(\omega_1) = P(\omega_2)$  (uniform priors)
- $P(\omega_1) + P(\omega_2) = 1$  (exclusivity and exhaustivity)

2

## Introduction

- Decision rule with only the prior information
  - Decide  $\omega_1$  if  $P(\omega_1) > P(\omega_2)$  otherwise decide  $\omega_2$
- Use of the class – conditional information
- $P(x | \omega_1)$  and  $P(x | \omega_2)$  describe the difference in lightness between populations of sea and salmon

3



**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  $\omega_i$ . 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.

4

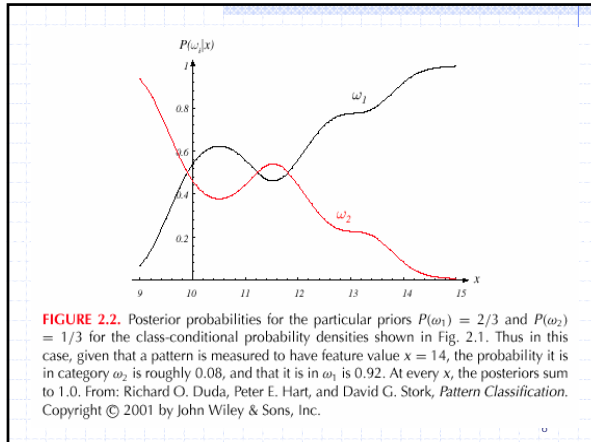
## Introduction

- Posterior, likelihood, evidence
  - $P(\omega_j | x) = P(x | \omega_j) \cdot P(\omega_j) / P(x)$
  - Where in case of two categories

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

- Posterior = (Likelihood · Prior) / Evidence

5



## Introduction

- Decision given the posterior probabilities  
 $X$  is an observation for which:
  - if  $P(\omega_1 | x) > P(\omega_2 | x)$   $\Rightarrow$  True state of nature =  $\omega_1$
  - if  $P(\omega_1 | x) < P(\omega_2 | x)$   $\Rightarrow$  True state of nature =  $\omega_2$

Therefore:

whenever we observe a particular  $x$ , the probability of error is :

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

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

## Introduction

- Minimizing the probability of error
- Decide  $\omega_1$  if  $P(\omega_1 | x) > P(\omega_2 | x)$ ; otherwise decide  $\omega_2$

Therefore:

$$P(\text{error} | x) = \min [P(\omega_1 | x), P(\omega_2 | x)]$$

(Bayes decision)

## Bayesian Decision Theory – Continuous Features

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions and not only decide on the state of nature
  - Introduce a loss of function which is more general than the probability of error

## Features

- Allowing actions other than classification primarily allows the possibility of rejection
- Refusing to make a decision in close or bad cases!
- The loss function states how costly each action taken is

## Preliminaries

- Let  $\{\omega_1, \omega_2, \dots, \omega_c\}$  be the set of  $c$  states of nature (or "categories")
- Let  $\{\alpha_1, \alpha_2, \dots, \alpha_a\}$  be the set of possible actions
- Let  $\lambda(\alpha_i | \omega_j)$  be the loss incurred for taking action  $\alpha_i$  when the state of nature is  $\omega_j$

## Preliminaries

Overall risk

$R = \text{Sum of all } R(\alpha_i | x) \text{ for } i = 1, \dots, a$

Conditional risk

Minimizing  $R \iff$  Minimizing  $R(\alpha_i | x)$  for  $i = 1, \dots, a$

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

for  $i, j = 1, \dots, a$

12

## Preliminaries

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



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

13

## Two-category classification

$\alpha_1$  : deciding  $\omega_1$

$\alpha_2$  : deciding  $\omega_2$

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

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

Conditional risk:

$$R(\alpha_1 | x) = \lambda_{11}P(\omega_1 | x) + \lambda_{12}P(\omega_2 | x)$$

$$R(\alpha_2 | x) = \lambda_{21}P(\omega_1 | x) + \lambda_{22}P(\omega_2 | x)$$

14

## Two-category classification

Our rule is the following:

if  $R(\alpha_1 | x) < R(\alpha_2 | x)$   
action  $\alpha_j$ ; "decide  $\omega_j$ " is taken

This results in the equivalent rule :

decide  $\omega_1$  if:

$$(\lambda_{21} - \lambda_{11}) P(x | \omega_1) P(\omega_1) >$$

$$(\lambda_{12} - \lambda_{22}) P(x | \omega_2) P(\omega_2)$$

and decide  $\omega_2$  otherwise

15

## Two-category classification

◆ 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  $\alpha_1$  (decide  $\omega_1$ )

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

16

## Two-category classification

◆ Optimal decision property

"If the likelihood ratio exceeds a threshold value independent of the input pattern  $x$ , we can take optimal actions"

17

## Exercise

Select the optimal decision where:

$$\Omega = \{\omega_1, \omega_2\}$$

$$P(x | \omega_1) \longrightarrow N(2, 0.5) \text{ (Normal distribution)}$$

$$P(x | \omega_2) \longrightarrow N(1.5, 0.2)$$

$$P(\omega_1) = 2/3$$

$$P(\omega_2) = 1/3$$

$$\lambda = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

18

## Minimum-Error-Rate Classification

• Actions are decisions on classes

If action  $\alpha_i$  is taken and the true state of nature is  $\omega_j$  then:

the decision is correct if  $i = j$  and in error if  $i \neq j$

• Seek a decision rule that minimizes the *probability of error* which is the *error rate*

19

## Minimum-Error-Rate Classification

• Introduction of the zero one loss function:

$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, \dots, c$$

Therefore, the conditional risk is:

$$\begin{aligned} R(\alpha_i | x) &= \sum_{j=1}^c \lambda(\alpha_i, \omega_j) P(\omega_j | x) \\ &= \sum_{j \neq i} P(\omega_j | x) = 1 - P(\omega_i | x) \end{aligned}$$

*"The risk corresponding to this loss function is the average probability error"*

20

## Minimum-Error-Rate Classification

• Minimize the risk requires maximize

$$P(\omega_i | x)$$

$$\text{(since } R(\alpha_i | x) = 1 - P(\omega_i | x)\text{)}$$

• For Minimum error rate

- Decide  $\omega_i$  if  $P(\omega_i | x) > P(\omega_j | x) \forall j \neq i$

21

## Minimum-Error-Rate Classification

• Regions of decision and zero one loss function, therefore:

$$\text{Let } \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then decide } \omega_1 \text{ if } : \frac{P(x | \omega_1)}{P(x | \omega_2)} > \theta_\lambda$$

• If  $\lambda$  is the zero one loss function which means:

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

$$\text{then } \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_0$$

$$\text{if } \lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \text{ then } \theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_0$$

22

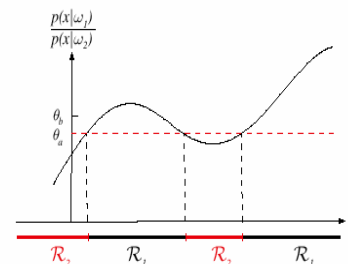


FIGURE 2.3. The likelihood ratio  $p(x|\omega_1)/p(x|\omega_2)$  for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold  $\theta_0$ . If our loss function penalizes miscategorizing  $\omega_2$  as  $\omega_1$  patterns more than the converse, we get the larger threshold  $\theta_0$ , and hence  $R_1$  becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

23

## Classifiers, Discriminant Functions and Decision Surfaces

### The multi-category case

- Set of discriminant functions  $g_i(x)$ ,  $i = 1, \dots, c$
- The classifier assigns a feature vector  $x$  to class  $\omega_i$  if:

$$g_i(x) > g_j(x) \quad \forall j \neq i$$

24

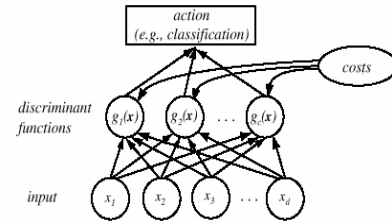


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes  $d$  inputs and  $c$  discriminant functions  $g_i(x)$ . A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction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

25

## Discriminant Functions

- Let  $g_i(x) = -R(\alpha_i | x)$   
(max. discriminant corresponds to min. risk!)
- For the minimum error rate, we take  
 $g_i(x) = P(\omega_i | x)$   
(max. discrimination corresponds to max. posterior!)

$$g_i(x) \equiv P(x | \omega_i) P(\omega_i)$$

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

(ln: natural logarithm!)

26

## Discriminant Functions

- Feature space divided into  $c$  decision regions  
if  $g_i(x) > g_j(x) \quad \forall j \neq i$  then  $x$  is in  $R_i$   
( $R_i$  means assign  $x$  to  $\omega_i$ )
- The two-category case
  - A classifier is a "dichotomizer" that has two discriminant functions  $g_1$  and  $g_2$
  - Let  $g(x) \equiv g_1(x) - g_2(x)$
  - Decide  $\omega_1$  if  $g(x) > 0$ ; Otherwise decide  $\omega_2$

27

## Discriminant Functions

- The computation of  $g(x)$

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

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

28

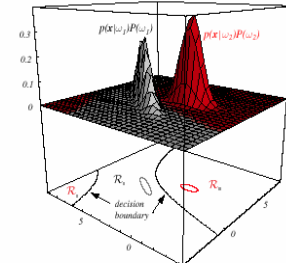


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region  $R_2$  is not simply connected. The ellipses mark where the density is  $1/e$  times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

29

## 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

30

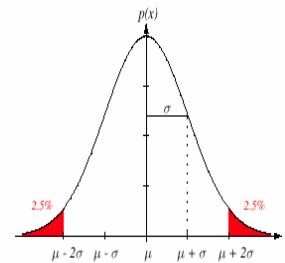


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range  $|x - \mu| \leq 2\sigma$ , as shown. The peak of the distribution has value  $p(\mu) = 1/\sqrt{2\pi}\sigma$ . From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by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

31

## The Normal Density

### 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:

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

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

$\Sigma = d \times d$  covariance matrix

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

32

## 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

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

33

## Discriminant Functions for the Normal Density

- 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 :

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

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

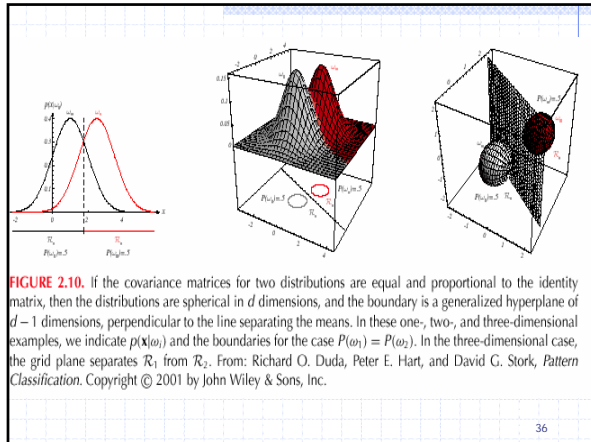
34

## Discriminant Functions for the Normal Density

- 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:

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

35



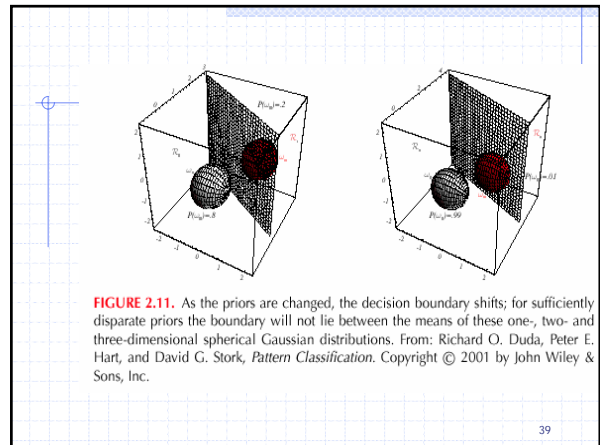
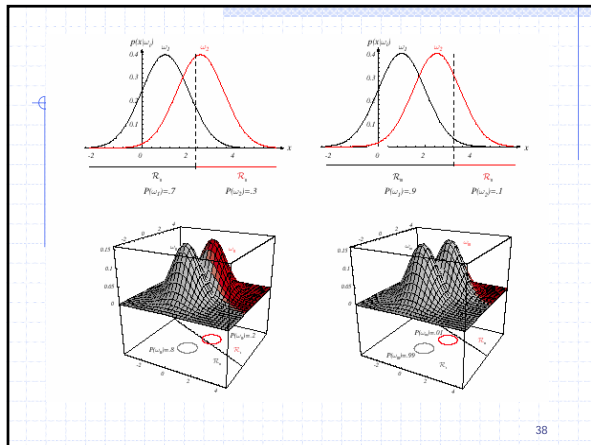
### Discriminant Functions for the Normal Density

- The hyperplane separating  $R_i$  and  $R_j$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

always orthogonal to the line linking the means!

if  $P(\omega_i) = P(\omega_j)$  then  $x_0 = \frac{1}{2}(\mu_i + \mu_j)$

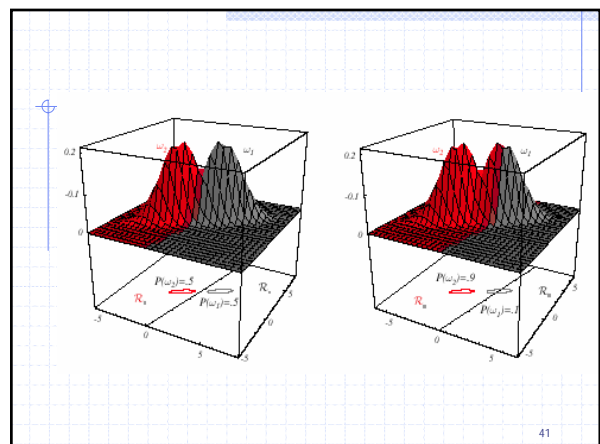


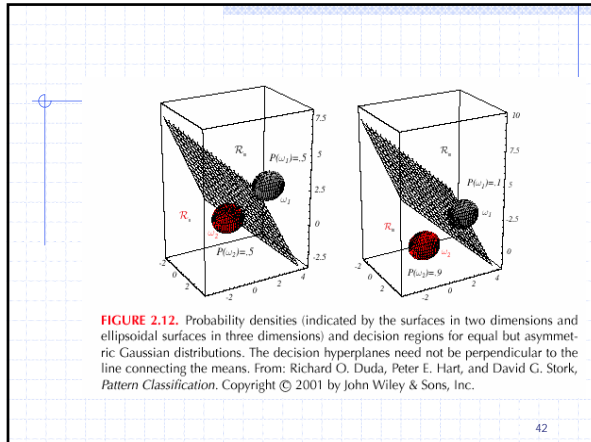
### Discriminant Functions for the Normal Density

- Case  $\Sigma_i = \Sigma$  (covariance of all classes are identical but arbitrary!)
  - Hyperplane separating  $R_i$  and  $R_j$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_j)' \Sigma^{-1} (\mu_i - \mu_j)} (\mu_i - \mu_j)$$

(the hyperplane separating  $R_i$  and  $R_j$  is generally not orthogonal to the line between the means!)





### Discriminant Functions for the Normal Density

- Case  $\Sigma_i = \text{arbitrary}$ 
  - The covariance matrices are different for each category

$$g_i(x) = x'W_i x + w_i'x = w_{i0}$$

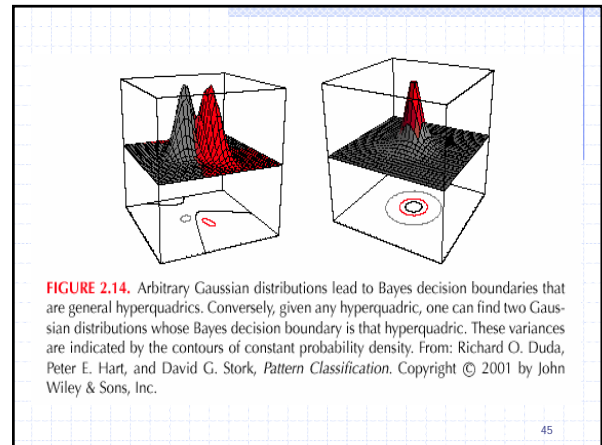
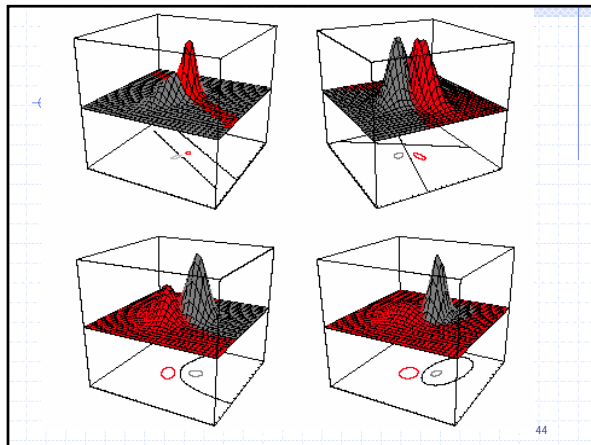
where :

$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i$$

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

(Hyperquadrics which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperbhyperboloids)



### Bayes Decision Theory – Discrete Features

- Components of  $x$  are binary or integer valued,  $x$  can take only one of  $m$  discrete values
 
$$V_1, V_2, \dots, V_m$$
- Case of independent binary features in 2 category problem
 

Let  $x = [x_1, x_2, \dots, x_d]'$  where each  $x_i$  is either 0 or 1, with probabilities:

$$p_i = P(x_i = 1 | \omega_1)$$

$$q_i = P(x_i = 1 | \omega_2)$$

### Bayes Decision Theory – Discrete Features

- The discriminant function in this case is:
 
$$g(x) = \sum_{i=1}^d w_i x_i + w_0$$

where :

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \quad i = 1, \dots, d$$

and :

$$w_0 = \sum_{i=1}^d \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

decide  $\omega_1$  if  $g(x) > 0$  and  $\omega_2$  if  $g(x) \leq 0$

## Exercise

Given:

$$\omega_1 = [34; 26; 46; 38]$$

$$\omega_2 = [30; 1-2; 5-2; 3-4]$$

Apply the discriminant function in case 3  
to find the decision boundary

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\Sigma_1^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix} \quad \Sigma_2^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$$

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

48