CS 559: Machine Learning Fundamentals and Applications 3rd Set of Notes

Instructor: Philippos Mordohai
Webpage: www.cs.stevens.edu/~mordohai
E-mail: Philippos.Mordohai@stevens.edu
Office: Lieb 215

Overview

- Making Decisions
- Parameter Estimation
  - Frequentist or Maximum Likelihood approach

Expected Utility

- You are asked if you wish to take a bet on the outcome of tossing a fair coin. If you bet and win, you gain $100. If you bet and lose, you lose $200. If you don't bet, the cost to you is zero.

  U(win, bet) = 100 U(lose, bet) = -200

  U(win, no bet) = 0 U(lose, no bet) = 0

- Your expected winnings/losses are:

  $U(\text{bet}) = p(\text{win}) \times U(\text{win, bet}) + p(\text{lose}) \times U(\text{lose, bet})$

  $= 0.5 \times 100 - 0.5 \times 200 = -50$

  $U(\text{no bet}) = 0$

- Based on making the decision which maximizes expected utility, you would therefore be advised not to bet.

Flow of Lecture and Entire Course

- Making optimal decisions based on prior knowledge (prev. slide)
- Making optimal decisions based on observations and prior knowledge
  - Given models of the underlying phenomena (last week and today)
  - Given training data with observations and labels (most of the semester)

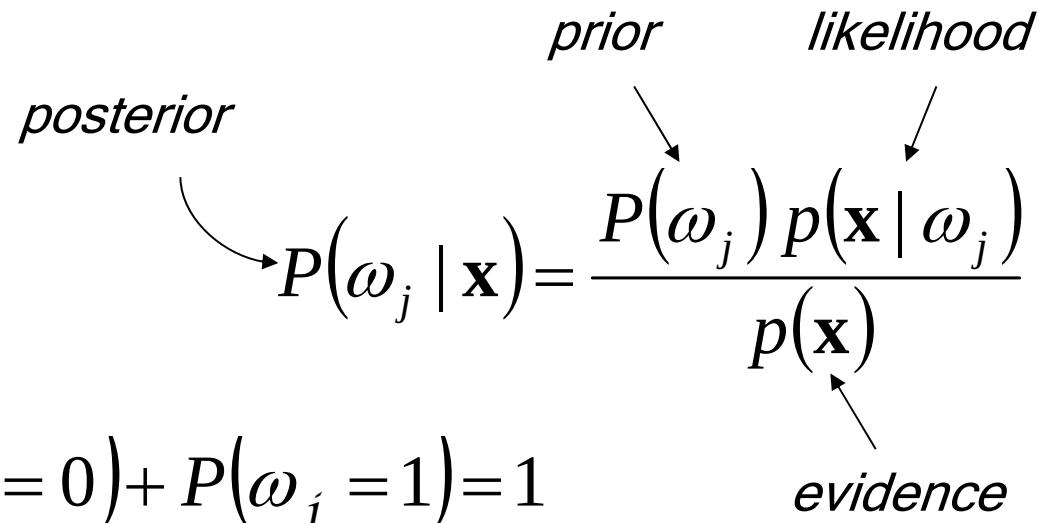
Bayesian Decision Theory

Adapted from:
Duda, Hart and Stork, Pattern Classification textbook

O. Veksler

E. Sudderth

Bayes' Rule



$$P(\omega_j = 0) + P(\omega_j = 1) = 1$$

$$p(\mathbf{x}) = p(\mathbf{x} \mid \omega_j = 1) P(\omega_j = 1) + p(\mathbf{x} \mid \omega_j = 0) P(\omega_j = 0)$$

$$p(\omega_j = 0 \mid \mathbf{x}) + p(\omega_j = 1 \mid \mathbf{x}) = 1$$

Bayes Rule - Intuition

The essence of the Bayesian approach is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence.

In other words, it allows scientists to combine new data with their existing knowledge or expertise.

From the Economist (2000)

Bayes Rule - Intuition

The canonical example is to imagine that a precocious newborn observes his first sunset, and wonders whether the sun will rise again or not. He assigns equal prior probabilities to both possible outcomes, and represents this by placing one white and one black marble into a bag. The following day, when the sun rises, the child places another white marble in the bag. The probability that a marble plucked randomly from the bag will be white (ie, the child's degree of belief in future sunrises) has thus gone from a half to two-thirds. After sunrise the next day, the child adds another white marble, and the probability (and thus the degree of belief) goes from two-thirds to three-quarters. And so on.

Gradually, the initial belief that the sun is just as likely as not to rise each morning is modified to become a near-certainty that the sun will always rise.

Bayesian Decision Theory

- Knowing the probability distribution of the categories
- We do not even need training data to design optimal classifiers

- Rare in real life

Prior

- Prior comes from prior knowledge, no data have been seen yet
- If there is a reliable source of prior knowledge, it should be used
- Some problems cannot even be solved reliably without a good prior
- However prior alone is not enough, we still need likelihood

Decision Rule based on Priors

- Model state of nature as a random variable, ω:
  - $\omega = \omega_1$: the event that the next sample is from category 1
  - $P(\omega_1)$ = probability of category 1
  - $P(\omega_2)$ = probability of category 2
  - $P(\omega_1) + P(\omega_2) = 1$
    - Exclusivity: ω_1 and ω_2 share no events
    - Exhaustivity: the union of all outcomes is the sample space (either ω_1 or ω_2 must occur)
- If all incorrect classifications have an equal cost:

➢ Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise, decide ω_2

Using Class-Conditional Information

- Use of the class-conditional information can improve accuracy
- $p(x \mid \omega_1)$ and $p(x \mid \omega_2)$ describe the difference in feature x between category 1 and category 2

Class-conditional Density vs. Likelihood

- Class-conditional densities are probability density functions p(x| ω) when class is fixed
- Likelihoods are values of p(x| ω) for a given x

- This is a subtle point. Think about it.

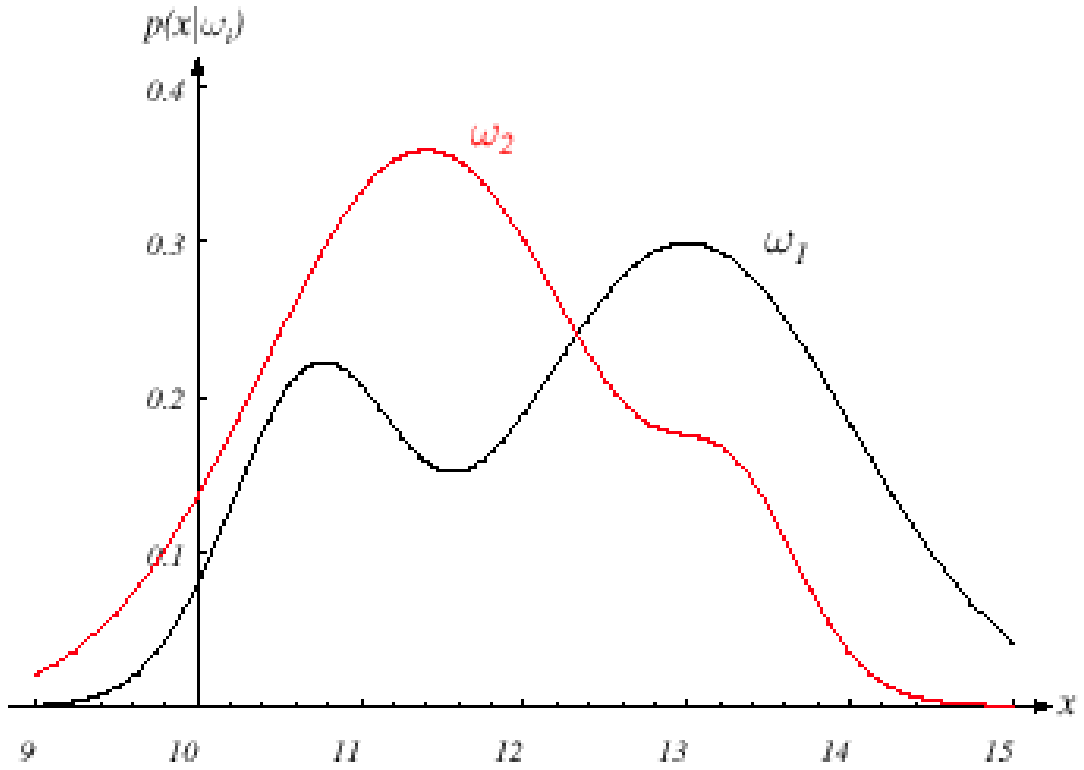


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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.

Posterior, Likelihood, Evidence

$$p(\omega_j \mid \mathbf{x}) = \frac{p(\omega_j)\, p(\mathbf{x} \mid \omega_j)}{p(\mathbf{x})}$$

– In the case of two categories

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

– Posterior = (Likelihood × Prior) / Evidence

Decision using Posteriors

- Decision given the posterior probabilities

  X is an observation for which:

  if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⇒ True state of nature = ω_1
  if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⇒ True state of nature = ω_2

  Therefore:

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

  $P(\text{error} \mid x) = P(\omega_1 \mid x)$ if we decide ω_2
  $P(\text{error} \mid x) = P(\omega_2 \mid x)$ if we decide ω_1

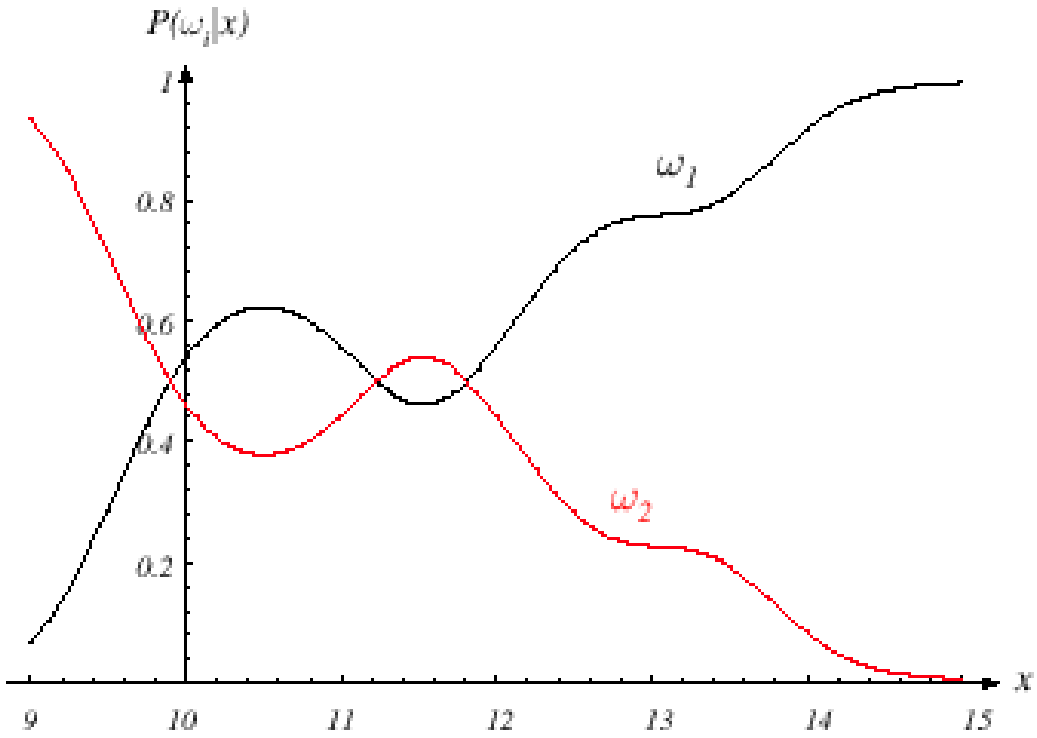


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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.

Probability of Error

- Minimizing the probability of error

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

Therefore:

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

(Bayes decision)

Decision Theoretic Classification

$\omega \in \Omega$: unknown class or category, finite set of options

$x \in \mathcal{X}$: observed data, can take values in any space

$a \in A$: action to chose one of the categories (or possibly to reject data)

$L(\omega,a)$: loss of action a given true class ω

Loss Function

- The loss function states how costly each action taken is
  - Opposite of Utility function: L = - U
- Most common choice is the 0-1 loss

$$L(y, a) = \mathbb{I}(y \neq a) = \begin{cases} 0 & \text{if } a = y \\ 1 & \text{if } a \neq y \end{cases}$$

- In regression, square loss is the most common choice

$L(y^{true}, y^{pred}) = (y^{true} - y^{pred})^2$

More General Loss Function

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

Notation

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

Overall Risk

R = Sum of all $R(\alpha_i \mid x)$ for i = 1,…,a

Conditional risk

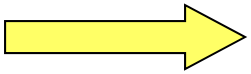
Minimizing R ⟺ Minimizing $R(\alpha_i \mid x)$ for i = 1,…, a
(select action α that minimizes risk as a function of x)

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

for i = 1,…,a

Minimize Overall Risk

Select the action α_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

Conditional Risk

- Two-category classification

α_1 : decide ω_1

α_2 : decide ω_2

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

loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

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

Decision Rule

Our rule is the following:

$$\text{if } R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$$

$$\text{action } \alpha_1\text{: decide } \omega_1$$

This results in the equivalent rule :

decide ω_1 if:

$(\lambda_{21}- \lambda_{11})\, P(x \mid \omega_1)\, P(\omega_1) > (\lambda_{12}- \lambda_{22})\, P(x \mid \omega_2)\, P(\omega_2)$
and decide ω_2 otherwise

Likelihood ratio

The preceding rule is equivalent to the following rule:

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

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

Optimal decision property

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

Exercise

Select the optimal decision where:

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

$P(x \mid \omega_1)$ ⟹ $N(2, 0.5)$ (Normal distribution)

$P(x \mid \omega_2)$ ⟹ $N(1.5, 0.2)$

$P(\omega_1) = 2/3$

$P(\omega_2) = 1/3$

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

Minimum-Error-Rate Classification

- Actions are decisions on classes

  If action α_i is taken and the true state of nature is ω_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 called the error rate

The Zero-one Loss Function

- Zero-one loss function:

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

Therefore, the conditional risk is:

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

$$= \sum_{j \neq i} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$

- The risk corresponding to this loss function is the average probability of error

Minimum Error Rate Decision Rule

- Minimizing the risk requires maximizing $P(\omega_i \mid x)$
  since $R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$

- For Minimum error rate

  – Decide ω_i if $P(\omega_i \mid x) > P(\omega_j \mid x)\ \forall j \neq i$

- Given the likelihood ratio and the zero-one loss function:

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

- If λ is the zero-one loss function which means:

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

$$\textit{then } \theta_\lambda=\frac{P(\omega_2)}{P(\omega_1)}=\theta_a$$

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

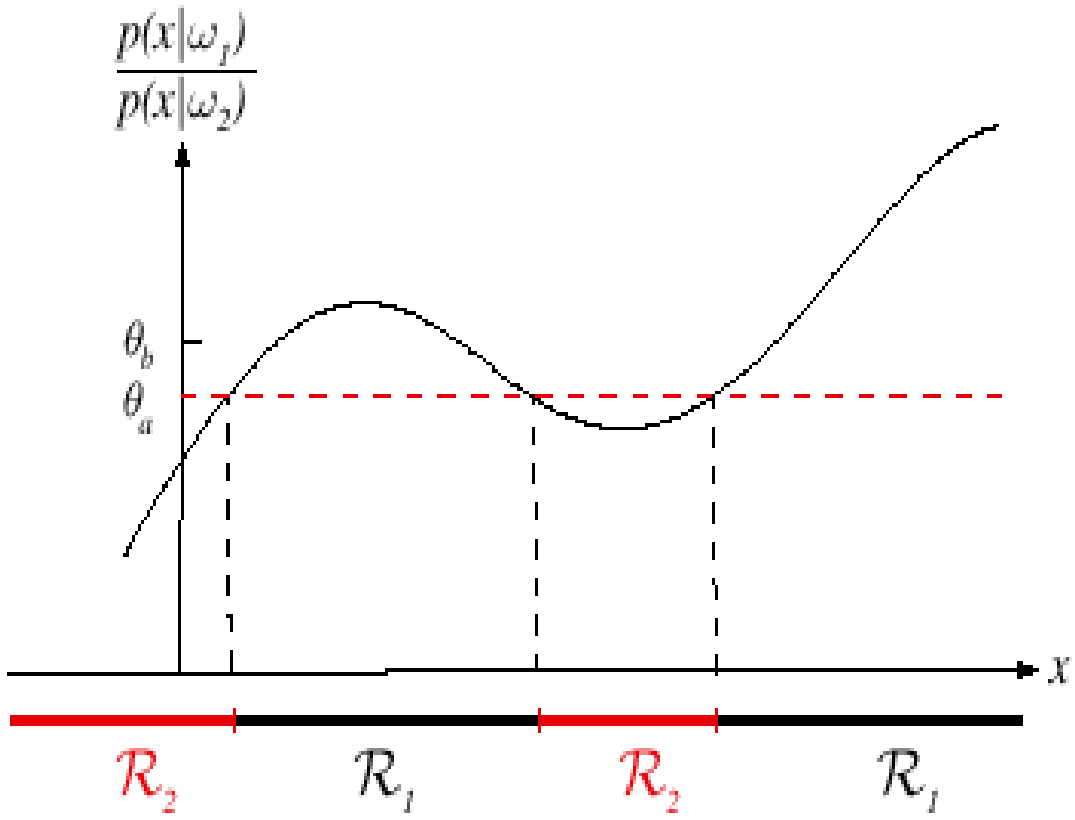


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Classifiers, Discriminant Functions and Decision Surfaces

- The multi-category case

  – Set of discriminant functions $g_i(x)$, i = 1,…, c

  – The classifier assigns a feature vector x to class ω_i if:

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

Max Discriminant Functions

- Let $g_i(x) = - R(\alpha_i \mid x)$
  (max. discriminant corresponds to min. risk)

- For the minimum error rate, we take

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

(max. discriminant corresponds to max. posterior)

$$g_i(x) \equiv P(x \mid \omega_i)\, P(\omega_i)$$

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

(ln: natural logarithm)

Decision Regions

- Feature space divided into c decision regions

  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 ω_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 ω_1 if $g(x) > 0$; Otherwise decide ω_2

Computation of g(x)

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

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

Discriminant Functions for the Normal Density

- 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

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

- Case $\Sigma_i = \sigma^2 \mathbf{I}$ ($\mathbf{I}$ is 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)

Prove it!

Linear Machines

– 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:

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

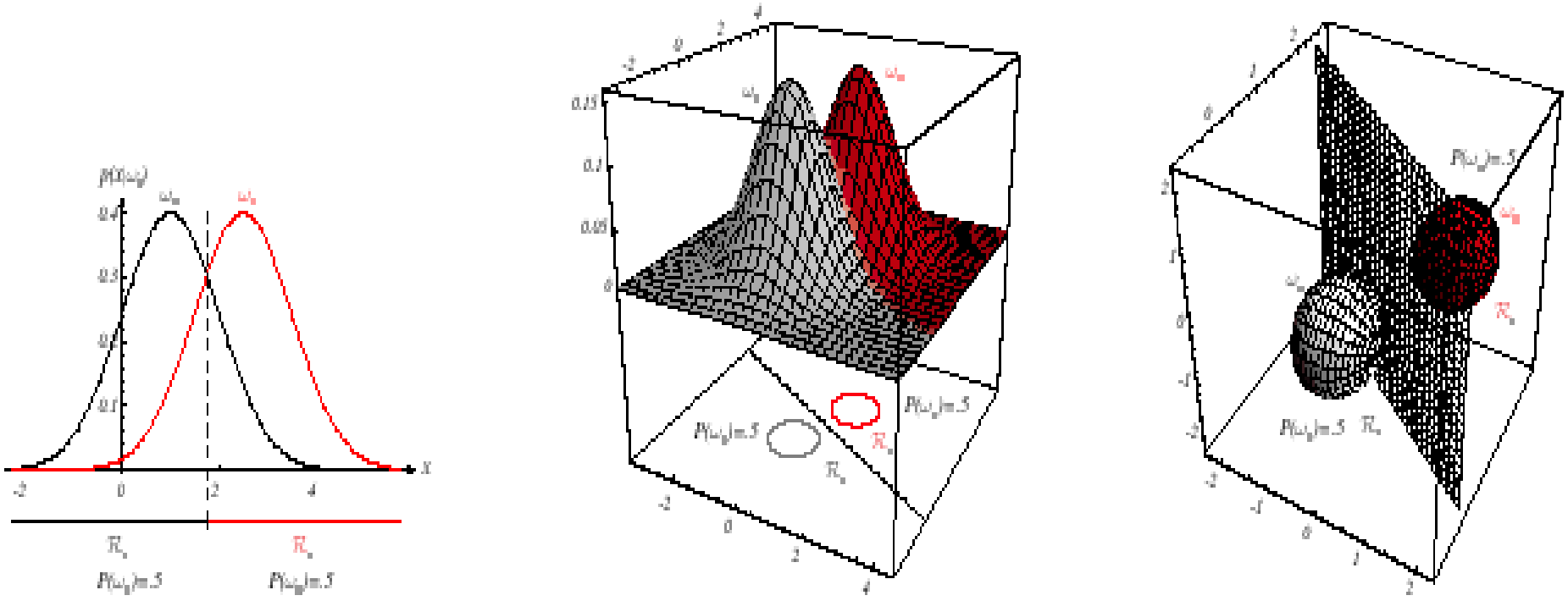


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, 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 $\mathcal{R}_j$

$$g_i(x) = w_i^t x + w_{i0} \quad \text{and} \quad g_j(x) = w_j^t x + w_{j0}$$

$$\text{Decision boundary}: g_i(x) = g_j(x)$$

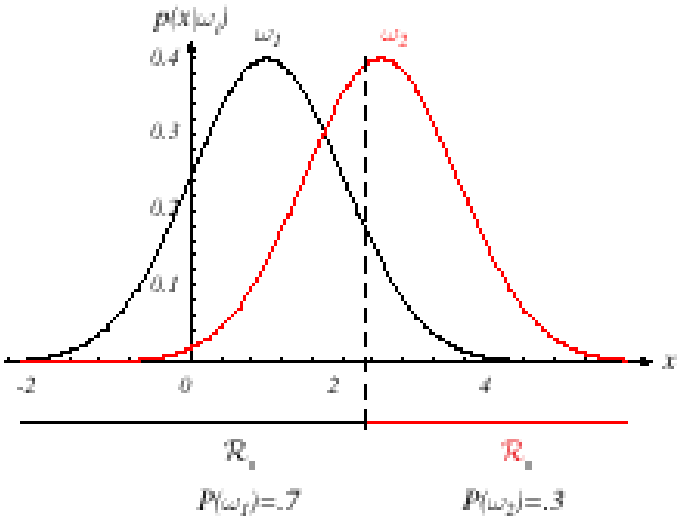
$$w^t(x - x_0) = 0$$

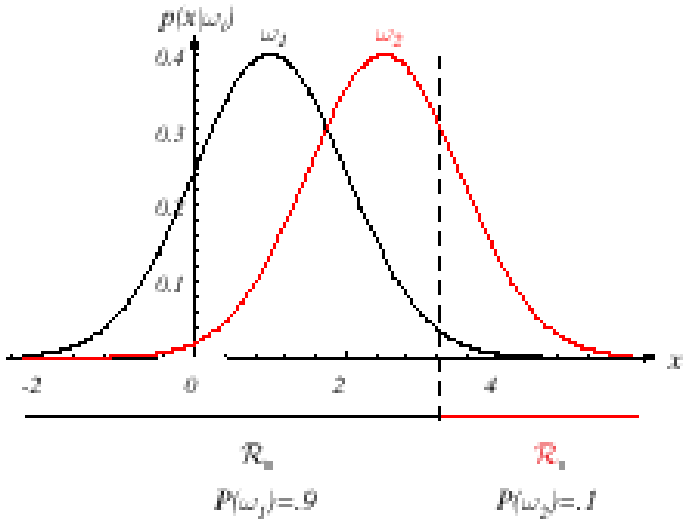
$$w = \mu_i - \mu_j$$

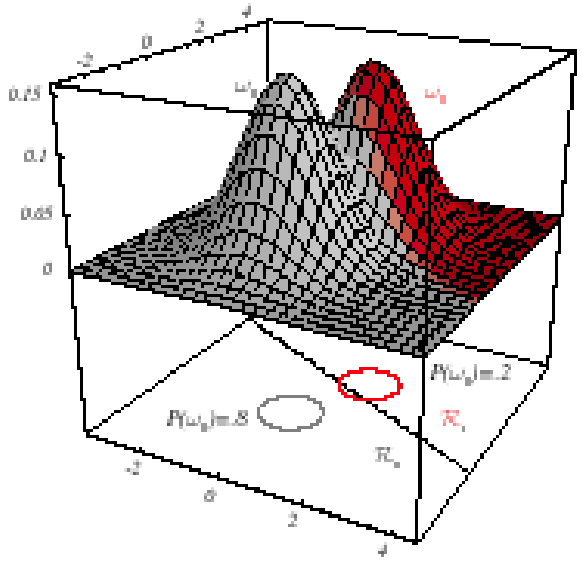
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\left\|\mu_i - \mu_j\right\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

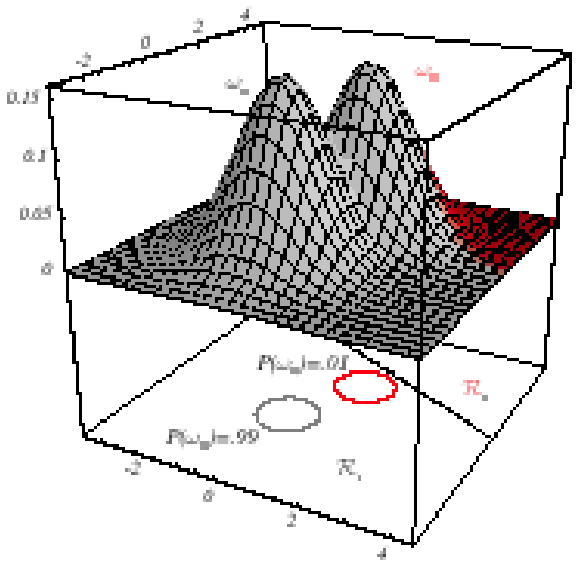
always orthogonal to the line linking the means

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


p(x|ω_j)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω1)=.7
P(ω2)=.3


p(x|ω_j)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
P(ω1)=.9
P(ω2)=.1


0.15
0.1
0.05
0
P(ω2)=.2
P(ω1)=.8


0.15
0.1
0.05
0
P(ω2)=.01
P(ω1)=.99

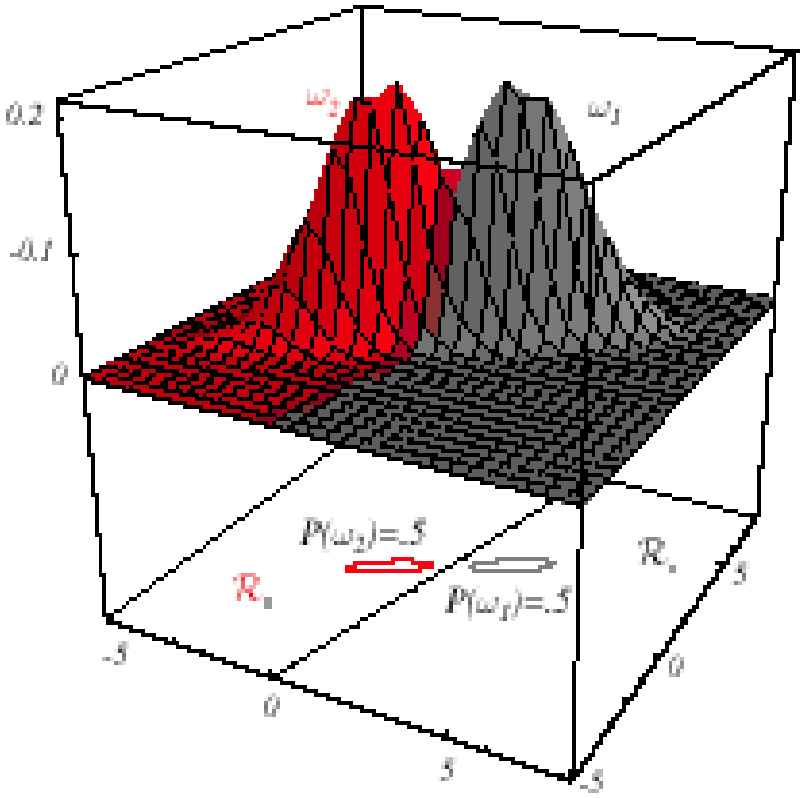
- Case $\Sigma_i = \Sigma$ (covariances of all classes are identical but arbitrary!)

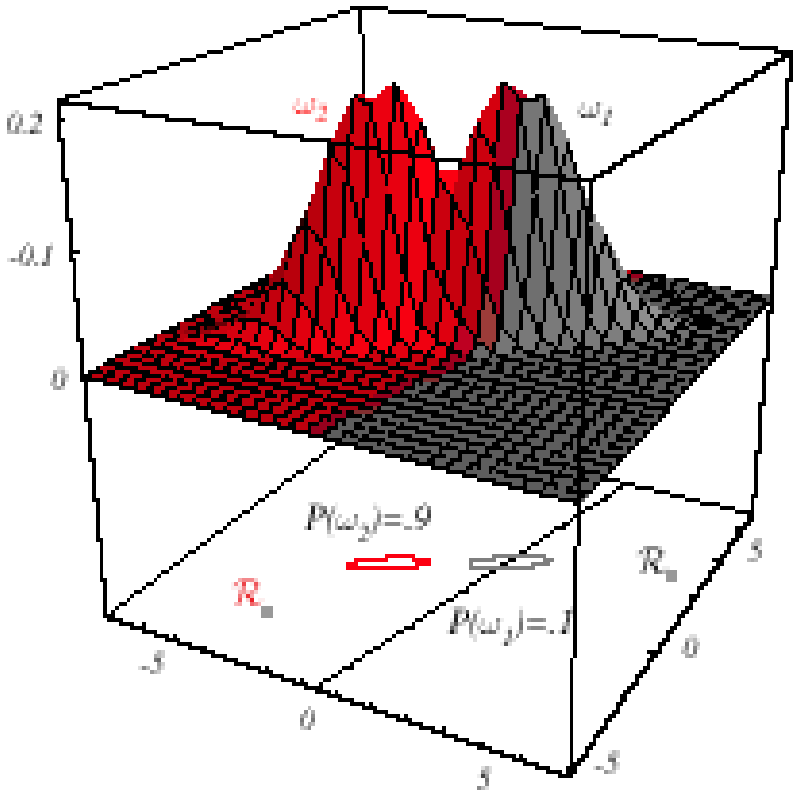
  – Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

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

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

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


0.2
-0.1
0
ω_2
ω_1
$P(\omega_2)=.5$
$P(\omega_1)=.5$
$\mathcal{R}_2$
$\mathcal{R}_1$
-5
0
5
5
0
-5


0.2
-0.1
0
ω_2
ω_1
$P(\omega_2)=.9$
$P(\omega_1)=.1$
$\mathcal{R}_2$
$\mathcal{R}_1$
-5
0
5
5
0
-5

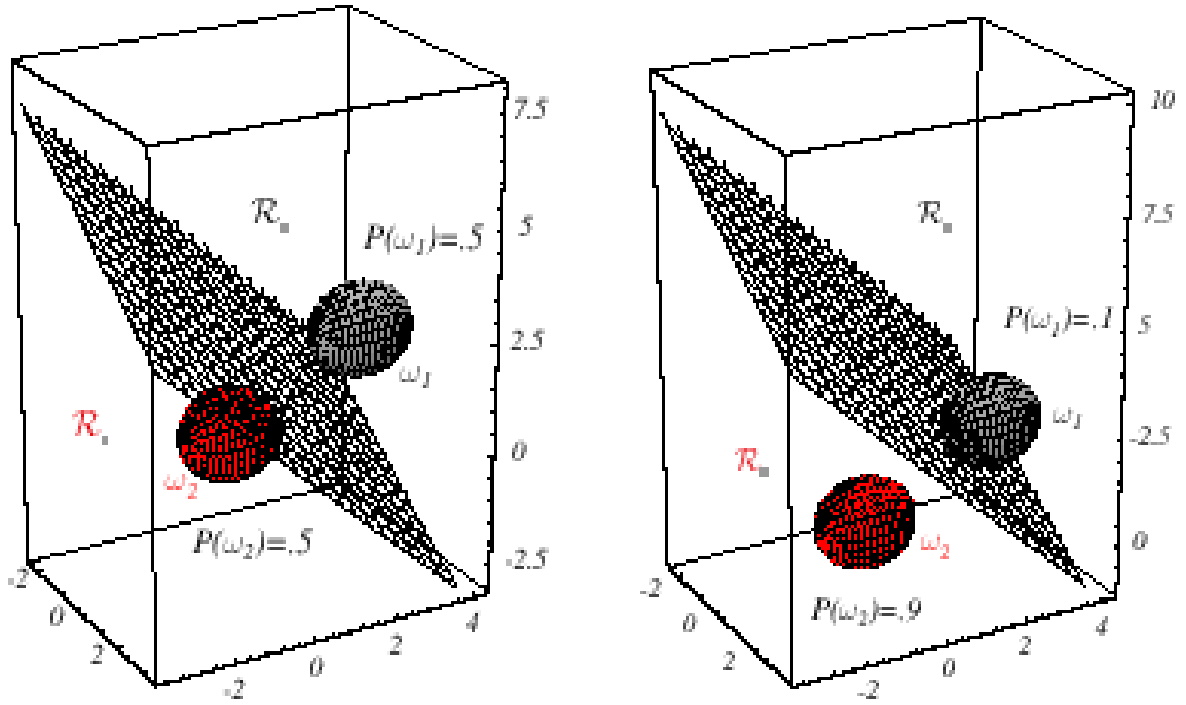


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

- Case Σ_i = arbitrary
  - The covariance matrices are different for each category

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

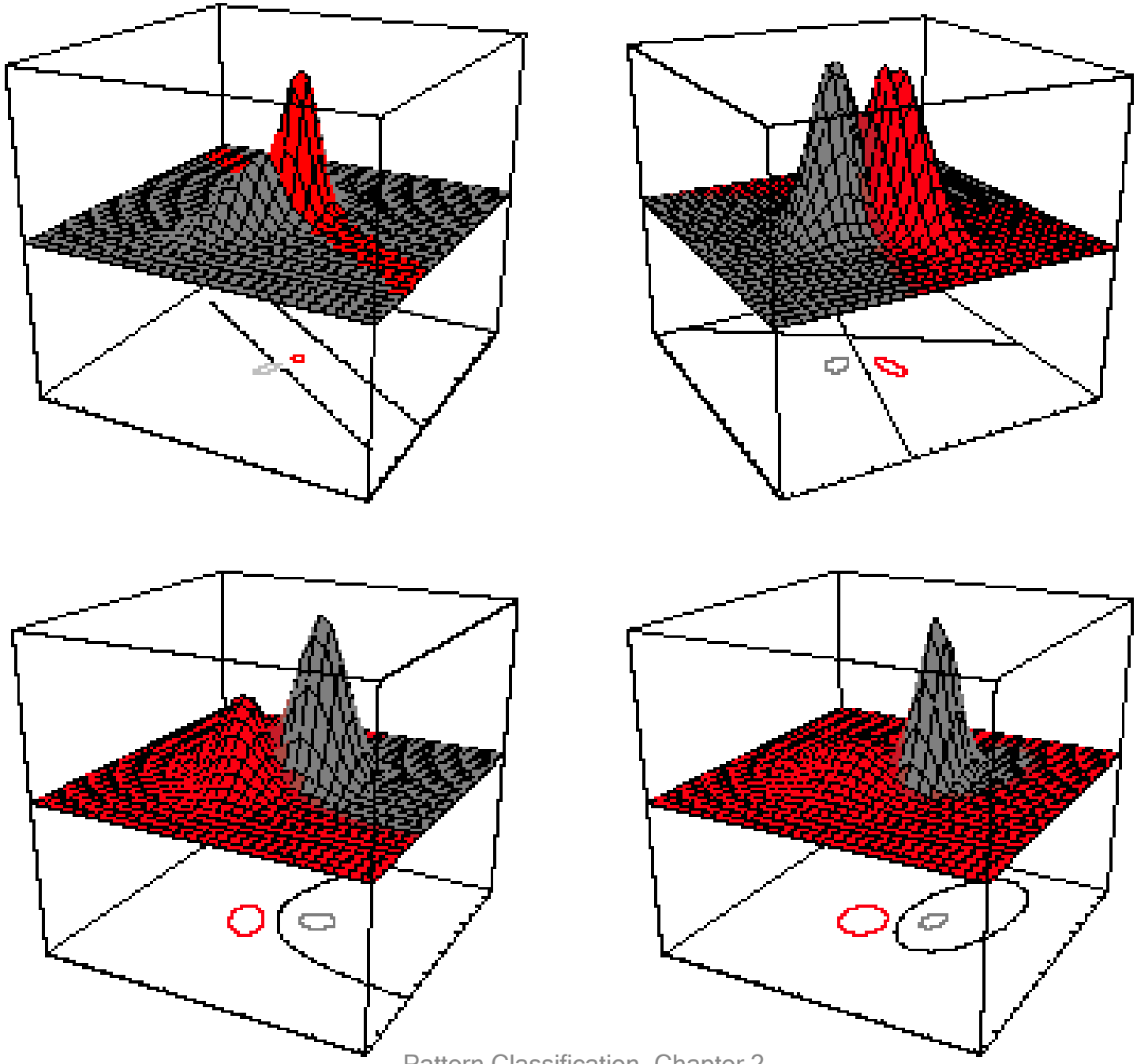
where :

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

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

$$\mathrm{w}_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

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



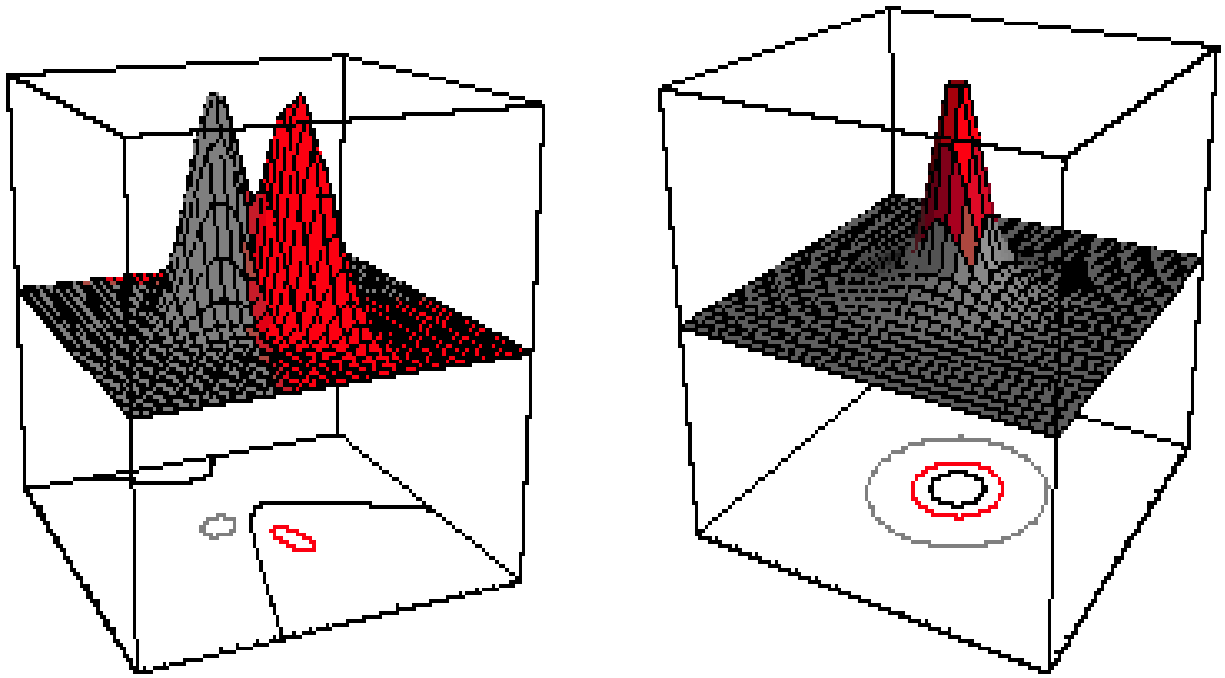


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Bayes Decision Theory - Discrete Features

- Components of x are binary or integer valued, x can take only one of m discrete values

$$v_1, v_2, \ldots, v_m$$

- Case of independent binary features in 2 category problem
- Let $x = [x_1, x_2, \ldots, x_d]^t$ where each x_i is either 0 or 1, with probabilities:

$$p_i = P(x_i = 1 \mid \omega_1)$$
$$q_i = P(x_i = 1 \mid \omega_2)$$

$$P(x \mid \omega_1) = \prod_{i=1}^{d} p_i^{x_i} (1 - p_i)^{1-x_i}$$

$$P(x \mid \omega_2) = \prod_{i=1}^{d} q_i^{x_i} (1 - q_i)^{1-x_i}$$

Likelihood ratio :

$$\frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1 - p_i}{1 - q_i}\right)^{1-x_i}$$

Discriminant function :

$$\mathrm{g(x)} = \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1 - p_i}{1 - q_i} \right] + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

where :

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

and :

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

decide ω_1 if $g(x) > 0$ and ω_2 if $g(x) \leq 0$

Exercise: DHS Problem 2.12

Let $\omega_{max}(x)$ be the state of nature for which $P(\omega_{max}|x) >= P(\omega_i|x)$ for all i=1,…,c

- Show that $P(\omega_{max}|x) >= 1/c$
- Show that for the minimum-error-rate decision rule the average probability of error is given by

$$P(error) = 1 - \int P(\omega_{\max} \mid x)\, p(x)\, dx$$

- Use these two results to show that $P(error) \leq (c-1)/c$
- Describe a situation for which $P(error) = (c-1)/c$

Discriminant Functions for the Normal Density

- Case $\Sigma_i = \sigma^2 \boldsymbol{I}$

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

where:

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

Discriminant Function Example

- 3 classes, each 2-dimensional Gaussian with

$$\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 4 \\ 6 \end{bmatrix} \quad \mu_3 = \begin{bmatrix} -2 \\ 4 \end{bmatrix} \quad \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

- Priors $P(c_1) = P(c_2) = \frac{1}{4}$ and $P(c_3) = \frac{1}{2}$
- Discriminant function is $g_i(x) = \frac{\mu_i^t}{\sigma^2} x + \left(-\frac{\mu_i^t \mu_i}{2\sigma^2} + \ln P(c_i) \right)$
- Plug in parameters for each class

$$g_1(x) = \frac{[1\ 2]}{3} x + \left(\frac{5}{6} - 1.38 \right) \quad g_2(x) = \frac{[4\ 6]}{3} x + \left(\frac{52}{6} - 1.38 \right)$$

$$g_3(x) = \frac{[-2\ 4]}{3} x + \left(\frac{20}{6} - 0.69 \right)$$

Discriminant Function Example

- Need to find out when $g_i(x) < g_j(x)$ for $i,j=1,2,3$
- Can be done by solving $g_i(x) = g_j(x)$ for $i,j=1,2,3$
- Let's take $g_1(x) = g_2(x)$ first

$$\frac{[1\ 2]}{3}x+\left(-\frac{5}{6}-1.38\right)=\frac{[4\ 6]}{3}x+\left(-\frac{52}{6}-1.38\right)$$

- Simplifying, $$\frac{[-3\ -4]}{3}\begin{bmatrix}x_1\\x_2\end{bmatrix}=-\frac{47}{6}$$

$$-x_1-\frac{4}{3}x_2=-\frac{47}{6}$$

line equation

Discriminant Function Example

- Next solve $g_2(x) = g_3(x)$

$$2x_1 + \frac{2}{3}x_2 = 6.02$$

- Almost finally solve $g_1(x) = g_3(x)$

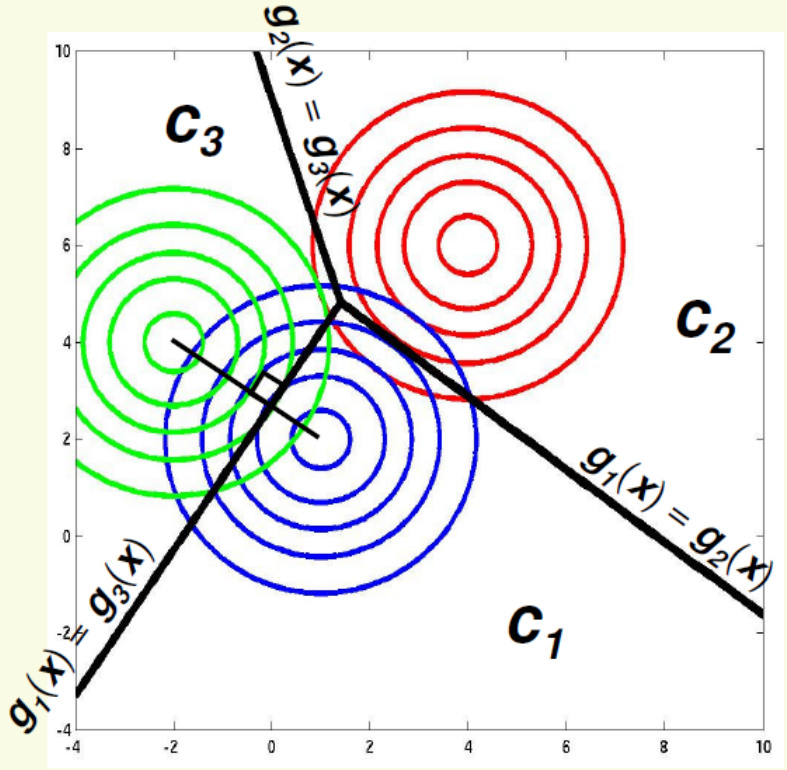
$$x_1 - \frac{2}{3}x_2 = -1.81$$

- And finally solve $g_1(x) = g_2(x) = g_3(x)$

$$x_1 = 1.4 \quad \text{and} \quad x_2 = 4.82$$

Discriminant Function Example

- Priors $P(c_1) = P(c_2) = \frac{1}{4}$ and $P(c_3) = \frac{1}{2}$



lines connecting means are perpendicular to decision boundaries

Maximum-Likelihood & Bayesian Parameter Estimation

Adapted from:
Duda, Hart and Stork, Pattern Classification textbook

O. Veksler

E. Sudderth
D. Batra

Introduction

- We could design an optimal classifier if we knew:
  - $p(\omega_i)$ (priors)
  - $p(x \mid \omega_i)$ (class-conditional densities)
  - Unfortunately, we rarely have this complete information!

- Design a classifier from training data

Supervised Learning in a Nutshell

- Training Stage:
  - Raw Data → x (Feature Extraction)
  - Training Data { (x,y) } → f (Learning)
- Testing Stage
  - Raw Data → x (Feature Extraction)
  - Test Data x → f(x) (Apply function, Evaluate error)

Statistical Estimation View

- Probabilities to the rescue:
  - x and y are *random variables*
  - $D = (x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N) \sim P(X,Y)$
- IID: Independent Identically Distributed
  - Both training & testing data sampled IID from P(X,Y)
  - Learn on training set
  - Have some hope of *generalizing* to test set

Parameter Estimation

- Use a priori information about the problem
- E.g.: Normality of $p(x \mid \omega_i)$

$$p(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

- Simplify problem
  - From estimating unknown distribution function
  - To estimating parameters

Why Gaussians?

- Why does the entire world seem to always be harping on about Gaussians?
  - Central Limit Theorem!
  - They're easy (and we like easy)
  - Closely related to squared loss (for regression)
  - Mixture of Gaussians is sufficient to approximate many distributions

Some properties of Gaussians

- Affine transformation
  - multiplying by scalar and adding a constant
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \quad \rightarrow Y \sim N(a\mu+b,a^2\sigma^2)$

- Sum of Independent Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \quad \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

Estimation techniques

- Maximum-Likelihood (ML) and Bayesian estimation
- Results are often identical, but the approaches are fundamentally different

- Frequentist View
  - limit N→∞ #(A is true)/N
  - limiting frequency of a repeating non-deterministic event

- Bayesian View
  - P(A) is your “belief” about A

Parameter Estimation

- Parameters in ML estimation are fixed but unknown!
- Best parameters are obtained by maximizing the probability of obtaining the samples observed
- Bayesian methods view the parameters as random variables having some known distribution
- In either approach, we use $p(\omega_i \mid x)$ for our classification rule

Independence Across Classes

- For each class ω_i we have a proposed density $p_i(x|\ \omega_i)$ with unknown parameters θ_i which we need to estimate
- Since we assumed independence of data across the classes, estimation is an identical procedure for all classes
- To simplify notation, we drop sub-indexes and say that we need to estimate parameters θ for density $p(x)$

Maximum-Likelihood Estimation

- Has good convergence properties as the sample size increases
- Simpler than alternative techniques

- General principle
  - Assume **c** datasets (classes) D_1, D_2, …, D_c drawn independently according to p(x| ω_j)

Maximum-Likelihood Estimation

- Assume that p(x| ω_j) has known parametric form determined by parameter vector θ_j
- Further assume that D_i gives no information about θ_j if $i \neq j$
  - Drop subscripts in remainder

Likelihood

- Use set of independent samples to estimate p(D | θ)
  - Let D = {x_1, x_2, …, x_n}
  - p(x_1,…, x_n | θ) = $\prod_{i=1}^{n}$ p(xi | θ); |D| = n

    Our goal is to determine $\hat{\theta}$ (value of θ that best agrees with observed training data)

- Note if D is fixed p(D| θ) is not a density

Example: Gaussian case

- Assume we have c classes and

  $p(x \mid \omega_j) \sim N(\mu_j, \Sigma_j)$

  $p(x \mid \omega_j) \equiv p(x \mid \omega_j, \theta_j)$ where:

$$\theta = (\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2, ..., \sigma_j^{11}, \sigma_j^{22}, \mathrm{cov}(x_j^m, x_j^n)...)$$

- Use the information provided by the training samples to estimate

  $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$, each θ_i (i = 1, 2, …, c) is associated with each category

- Suppose that D contains n samples, $x_1, x_2, \ldots, x_n$

$$p(D \mid \theta) = \prod_{k=1}^{k=n} p(x_k \mid \theta)$$

- $p(D \mid \theta)$ is called the likelihood of θ w.r.t the set of samples
- ML estimate of θ is, by definition the value $\hat{\theta}$ that maximizes $p(D \mid \theta)$

  "It is the value of θ that best agrees with the actually observed training sample"

- Optimal estimation
  - Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let ∇_θ be the gradient operator

$$\nabla_\theta = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_p} \right]^t$$

  - We define $l(\theta)$ as the log-likelihood function

$$l(\theta) = \ln p(D \mid \theta)$$

  - New problem statement:
    determine θ that maximizes the log-likelihood

$$\hat{\theta} = \arg \max_{\theta} l(\theta)$$

Necessary conditions for an optimum:

$$\nabla_\theta l = \sum_{k=1}^{k=n} \nabla_\theta \ln p(x_k \mid \theta)$$

$$\nabla_\theta l = 0$$

- Local or global maximum
- Local or global minimum
- Saddle point
- Boundary of parameter space

Example of ML estimation: unknown μ

– $p(x_i \mid \mu) \sim N(\mu, \Sigma)$
(Samples are drawn from a multivariate normal population)

$$\ln p(x_k \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d |\Sigma|\right] - \frac{1}{2}(x_k - \mu)^t \Sigma^{-1}(x_k - \mu)$$

$$\nabla_\theta \ln p(x_k \mid \mu) = \Sigma^{-1}(x_k - \mu)$$

$\theta = \mu$ therefore:

- The ML estimate for μ must satisfy:

$$\sum_{k=1}^{k=n} \Sigma^{-1}(x_k - \hat{\mu}) = 0$$

- Multiplying by Σ and rearranging, we obtain:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{k=n} x_k$$

Just the arithmetic average of the samples of the training samples!

Conclusion:

If $p(x_k \mid \omega_j)$ (j = 1, 2, …, c) is assumed to be Gaussian in a *d*-dimensional feature space, then we can estimate the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_c)^t$ and perform optimal classification!

- Example of ML estimation: unknown μ *and* σ (univariate)

$$\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$$

$$l = \ln\ p(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \begin{pmatrix} \frac{\partial}{\partial \theta_1}(\ln\ p(x_k \mid \theta)) \\ \frac{\partial}{\partial \theta_2}(\ln\ p(x_k \mid \theta)) \end{pmatrix} = 0$$

$$\begin{cases} \frac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} = 0 \end{cases}$$

Summation:

$$\begin{cases} \sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2}(x_k - \theta_1) = 0 & (1) \\ -\sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{k=n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

Combining (1) and (2), one obtains:

$$\mu = \sum_{k=1}^{k=n} \frac{x_k}{n} \qquad \sigma^2 = \frac{\sum_{k=1}^{k=n} (x_k - \mu)^2}{n}$$

Bias

- ML estimate for σ^2 is biased

$$E\left[\frac{1}{n}\Sigma(x_i - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

- For one sample, the estimated variance is always zero => under-estimate
- An elementary unbiased estimator for Σ is:

$$\underbrace{C = \frac{1}{n-1}\sum_{k=1}^{k=n}(x_k - \mu)(x_k - \hat{\mu})^t}_{\text{Sample covariance matrix}}$$

- Ultimately, interested in estimate that maximizes classification performance

Model Error

- What if we assume class distribution to be $N(\mu,1)$, but true distribution is $N(\mu,10)$?
  - ML estimate: $\theta=\mu$ is the correct mean
- Will this θ result in best classifier performance?
  - NO