3. Bayesian Decision Theory

Bayesian Decision Theory

- Fundamental statistical approach to the problem of pattern classification

- Assumptions:
    1. Decision problem is posed in probabilistic terms
    2. Ideal case: probability structure underlying the categories is known perfectly

Statistical Decision Theory

- What is a pattern?

- In statistical pattern recognition, a pattern is a d-dimensional feature vector

$$\mathbf{x} = (X_1, X_2, ..., X_d)^T$$

Statistical Decision Theory

- **The sea bass/salmon example**

- **State of nature**
- **Prior**

- **State of nature is a random variable (ω):**

  $\omega = \omega_1$ for sea bass; $\omega = \omega_2$ for salmon.
- **The catch of salmon and sea bass is equiprobable**

  $P(\omega_1) = P(\omega_2)$ (Prior)

  $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

Statistical Decision Theory

- Decision rule with *only the prior information*
- Decide ω_1 if $P(\omega_1) > P(\omega_2)$

  otherwise decide ω_2

- Use of the class-conditional information

- $p(x \mid \omega_1)$ and $p(x \mid \omega_2)$ describe the difference in lightness between populations of sea and salmon

Statistical Decision Theory

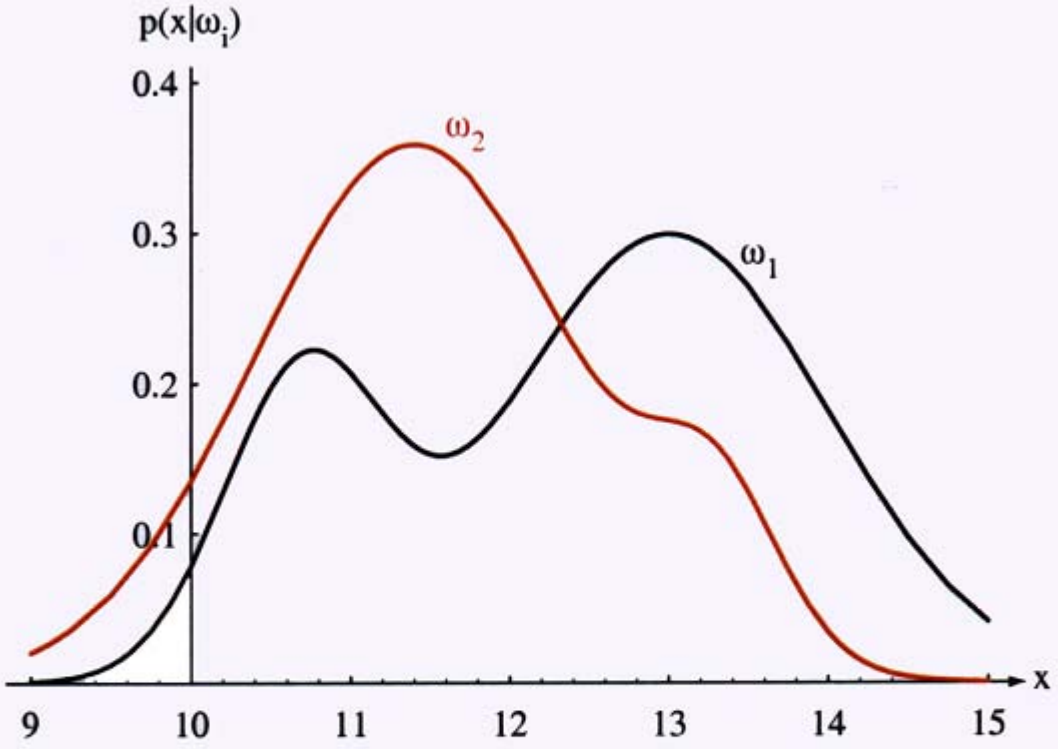


Figure 2.1: Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the length of a fish, the two curves might describe the difference in length of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0.

Statistical Decision Theory

- ***Bayes formula (demo)*:**

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

Where in case of two categories

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

$$posterior = \frac{likelihood \times prior}{evidence}$$

Statistical Decision Theory

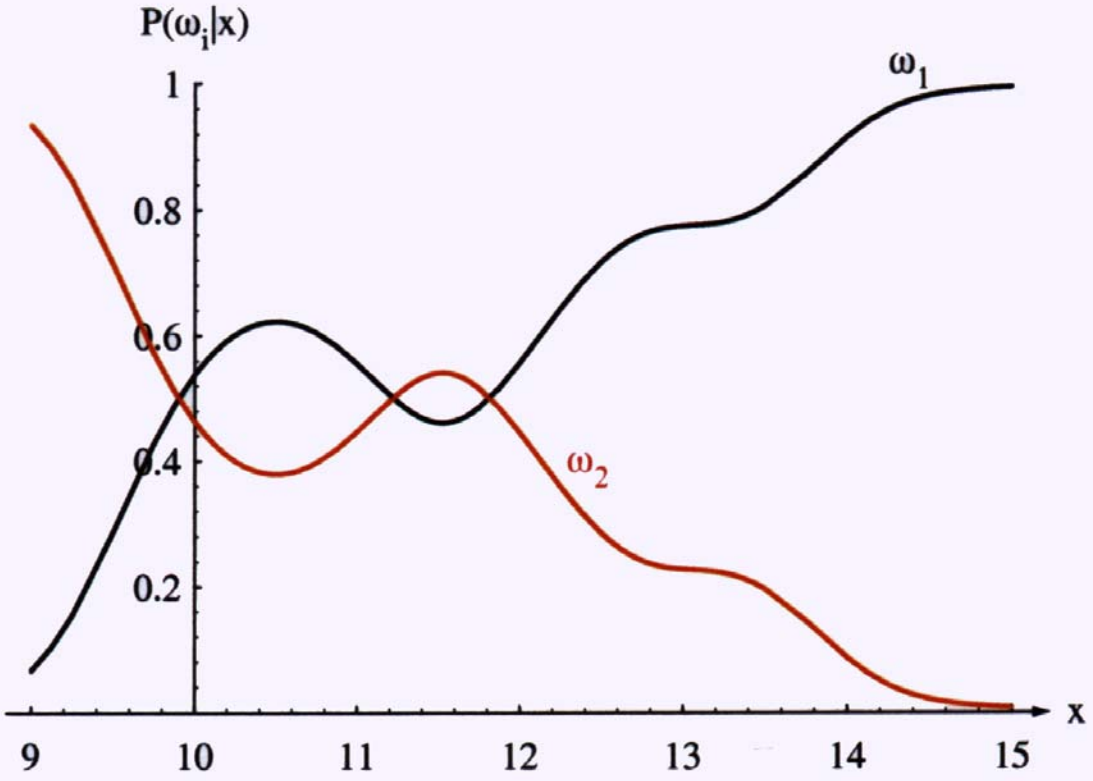


Figure 2.2: Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0.

Statistical Decision Theory

•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

Statistical Decision Theory

- 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 rule)

Probability of Error

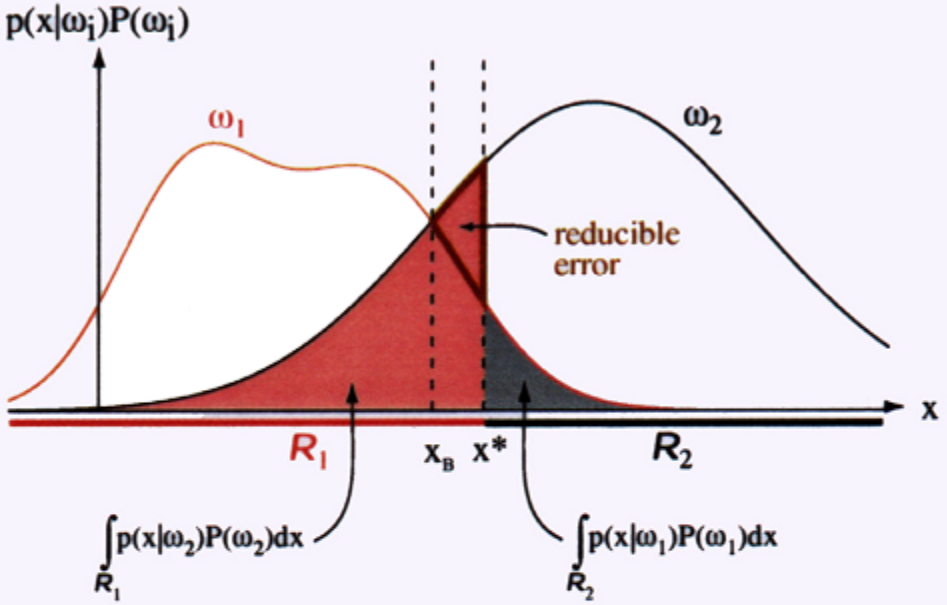


Figure 2.17: Components of the probability of error for equal priors and (non-optimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 68. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible — this is the Bayes decision and gives the Bayes error rate.

Statistical Decision Theory

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 function which is more general than the probability of error

Statistical Decision Theory

- 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

Statistical Decision Theory

- Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature ("categories")

- Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of a 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

Statistical Decision Theory

- **Overall risk (expected loss)**
  R = Sum of all R(α_i | x) for i = 1,…,a

decision function

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

- **Minimizing the conditional risk R(α_i | x) for i = 1,…, a**

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

Minimum Risk Classification

- ***Bayes decision rule:*** **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.**

Two-Category Classification

- **Two-category classification**
  - α_1 : deciding ω_1
  - α_2 : deciding 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:**

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

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

Two-Category Classification

Our rule is the following:

if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$

action α_1: "decide ω_1" is taken

This results in the equivalent rule :

decide ω_1 if:

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

and decide ω_2 otherwise

Two-Category Classification

- ***Likelihood ratio***:

The preceding rule is equivalent to the following rule (assuming that $\lambda_{21} > \lambda_{11}$):

If $$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

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

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"

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 ≠ j

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

Minimum-Error-Rate Classification

Introduction of the zero-one loss function:

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

Therefore, the conditional risk is:

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

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

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

Minimum-Error-Rate Classification

- **Minimizing the risk requires maximizing** $P(\omega_j|\mathbf{x})$ **since** $R(\alpha_j|\mathbf{x}) = 1 - P(\omega_j|\mathbf{x})$

- **For *Minimum error rate***

Decide ω_i if $P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x})$ for all $j \neq i$

Minimum-Error-Rate Classification

- **Investigate the loss function:**

Let $\theta_\lambda = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$ **then decide** ω_1 if $\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \theta_\lambda$

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

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

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

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

Decision Regions: Effect of Loss Functions

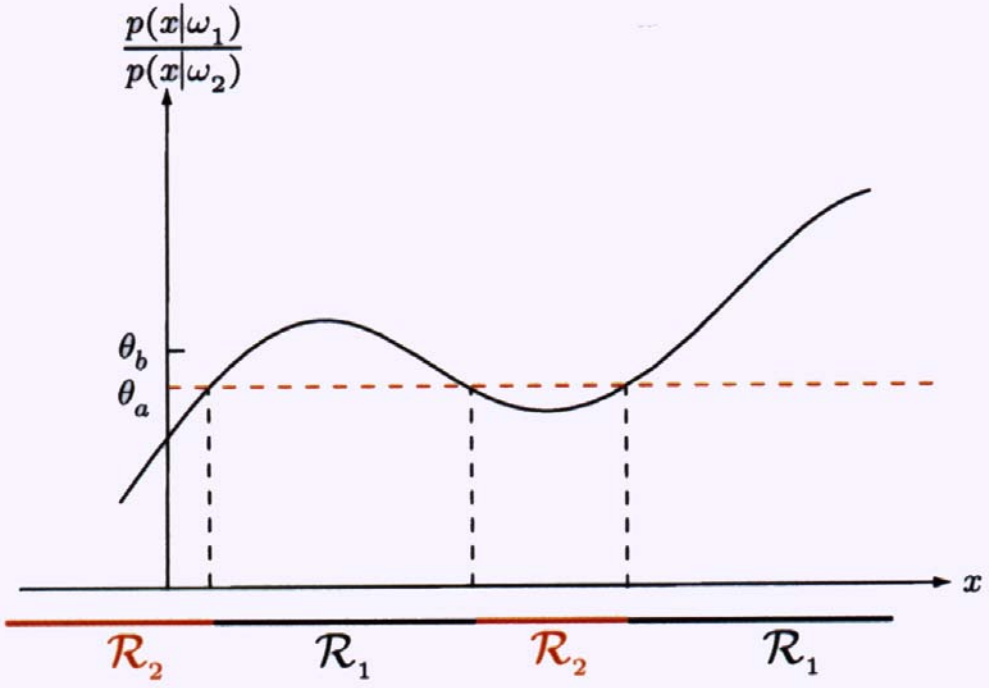


Figure 2.3: The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, (i.e., $\lambda_{12} > \lambda_{21}$), we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller.

Classifiers, Discriminant Functions and Decision Surfaces

THE MULTICATEGORY CASE

Set of *discriminant functions* $g_i(x)$, $i = 1,\ldots, c$

The classifier assigns a feature vector x to class ω_i

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

Classifiers, Discriminant Functions and Decision Surfaces

- 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. discrimination corresponds to max. posterior)

In this case, we can also write:

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

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

General Statistical Classifier

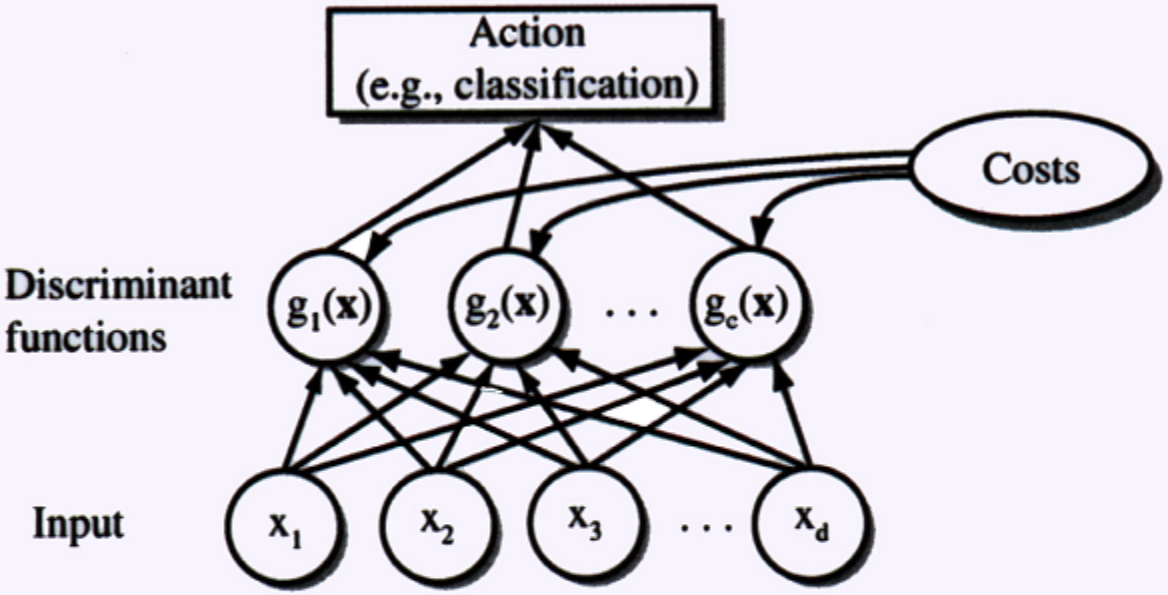


Figure 2.5: The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident.

Classifiers, Discriminant Functions and Decision Surfaces

- **Feature space divided into c *decision regions***

  if $g_i(x) > g_j(x)\ \forall j \neq i$ then x is in R_i

  R_i means assign x to ω_i

- **The regions are separated by *decision boundaries***

The Two-Category Case

- **A classifier is a *dichotomizer* with 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

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

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

Classifiers, Discriminant Functions and Decision Surfaces

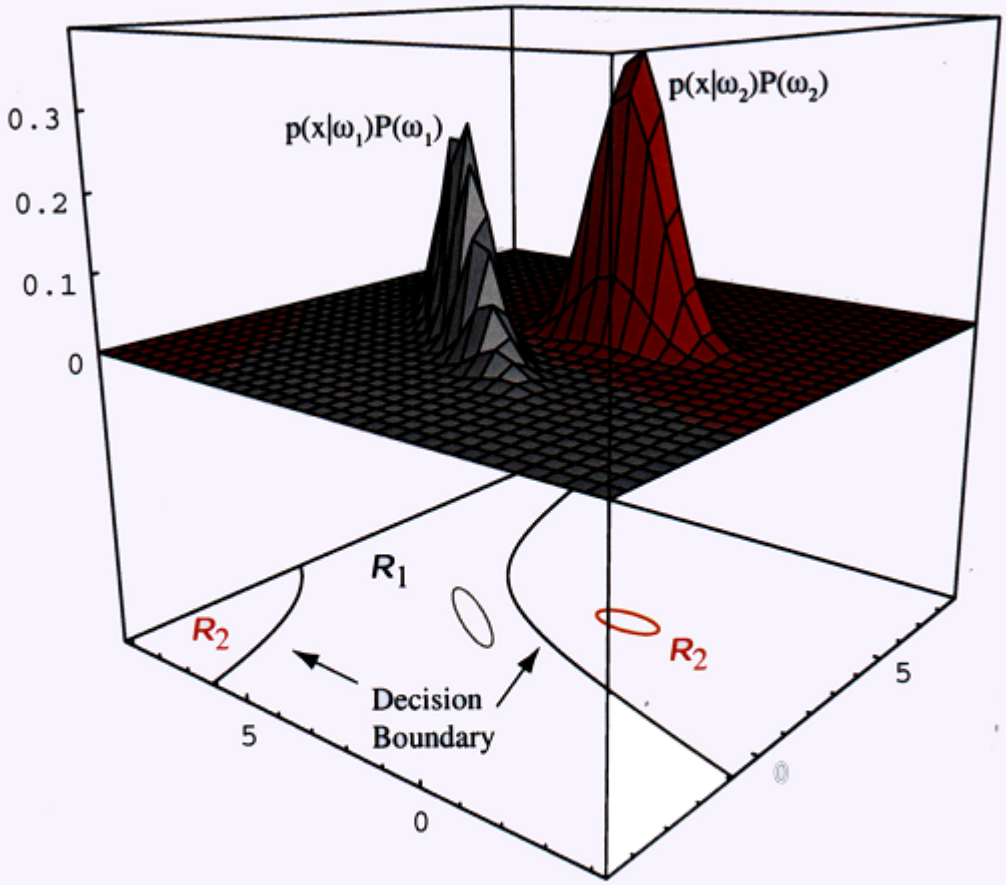


Figure 2.6: In this two-dimensional two-category classifier, the probability densities are Gaussian (with $1/e$ ellipses shown), the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected.

The Normal Density

Univariate density

- **Density which is analytically tractable**
  - **Continuous density**
  - **A lot of processes are asymptotically Gaussian**
  - **Handwritten characters, speech sounds are examples or prototypes 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 **= the variance of x**

The Normal Density

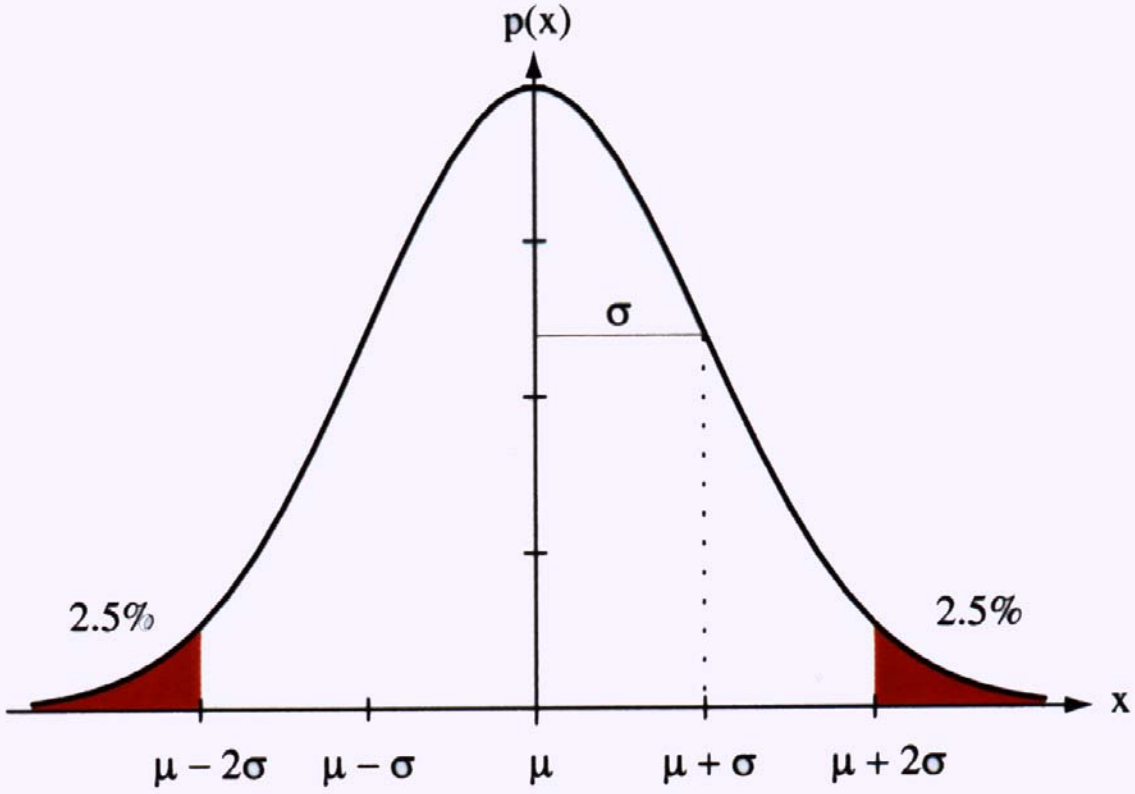


Figure 2.7: A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$.

The Normal Density

- Multivariate density
- Multivariate normal density in d dimensions is:

$$p(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$$

where:

$$\mathbf{x} = (x_1, x_2, ..., x_d)^T$$

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

Σ = d*d covariance matrix

$|\Sigma|$ and Σ^{-1} are the 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

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

- Case of multivariate normal distribution

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Discriminant and Classification for Different Cases

- **Case** $\Sigma_i = \sigma^2 \mathbf{I}$
  - **Features are statistically independent**
  - **Each feature has the same variance** σ^2

$$|\Sigma_i| = \sigma^{2d} \qquad \Sigma_i^{-1} = (1/\sigma^2)\mathbf{I}$$

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

$$\downarrow$$

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

where: $\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \qquad w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + \ln P(\omega_i)$

$\downarrow$

threshold for the category *i*

Discriminant and Classification for Different Cases

- 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:

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

Equal Covariances

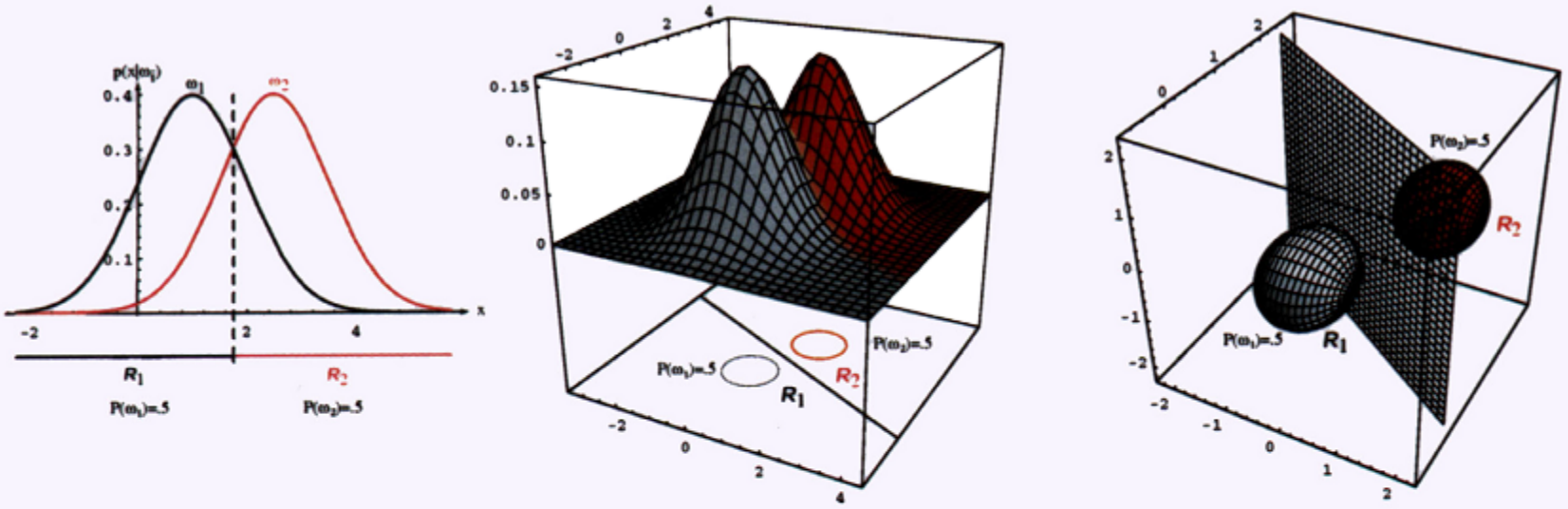


Figure 2.10: If the covariances of two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these 1-, 2-, and 3-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the 3-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$.

Discriminant and Classification for Different Cases

$$\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0)=0 \qquad \mathbf{w}=\boldsymbol{\mu}_i-\boldsymbol{\mu}_j$$

- **The hyperplane separating R_i and R_j**

$$\mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\sigma^2}{\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

always orthogonal to the line linking the means

$$\text{If } P(\omega_i)=P(\omega_j) \text{ then } \mathbf{x}_0=\tfrac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)$$

Shift in Priors

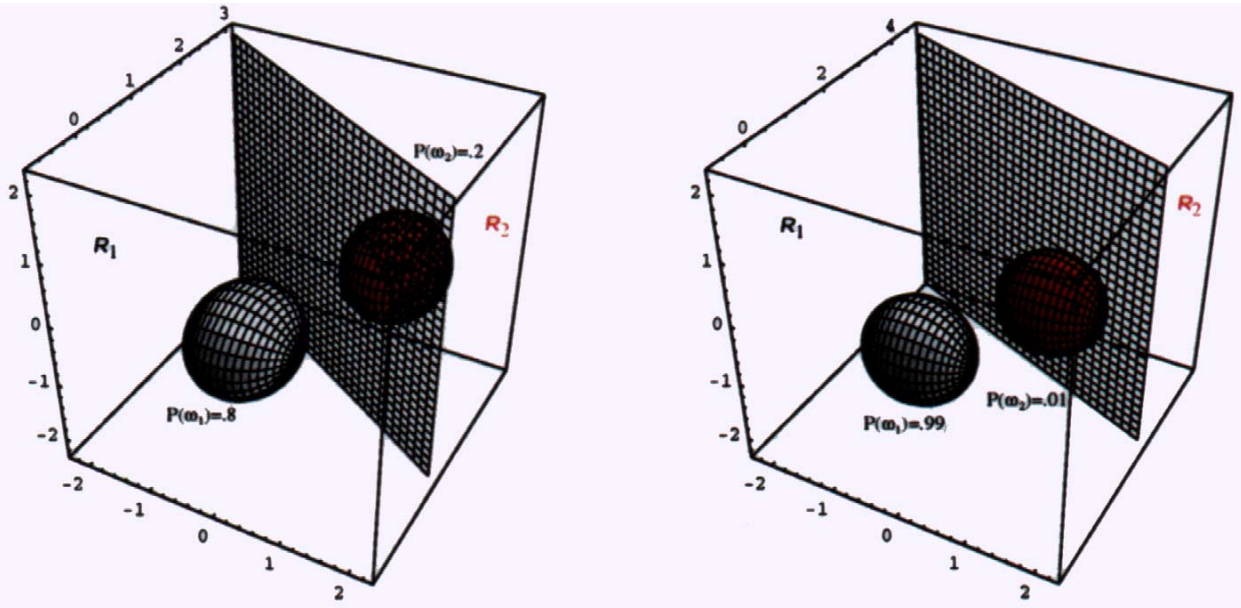
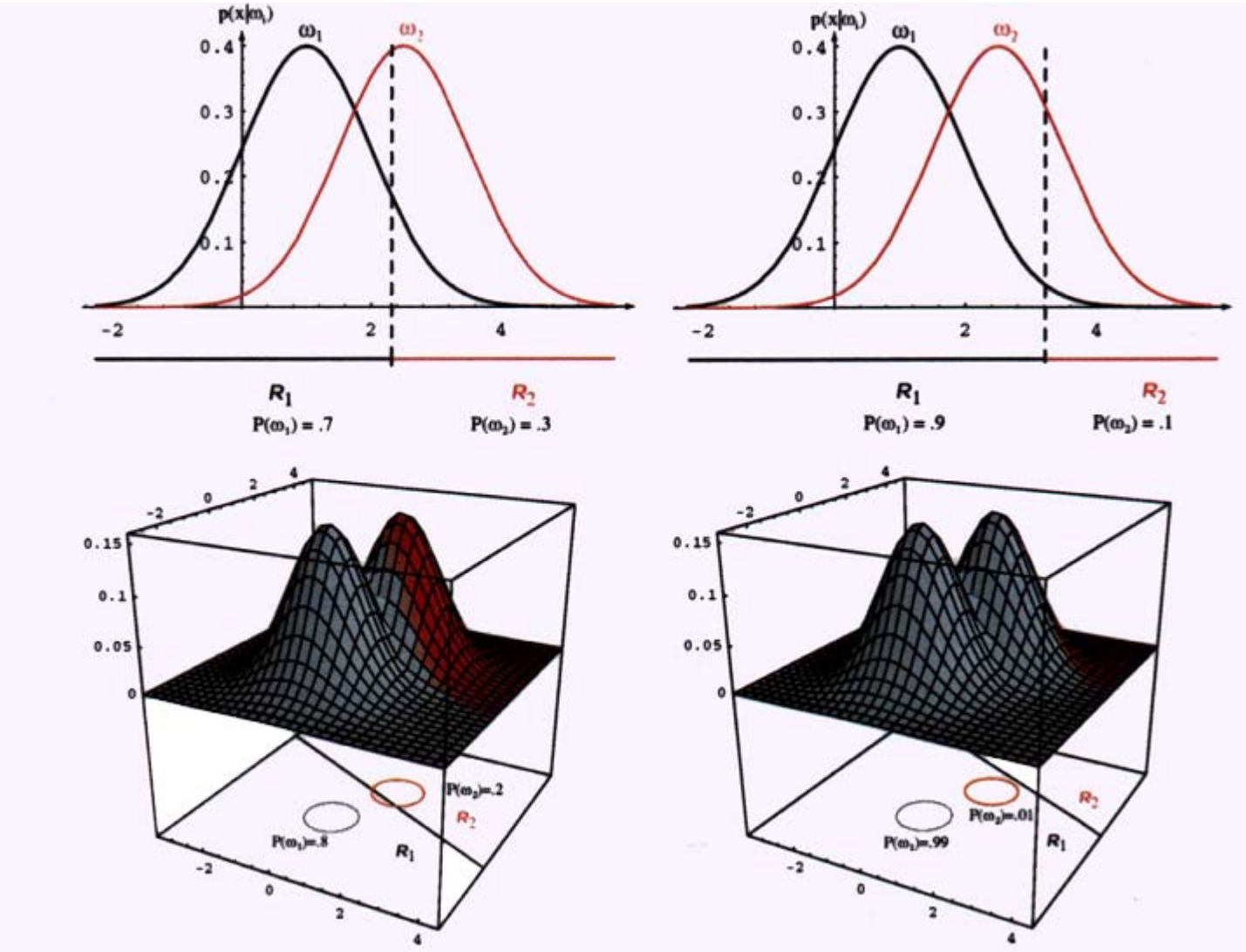


Figure 2.11: As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these 1-, 2- and 3-dimensional spherical Gaussian distributions.

Shift in Priors



Discrimination and Classification for Different Cases

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

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \Sigma^{-1}\boldsymbol{\mu}_i \qquad w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$

Hyperplane separating R_i and R_j

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\Sigma^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

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

Decision Surfaces

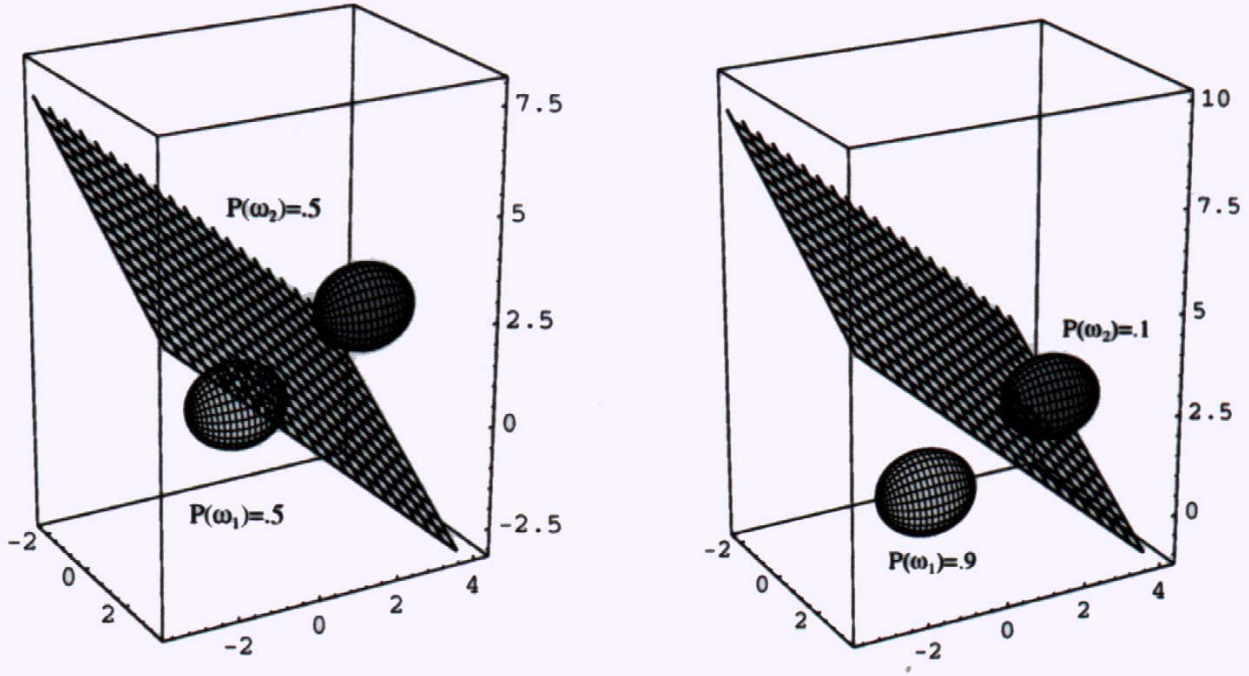
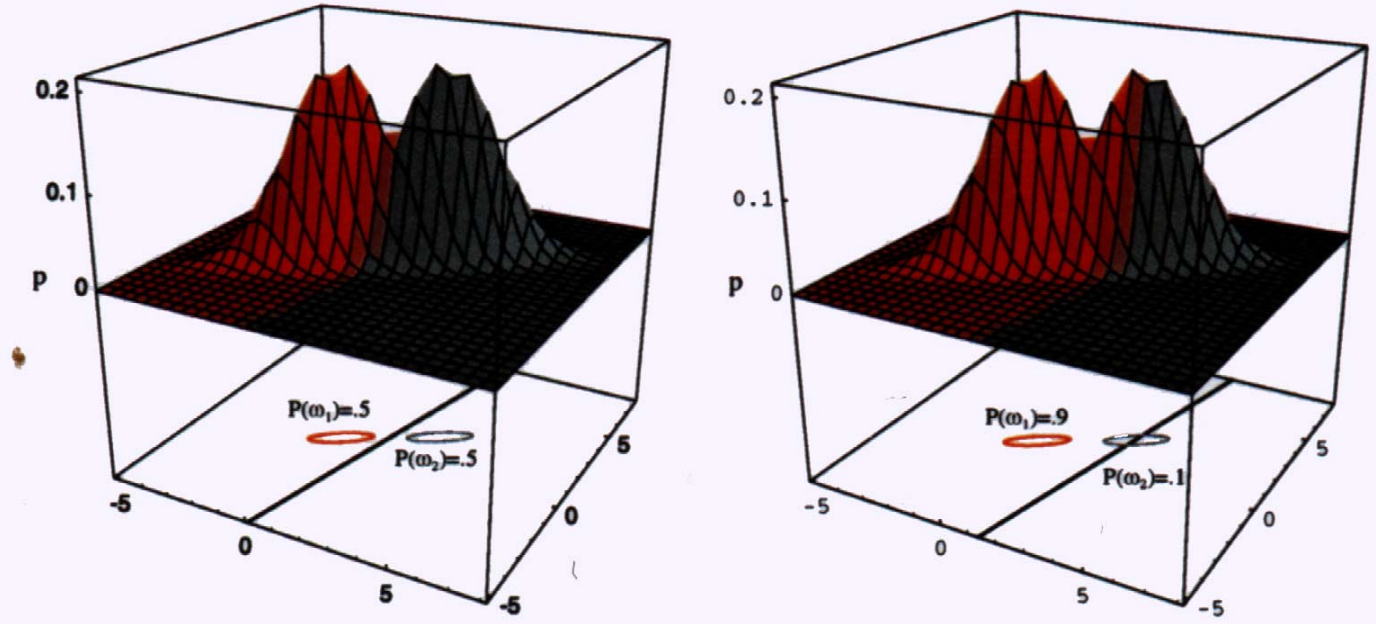


Figure 2.12: Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means.

Decision Surfaces



Discrimination and Classification for Different Cases

Case Σ_i = arbitrary

- **The covariance matrices are different for each category**

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

where:

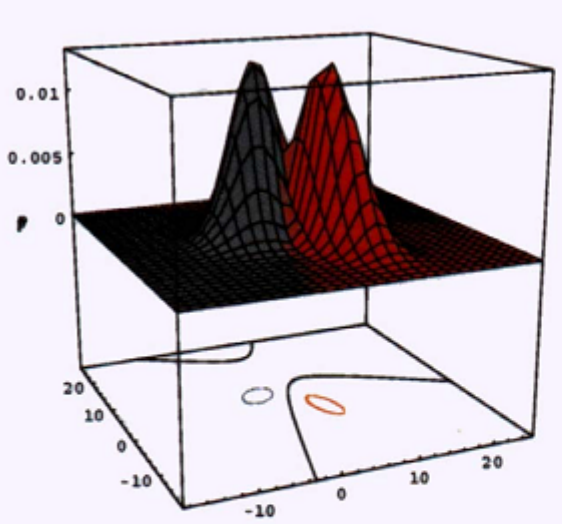
$$\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$$

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

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T \Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

(Hyperquadratics which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, etc.)

Decision Boundaries



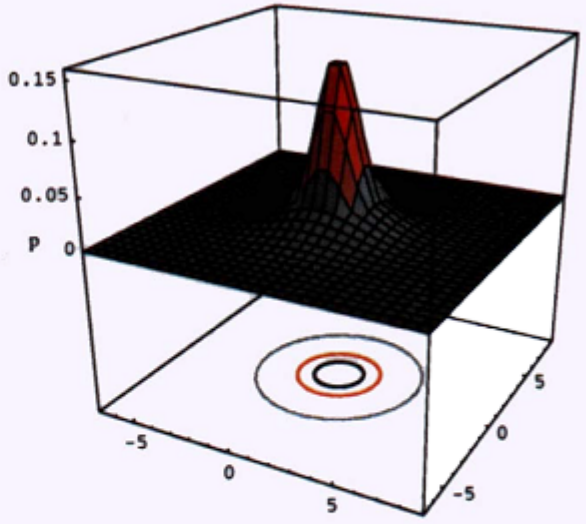
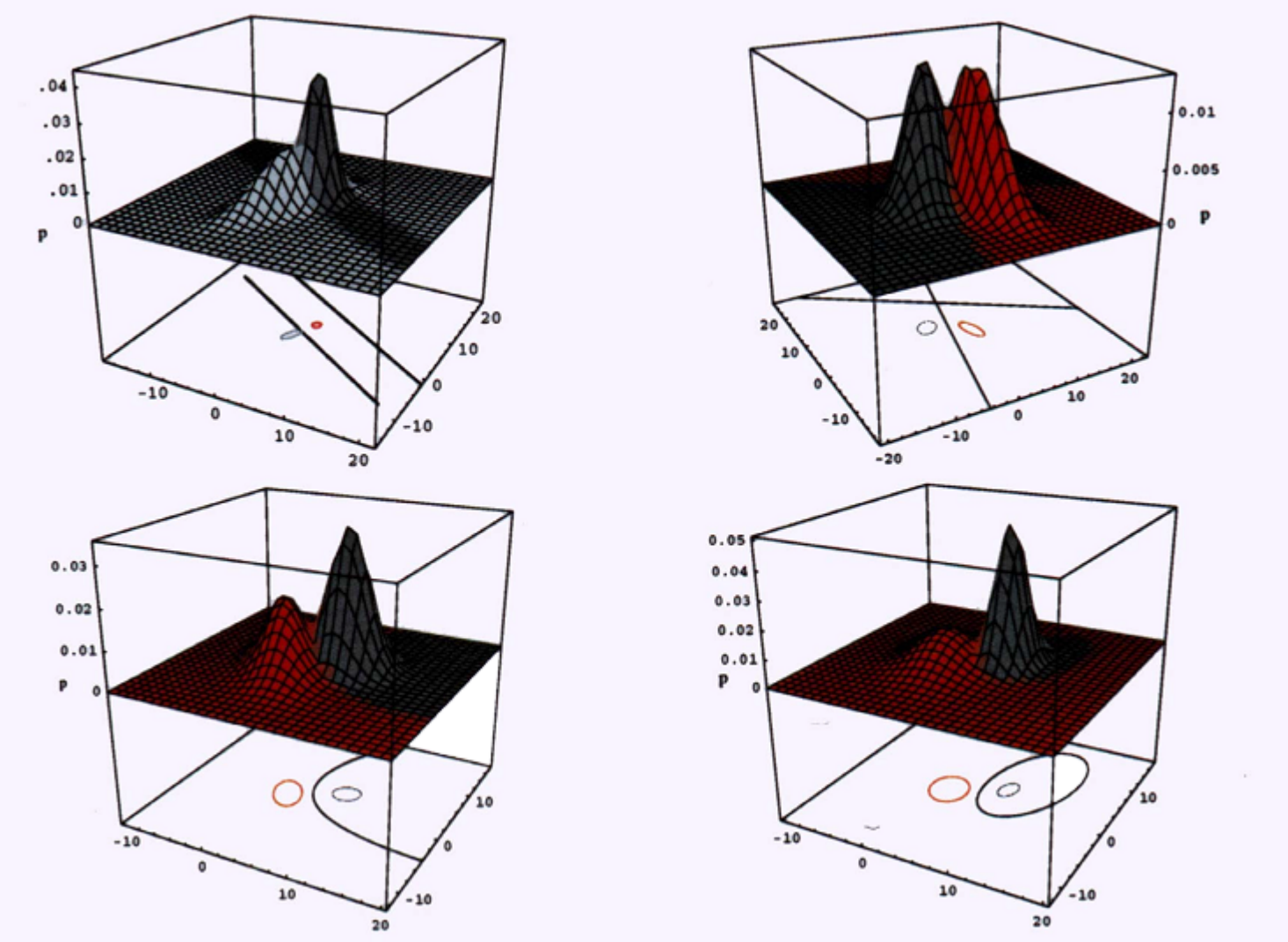


Figure 2.14: Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadratic, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric.

Decision Boundaries



Decision Boundaries

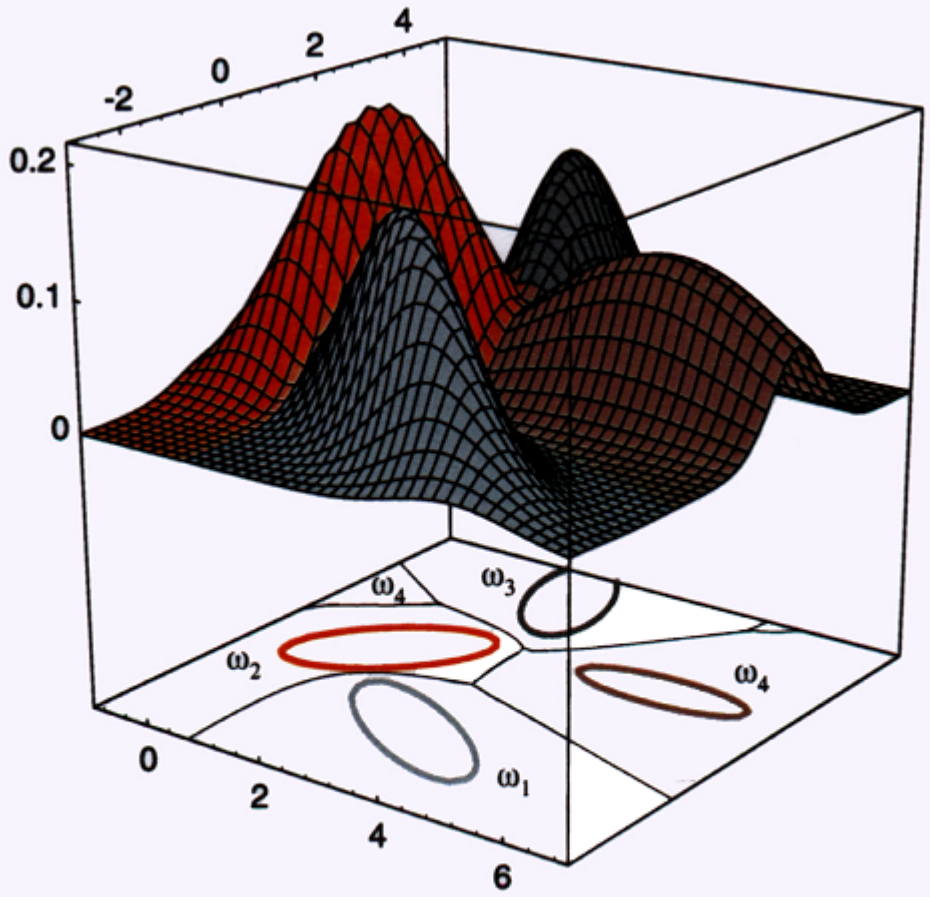


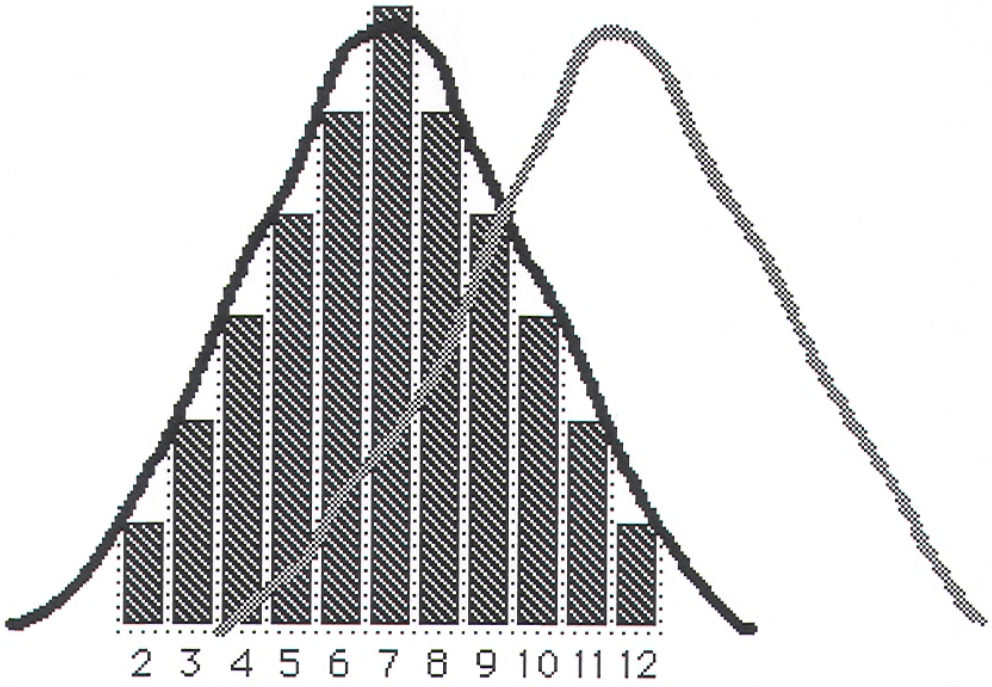
Figure 2.16: The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex.

Funny Dice Example

◊ Funny dice example
- Two players,
- Two pairs of dice,
  - One normal
  - One Augmented (2 extra spots on each side)
- Player 1 selects a pair of dice at random and rolls,
- Announces the total showing.
- Player 2 names the pair of dice used, with a $1 bet.

To determine a "decision boundary" in this case, one can determine how likely each possible outcome is (2 through 12 for the normal die, 6 through 16 for the augmented die). The result of this can be displayed as a histogram as follows:

Funny Dice Example



The decision boundary can then be set based upon choosing the most likely outcome in any given case. For example, if a 6 is the outcome, it is seen from the above to have more likely come from the normal dice. We shall generalize and formalize this idea next.

Funny Dice Example

Now let's complicate the funny dice rules. Suppose the dice to be rolled are selected at random from 80 standard pairs and 20 augmented pairs. You bet $1 on each play.

-if you guess wrong, you lose your dollar
-if you guess correctly when a standard pair is drawn, you win $1
-if you guess correctly when an augmented pair is drawn, you win $5

Possible strategies are as follows:

◊ **Maximum likelihood:**

$$g_s^1(x) = p(x|S)$$
$$g_a^1(x) = p(x|A)$$

Funny Dice Example

◊ **Minimum *a posteriori* probability of error**

It is $p(\omega_i|x)$, the so-called *a posteriori* probability, the probability of error after knowing the value of x, that we wish to maximize in this case. Bayes' Theorem states that,

$$p(\omega_i|x) = \frac{p(x|\omega_i)p(\omega_i)}{p(x)} = \frac{p(x,\omega_i)}{p(x)}$$

Note that p(x) is the same for any i, thus pick the larger of $g_i^2(x)$,

$g_s^2(x) = p(x,S) = p(x|S)p(S) = p(x|S)*0.8$

$g_a^2(x) = p(x,A) = p(x|A)p(A) = p(x|A)*0.2$

This strategy is referred to as the Bayes' Rule Strategy. Note that for equal class prior probabilities $p(\omega_i)$, this reduces to the Maximum likelihood strategy.

Funny Dice Example

◊ **Minimum risk**

The expected loss for each class is given by

$$L_{\underline{\omega}}(x) = \sum_{\omega_j} \lambda(\underline{\omega}|\omega_j)\, p(\omega_j|x)$$ where λ is the loss on any one outcome. ($\underline{\omega}$ is your answer, ω_j the true result.) So,

$$L_S(x) = \lambda(S|S)\, p(S|x) + \lambda(S|A)\, p(A|x)$$

$$L_A(x) = \lambda(A|S)\, p(S|x) + \lambda(A|A)\, p(A|x)$$

For the discriminant functions, we can use

$$g_S^3(x) = - L_S(x)$$

$$g_A^3(x) = - L_A(x)$$

For this problem, $\lambda(S|S) = -1$ (win = negative loss),

$\lambda(S|A) = 1$, $\lambda(A|S) = 1$, and $\lambda(A|A) = -5$

Funny Dice Example

Or we can use the equivalent discriminant functions:

$$g_{\underline{\omega}}^{3}(x) = -\sum \lambda(\underline{\omega}|\omega_i)\, p(x|\omega_i)\, p(\omega_i)$$

which yields

$$g_S^{3}(x) = \; p(x|S)\, p(S) - p(x|A)\, p(A)$$
$$g_A^{3}(x) = -\, p(x|S)\, p(S) + 5\, p(x|A)\, p(A)$$

The next step is to determine the decision rules. To do so it is convenient to establish a table, as shown.

Funny Dice Example

DECISION TABLE

x	p(x|S)	p(x|A)	Max Like. Decision	p(x|S)p(S)	p(x|A)p(A)	Max. a post Decision	gS3(x)	gA3(x)	Min Risk Decision
1	0	0		0	0				
2	0.028	0	S	0.022	0	S	0.022	-0.022	S
3	0.056	0	S	0.044	0	S	0.044	-0.044	S
4	0.083	0	S	0.067	0	S	0.067	-0.067	S
5	0.111	0	S	0.089	0	S	0.089	-0.089	S
6	0.139	0.028	S	0.111	0.006	S	0.106	-0.083	S
7	0.167	0.056	S	0.133	0.011	S	0.122	-0.078	S
8	0.139	0.083	S	0.111	0.017	S	0.094	-0.028	S
9	0.111	0.111	S	0.089	0.022	S	0.067	0.022	S
10	0.083	0.139	A	0.067	0.028	S	0.039	0.072	A
11	0.056	0.167	A	0.044	0.033	S	0.011	0.122	A
12	0.028	0.139	A	0.022	0.028	A	-0.006	0.117	A
13	0	0.111	A	0	0.022	A	-0.022	0.111	A
14	0	0.083	A	0	0.017	A	-0.017	0.083	A
15	0	0.056	A	0	0.011	A	-0.011	0.056	A
16	0	0.028	A	0	0.006	A	-0.006	0.028	A

From it the following decision rules can be established.

Funny Dice Example

Maximum Likelihood: Decide

$d_1(x)$ $\quad X \varepsilon S$ for $2 \leq x \leq 9$,

$\quad X \varepsilon A$ for $10 \leq x \leq 16$

Minimum Error Probability: Decide

$d_2(x)$ $\quad X \varepsilon S$ for $2 \leq x \leq 11$,

$\quad X \varepsilon A$ for $12 \leq x \leq 16$

Minimum Risk: Decide

$d_3(x)$ $\quad X \varepsilon S$ for $2 \leq x \leq 9$,

$\quad X \varepsilon A$ for $10 \leq x \leq 16$

(It is coincidental that Maximum Likelihood and Minimum Risk are the same.)

Funny Dice Example

To evaluate the three strategies, calculate the probability of being correct on any given play and the expected winings after 100 plays. This may be done as follows.

$$\text{pr(correct)} = \sum^{\omega_j} \text{pr}(X \varepsilon R_i | \omega_i) \quad R_i: g_i(X) \geq g_j(X) \quad i \neq j$$

For the maximum likelihood and minimum risk cases:

$$\text{pr(correct)} = \text{pr}(x \leq 9 \text{ and } \omega = S) + \text{pr}(x \geq 10 \text{ and } \omega = A)$$

$$= \sum_{x=2}^{9} \text{pr}(x,S) + \sum_{x=10}^{16} \text{pr}(x,A)$$

Funny Dice Example

$$= \sum_{x=2}^{9} pr(x|S)\, pr(S) + \sum_{x=10}^{16} pr(x|A)\, pr(A)$$

A similar equation is used for the minimum error rule except summing over the limits defined by that decision rule.

Funny Dice Example

For the winnings after 100 plays, calculate the expected (average) loss using the previous loss formula, them multiply -1 (to get expected winnings per play instead of loss) and 100. This may be done as follows. For a given class, ω_i, and decision rule, d(x), the expected loss is,

$$L_{d(x)}(x) = \sum^{\omega_j} \lambda(d(x)|\omega_i)\ p(\omega_i|x)$$

Then the expected loss over all possible outcomes (i.e. classes and x's) is,

$$E[L_{d(x)}(x)] = \sum_{x=2}^{16} L_{d(x)}\ p(x)$$

$$= \sum_{x=2}^{16} \sum^{\omega_j} \lambda(d(x)|\omega_i)\ p(\omega_i|x)\ p(x)$$

$$= \sum_{x=2}^{16} \sum^{\omega_j} \lambda(d(x)|\omega_i)\ p(x|\omega_i)\ p(\omega_i)$$

Funny Dice Example

The results for both p(correct) and the expected winnings turn out as follows:

Maximum Likelihood and Minimum Risk,
p(correct) = 0.811, $/100 plays = +$120

Minimum Error Probability
p(correct) = 0.861 $/100 plays = + $105.55

Thus, though the latter decision rule will produce more correct responses, the former ones will produce a larger dollar win. However, remember that it is coincidental that the Maximum Likelihood strategy also produced minimum risk in this case.

Classifiers and Error Computation

- Bayes Classifier
- Neyman-Pearson Classifier
- Minimax Test (Minimax Classifier)
- Error Compuation and Error Bounds
- Linear Classifiers

Bayesian 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 the two-category problem

Let $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$ where each x_i is either 0 or 1, with probabilities:

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

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

Bayesian Decision Theory – Discrete Features

The discriminant function in this case is:

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

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

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

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