

Aprenentatge Automàtic 1

GCED

Lluís A. Belanche

`belanche@cs.upc.edu`



Soft Computing Research Group
Dept. de Ciències de la Computació (Computer Science)
Universitat Politècnica de Catalunya

2019-2020

**LECTURE 5: Classification theory and linear classification models
(I)**

Bayesian decision theory

Introduction: Bayes' formula

Thomas Bayes: XVIII-century priest. His works on the celebrated formula were found upon his death

Discrete random variables. Let A a discrete r.v. with pmf P_A . We use the shorthand notation $P(a)$ to mean $P_A(A = a)$. Similarly we write $P(b|a)$ to mean $P_{B|A}(B = b|A = a)$, etc, where

$$P(b|a) = \frac{P(b, a)}{P(a)}, \quad P(a) > 0$$

(**prior, joint** and **conditional** probabilities)

Bayesian decision theory

Introduction: Bayes' formula

Let $\{a_1, \dots, a_n\}, \{b_1, \dots, b_m\}$ be the possible values that A, B can take.
Then, for any $a \in \{a_1, \dots, a_n\}$:

$$P(a) = \sum_{i=1}^m P(a, b_i) = \sum_{i=1}^m P(a|b_i)P(b_i)$$

Since $P(a, b) = P(b, a)$, it follows that, for any a_k, b_j :

$$P(b_j|a_k) = \frac{P(a_k|b_j)P(b_j)}{\sum_{i=1}^m P(a_k|b_i)P(b_i)}, \quad \text{with } \sum_{j=1}^m P(b_j|a_k) = 1$$

(**posterior** probabilities)

Bayesian decision theory

Introduction: Bayes' formula

Continuous random variables. Let X, Y two continuous r.v. with pdfs p_X, p_Y and joint density p_{XY} . We use the shorthand notation $p(x)$ to mean $p_X(X = x)$, etc.

$$p(x) = \int_{\mathbb{R}} p(x, y) dy; \quad p(y) = \int_{\mathbb{R}} p(x, y) dx$$

Therefore:

$$p(y|x) = \frac{p(x|y)p(y)}{\int_{\mathbb{R}} p(x|y)p(y) dy}, \text{ with } \int_{\mathbb{R}} p(y|x) dy = 1$$

Bayesian decision theory

Introduction: Bayes' formula

Mixed random variables. Suppose X is a continuous r.v. and Y is a discrete r.v. with values in $\{y_1, \dots, y_m\}$.

In this case, $p(\cdot|y_i)$ is a continuous r.v. and $P(\cdot|x)$ is a discrete r.v. Moreover,

$$P(y_j|x) = \frac{p(x|y_j)P(y_j)}{\sum_{i=1}^m p(x|y_i)P(y_i)}, \quad \text{with } \sum_{j=1}^m P(y_j|x) = 1$$

Bayesian decision theory

Decision rules

We are interested in determining the class or category of objects of nature according to Ω , a discrete r.v. with values $\{\omega_1, \omega_2\}$ that represent the two possible classes.

The prior probabilities are $P(\omega_1), P(\omega_2)$. How should we classify objects?

rule 1:

if $P(\omega_1) > P(\omega_2)$ then class of object is ω_1 else is ω_2

This rule classifies all objects into the same class; therefore it makes errors!

$$P_e(\text{rule1}) = \min\{P(\omega_1), P(\omega_2)\}$$

useful only if $P(\omega_1) \ll P(\omega_2)$ or $P(\omega_1) \gg P(\omega_2)$.

Bayesian decision theory

Decision rules

- Suppose now that X is a discrete r.v. taking values in $\{x_1, \dots, x_d\}$ that measures a **feature** of objects
- Now $P(\omega_i|x) = P(x|\omega_i)P(\omega_i)/P(x)$ is the **posterior** probability that an object with measured feature x belongs to class $P(\omega_i), i = 1, 2$
- Moreover $P(x) = P(x|\omega_1)P(\omega_1) + P(x|\omega_2)P(\omega_2)$

Upon observing x , the Bayes formula converts **prior** class probabilities $P(\omega_i)$ into **posterior** probabilities $P(\omega_i|x)$. How should we classify objects now?

rule 2:

if $P(\omega_1|x) > P(\omega_2|x)$ then class of object is ω_1 else class is ω_2

$$P_e(\text{rule2}) = \sum_{i=1}^d \min\{P(\omega_1|x_i), P(\omega_2|x_i)\}P(x_i)$$

(this rule is known as the **Bayes rule** or the **Bayes classifier**)

Bayesian decision theory

Decision rules

Lemma. For all $a, b, c, d \in \mathbb{R}$, $\min(a, b) + \min(c, d) \leq \min(a + c, b + d)$

Proposition 1 $P_e(\text{rule2}) \leq P_e(\text{rule1})$

Proof.

$$\begin{aligned} & \sum_{i=1}^d \min\{P(\omega_1|x_i), P(\omega_2|x_i)\}P(x_i) \\ &= \sum_{i=1}^d \min\{P(x_i)P(\omega_1|x_i), P(x_i)P(\omega_2|x_i)\} && \text{(Bayes formula)} \\ &= \sum_{i=1}^d \min\{P(x_i|\omega_1)P(\omega_1), P(x_i|\omega_2)P(\omega_2)\} && \text{(iterated lemma)} \\ &\leq \min\left\{\sum_{i=1}^d P(x_i|\omega_1)P(\omega_1), \sum_{i=1}^d P(x_i|\omega_2)P(\omega_2)\right\} \\ &= \min\left\{P(\omega_1) \sum_{i=1}^d P(x_i|\omega_1), P(\omega_2) \sum_{i=1}^d P(x_i|\omega_2)\right\} \\ &= \min\{P(\omega_1), P(\omega_2)\} \end{aligned}$$

The probabilities of error are equal only if $P(x_i|\omega_1) = P(x_i|\omega_2)$ for all i .

Bayesian decision theory

Example 1 We have a **conveyor belt** carrying two classes of pills, suitable for two different diseases (ω_1 and ω_2). These pills go in two colors {yellow, white}.

$$P(\omega_1) = \frac{1}{3}, P(\omega_2) = \frac{2}{3}$$

$$P(\text{yellow}|\omega_1) = \frac{1}{5}, P(\text{white}|\omega_1) = \frac{4}{5};$$

$$P(\text{yellow}|\omega_2) = \frac{2}{3}, P(\text{white}|\omega_2) = \frac{1}{3}$$

$$P(\text{yellow}) = P(\omega_1)P(\text{yellow}|\omega_1) + P(\omega_2)P(\text{yellow}|\omega_2) = \frac{1}{3} \cdot \frac{1}{5} + \frac{2}{3} \cdot \frac{2}{3} = \frac{23}{45}$$

$$P(\text{white}) = P(\omega_1)P(\text{white}|\omega_1) + P(\omega_2)P(\text{white}|\omega_2) = \frac{1}{3} \cdot \frac{4}{5} + \frac{2}{3} \cdot \frac{1}{3} = \frac{22}{45}$$

$$P(\omega_1|\text{yellow}) = \frac{P(\text{yellow}|\omega_1)P(\omega_1)}{P(\text{yellow})} = \left(\frac{1}{5} \cdot \frac{1}{3}\right) / \frac{23}{45} = \frac{3}{23}; \quad P(\omega_2|\text{yellow}) = 1 - P(\omega_1|\text{yellow}) = \frac{20}{23}$$

$$P(\omega_1|\text{white}) = \frac{P(\text{white}|\omega_1)P(\omega_1)}{P(\text{white})} = \left(\frac{4}{5} \cdot \frac{1}{3}\right) / \frac{22}{45} = \frac{6}{11}; \quad P(\omega_2|\text{white}) = 1 - P(\omega_1|\text{white}) = \frac{5}{11}$$

$$\Rightarrow P_e = \frac{23}{45} \cdot \frac{3}{23} + \frac{22}{45} \cdot \frac{5}{11} = \frac{13}{45} < \frac{1}{3} = \min \left\{ \frac{1}{3}, \frac{2}{3} \right\}$$

Bayesian decision theory

Continuous variables

The next step is to consider a r.v. X with pdf $p(x)$ that measures a *continuous* feature of the objects. Let \mathcal{P} be the support of p , i.e. $\mathcal{P} = \{x \in \mathbb{R} \mid p(x) > 0\}$.

In this setting, $p(x|\omega_i), i = 1, 2$ are the conditional densities of x for every class.

Proposition 2 $P_e(\text{rule2}) \leq P_e(\text{rule1})$

Proof.

$$\begin{aligned} & \int_{\mathcal{P}} \min\{P(\omega_1|x), P(\omega_2|x)\} p(x) dx \\ &= \int_{\mathcal{P}} \min\{p(x)P(\omega_1|x), p(x)P(\omega_2|x)\} dx && \text{(Bayes formula)} \\ &= \int_{\mathcal{P}} \min\{p(x|\omega_1)P(\omega_1), p(x|\omega_2)P(\omega_2)\} dx && \text{(standard result for integrals)} \\ &\leq \min\left\{\int_{\mathcal{P}} p(x|\omega_1)P(\omega_1) dx, \int_{\mathcal{P}} p(x|\omega_2)P(\omega_2) dx\right\} \\ &= \min\{P(\omega_1) \int_{\mathcal{P}} p(x|\omega_1) dx, P(\omega_2) \int_{\mathcal{P}} p(x|\omega_2) dx\} \\ &= \min\{P(\omega_1), P(\omega_2)\} \end{aligned}$$

The probabilities of error are equal only if $p(\cdot|\omega_1) = p(\cdot|\omega_2)$.

Bayesian decision theory

Example 2 We have a **conveyor belt** carrying two classes of pills, suitable for two different diseases (ω_1 and ω_2). This time the pills go in two colors, shaded in $[0, 2]$, with probabilities:

$$P(\omega_1) = \frac{1}{3}, P(\omega_2) = \frac{2}{3}, p(x|\omega_1) = \frac{2-x}{2}, p(x|\omega_2) = \frac{x}{2}.$$

$$p(x) = P(\omega_1)p(x|\omega_1) + P(\omega_2)p(x|\omega_2) = \frac{1}{3} \cdot \frac{2-x}{2} + \frac{2}{3} \cdot \frac{x}{2} = \frac{2+x}{6}$$

$$P(\omega_1|x) = \frac{p(x|\omega_1)P(\omega_1)}{p(x)} = \left(\frac{2-x}{2} \cdot \frac{1}{3}\right) / \frac{2+x}{6} = \frac{2-x}{2+x}; \quad P(\omega_2|x) = 1 - P(\omega_1|x) = 1 - \frac{2-x}{2+x} = \frac{2x}{2+x}$$

$$\Rightarrow P_e = P(\omega_2) \int_0^{2/3} p(x|\omega_2) dx + P(\omega_1) \int_{2/3}^2 p(x|\omega_1) dx = \frac{2}{9} < \frac{1}{3} = \min \left\{ \frac{1}{3}, \frac{2}{3} \right\}$$

$$\left(\frac{2}{3} \text{ is the solution of } \frac{2-x}{2+x} = \frac{2x}{2+x} \right)$$

Bayesian decision theory

The Bayes classifier

The Bayes classifier can be extended in two ways:

1. Consider a vector $X = (X_1, \dots, X_d)^\top$ of continuous r.v. with pdf $p(\mathbf{x}) = p(x_1, \dots, x_d)$ that measures d continuous features
2. Consider a finite number of classes Ω , a discrete r.v. with values $\omega_1, \dots, \omega_K$, that represent the possible classes ($K \geq 2$)

Therefore we have new probabilities $p(\mathbf{x}|\omega_i), P(\omega_i|\mathbf{x}), 1 \leq i \leq K$.

The new Bayes rule says:

the class $\hat{\omega}(\mathbf{x})$ of object \mathbf{x} is ω_k when $k = \arg \max_{i=1, \dots, K} P(\omega_i|\mathbf{x})$

The sets $\mathcal{R}_k = \{\mathbf{x} | \hat{\omega}(\mathbf{x}) = k\}$ are called **regions** (and depend on the specific classifier)

Bayesian decision theory

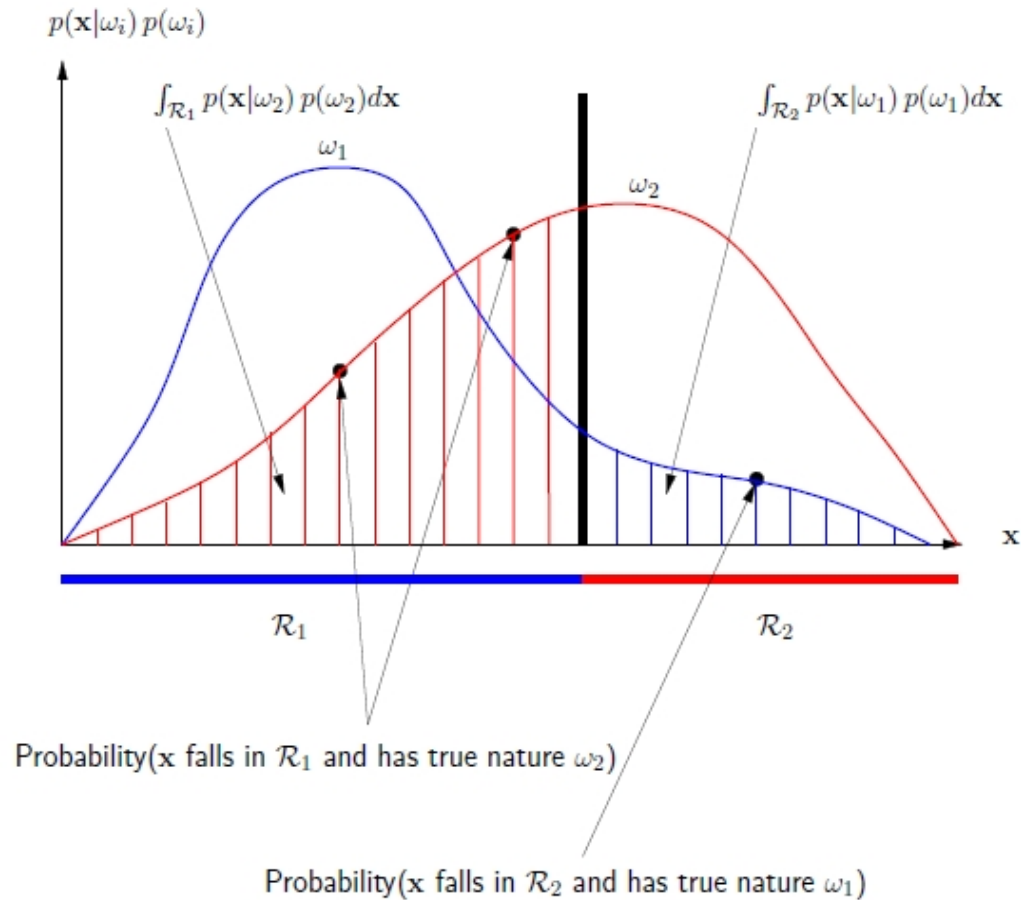
Illustration of the optimal classifier (two-class case)

Let us assume a classifier with regions $\mathcal{R}_1, \mathcal{R}_2$:

$$\begin{aligned} P_e &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\ &= P(\mathbf{x} \in \mathcal{R}_2 | \omega_1)P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1 | \omega_2)P(\omega_2) \\ &= \int_{\mathcal{R}_2} p(\mathbf{x} | \omega_1)P(\omega_1) d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x} | \omega_2)P(\omega_2) d\mathbf{x} \\ &\geq \int_{\mathcal{P}} \min \{p(\mathbf{x} | \omega_1)P(\omega_1), p(\mathbf{x} | \omega_2)P(\omega_2)\} d\mathbf{x} \\ &= P_e(\text{Bayes}) \end{aligned}$$

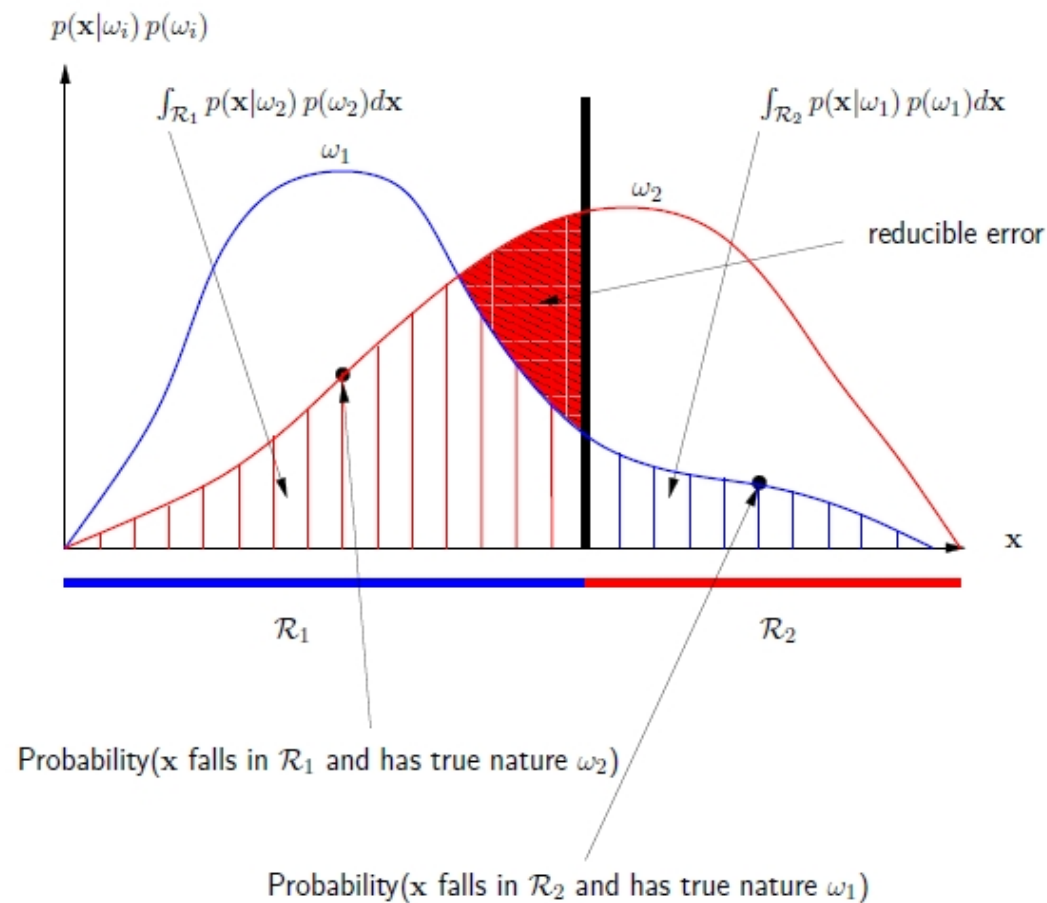
Bayesian decision theory

Graphical illustration



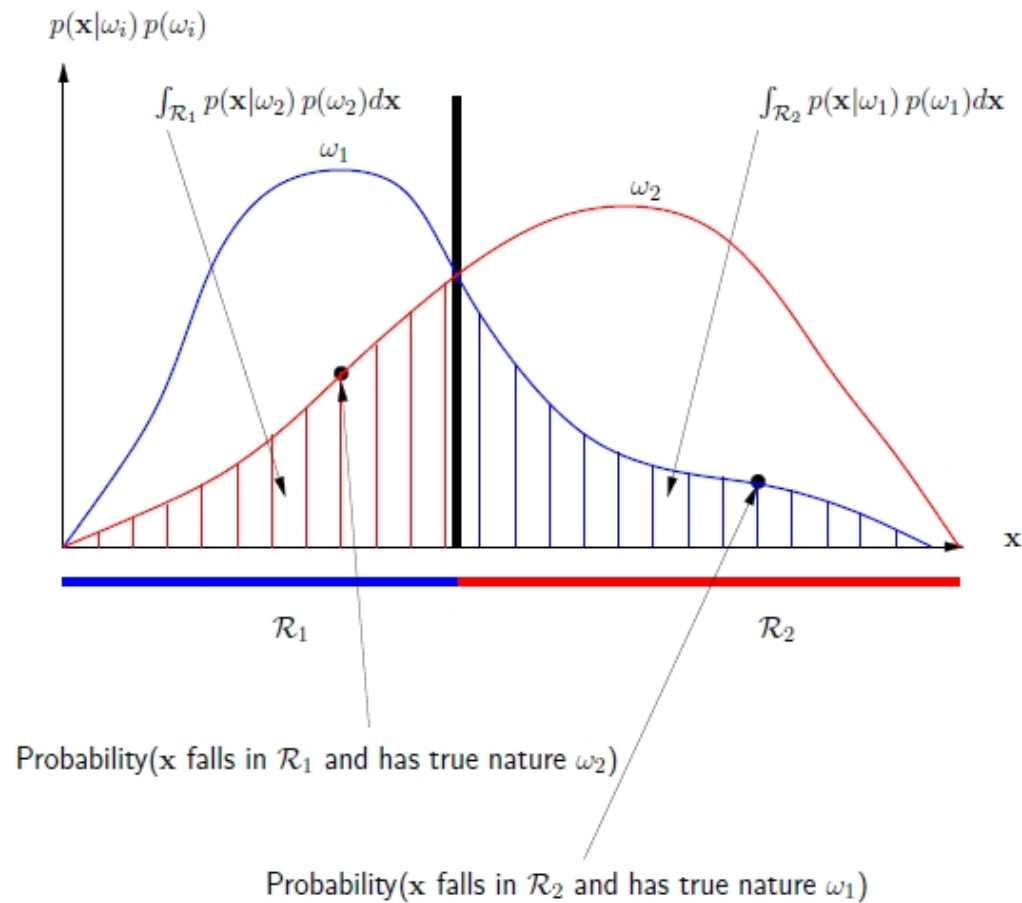
Bayesian decision theory

Graphical illustration



Bayesian decision theory

Graphical illustration



Bayesian decision theory

The Bayes classifier

The Bayes classifier can also have a **rejection class** (illustrated here for two classes); fix $\epsilon \in (0, 1)$:

if $P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) > \epsilon$ **then** class of object is ω_1

else if $P(\omega_2|\mathbf{x}) - P(\omega_1|\mathbf{x}) > \epsilon$ **then** class of object is ω_2

else do not classify

For every feature vector \mathbf{x} we take one of three possible **actions**.

Bayesian decision theory

The Bayes classifier

Consider a finite set of actions $A = \{a_1, \dots, a_m\}$. For each $a_i \in A$, denote by $l(a_i|\omega_j)$ the *loss* for choosing a_i when x is known to be in ω_j .

(this is a simplified setting in which the loss does not depend on x)

Example 3 Let $m = K + 1$ and let a_i stand for “classify x into class ω_i ” for $1 \leq i \leq K$; let a_{K+1} stand for “do not classify x ”. A possible set of losses is:

$$\begin{cases} l(a_i|\omega_j) = 1 & \text{for } 1 \leq i, j \leq K, i \neq j \\ l(a_i|\omega_i) = 0 & \text{for } 1 \leq i \leq K \\ l(a_{K+1}|\omega_j) = \frac{1}{2} & \text{for } 1 \leq j \leq K \end{cases}$$

... which suggests that a decision not to classify is less costly than a misclassification.

Bayesian decision theory

The notion of risk

For a given feature vector \mathbf{x} , define the **conditional risk** of an action as:

$$r(a_i|\mathbf{x}) := \sum_{j=1}^K l(a_i|\omega_j)P(\omega_j|\mathbf{x})$$

A **decision rule** is any function $a : \mathcal{P} \in \mathbb{R}^n \rightarrow A$ that assigns an action $a(\mathbf{x})$ to every \mathbf{x} s.t. $p(\mathbf{x}) > 0$. Define the **total risk** of a decision rule as:

$$R(a) := \int_{\mathcal{P}} r(a(\mathbf{x})|\mathbf{x})p(\mathbf{x}) d\mathbf{x}$$

Bayesian decision theory

The notion of risk

We are interested in the decision rule that minimizes the total risk.
Consider the rule

$$\hat{a}(\mathbf{x}) = \arg \min_{1 \leq j \leq m} r(a_j | \mathbf{x})$$

(you may recognize it as the Bayes rule!)

Given that this rule minimizes the argument of the integral for every possible \mathbf{x} , it follows that the Bayes rule has the lowest possible risk.

The value $R(\hat{a})$ is called the **Bayes risk**.

Bayesian decision theory

The notion of risk

Example 4 Yet again the conveyor belt carrying two classes of pills that go in two shaded colors (yellow,white), i.e. a scalar feature $x \in [0, 2]$, with probabilities:

$$P(\omega_1) = \frac{2}{3}, P(\omega_2) = \frac{1}{3}, p(x|\omega_1) = \frac{2-x}{2}, p(x|\omega_2) = \frac{1}{2}.$$

and three possible actions:

a_1 — classify as ω_1 ,
 a_2 — classify as ω_2 ,
 a_3 — do not classify

Let the loss matrix $l_{ij} \equiv l(a_i|\omega_j)$ be:

	ω_1	ω_2
a_1	0	1
a_2	1	0
a_3	$\frac{1}{4}$	$\frac{1}{4}$

-
1. Compute the optimal decision rule and its associated risk
 2. Give the probability that an object is *not* classified

Bayesian decision theory

The notion of risk

We have

$$p(x) = \frac{2}{3} \cdot \frac{2-x}{2} + \frac{1}{3} \cdot \frac{1}{2} = \frac{5-2x}{6}$$

$$P(\omega_1|x) = \frac{2}{3} \cdot \frac{2-x}{2} / \frac{5-2x}{6} = \frac{4-2x}{5-2x}; \quad P(\omega_2|x) = 1 - P(\omega_1|x) = 1 - \frac{4-2x}{5-2x} = \frac{1}{5-2x}$$

The conditional risks are:

$$r_1(x) \equiv r(a_1|x) = 0 \cdot P(\omega_1|x) + 1 \cdot P(\omega_2|x) = \frac{1}{5-2x}$$

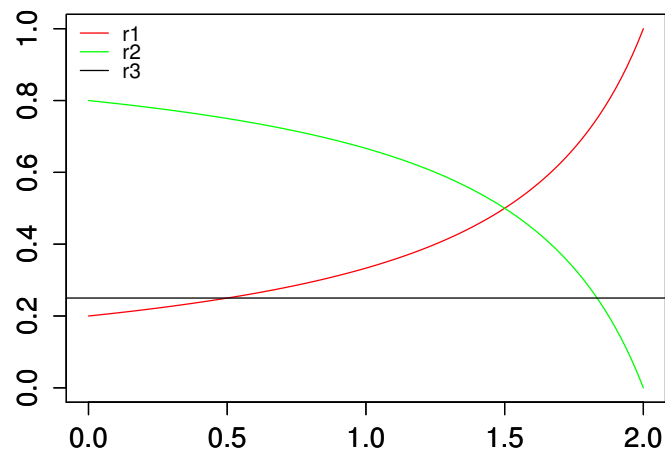
$$r_2(x) \equiv r(a_2|x) = 1 \cdot P(\omega_1|x) + 0 \cdot P(\omega_2|x) = \frac{4-2x}{5-2x}$$

$$r_3(x) \equiv r(a_3|x) = \frac{1}{4} \cdot P(\omega_1|x) + \frac{1}{4} \cdot P(\omega_2|x) = \frac{1}{4}$$

Now the Bayes rule chooses for each x the action with minimum conditional risk.

Bayesian decision theory

The notion of risk



$0 \leq x \leq \frac{1}{2} \Rightarrow$ take action $a_1 \Rightarrow$ choose ω_1

$\frac{1}{2} \leq x \leq \frac{11}{6} \Rightarrow$ take action $a_3 \Rightarrow$ do not classify

$\frac{11}{6} \leq x \leq 2 \Rightarrow$ take action $a_2 \Rightarrow$ choose ω_2

Bayesian decision theory

The notion of risk

- The total Bayes risk is:

$$\begin{aligned} R &= \int_0^{\frac{1}{2}} r_1(x)p(x) dx + \int_{\frac{1}{2}}^{\frac{11}{6}} r_3(x)p(x) dx + \int_{\frac{11}{6}}^2 r_2(x)p(x) dx \\ &= \frac{1}{12} + \frac{1}{3} + \frac{1}{6} + \ln \frac{\sqrt{3}}{2} \approx 0,44 \end{aligned}$$

- The probability that an object is *not* classified is:

$$\int_{\frac{1}{2}}^{\frac{11}{6}} \frac{5-2x}{6} dx = \frac{16}{27} \approx 0,59$$

Bayesian decision theory

0/1 losses

In many applications the 0/1 loss is used (usually in absence of more precise information):

$$l_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

Consider K classes and actions a_i — classify \mathbf{x} into ω_i . Then:

$$r(a_i|\mathbf{x}) = \sum_{j=1}^K l_{ij}P(\omega_j|\mathbf{x}) = \sum_{j=1, i \neq j}^K P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

Bayesian decision theory

Discriminant functions

- Functions of the form $g_k : \mathcal{P} \rightarrow \mathbb{R}$ are a useful tool to build an abstract classifier
- An object x is assigned to class ω_i when $g_i(x)$ is the highest among the values $g_1(x), \dots, g_K(x)$. Examples:
 - $g_k(x) = P(\omega_k|x)$
 - $g_k(x) = P(\omega_k)p(x|\omega_k)$
 - $g_k(x) = -r(a_k|x)$
- If g_k is a discriminant function, then so is $h \circ g_k$, for any strictly monotonic function h .
- For two classes, we can use a single discriminant function, called a **dichotomizer**:
 1. define $g(x) := g_1(x) - g_2(x)$
 2. assign x to class ω_1 if $g(x) > 0$ and to class ω_2 if $g(x) < 0$

Bayesian decision theory

The Gaussian Distribution

A normally distributed d -variate random vector $X = (X_1, \dots, X_d)^\top$ has pdf:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

where $\boldsymbol{\mu}$ is the **mean vector** and $\Sigma_{d \times d} = (\sigma_{ij}^2)$ is the real symmetric and positive definite (p.d.) **covariance matrix**.

- $\mathbb{E}[X] = \boldsymbol{\mu}$ and $\mathbb{E}[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^\top] = \Sigma$.
- $\text{CoVar}[X_i, X_j] = \sigma_{ij}^2$ and $\text{Var}[X_i] = \sigma_{ii}^2$

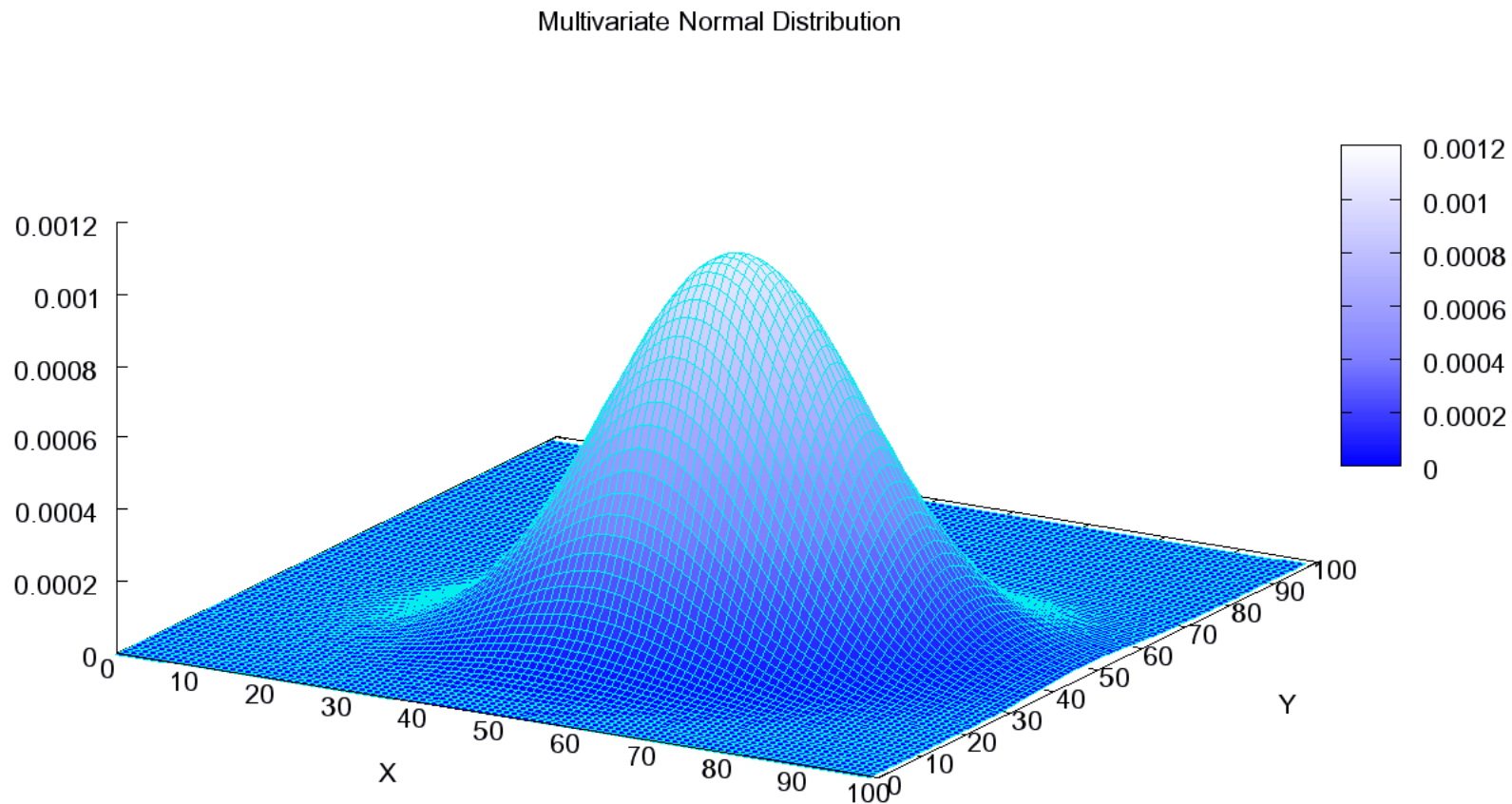
if $X \sim N(\boldsymbol{\mu}, \Sigma)$, then X_i, X_j are statistically independent $\iff \text{CoVar}[X_i, X_j] = 0$

(in general, only the left-to-right implication holds)

Notation: It is convenient to write $p(\mathbf{x}) = N(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ or $X \sim N(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$.

Bayesian decision theory

The Gaussian Distribution



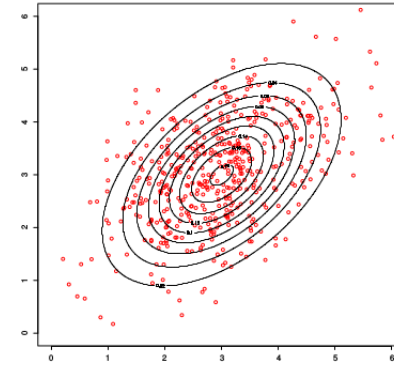
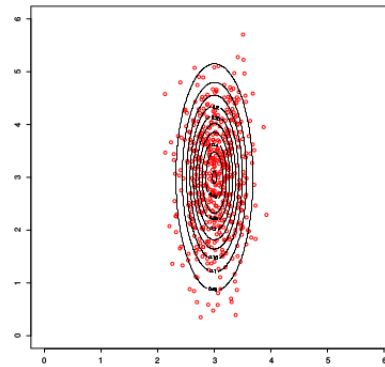
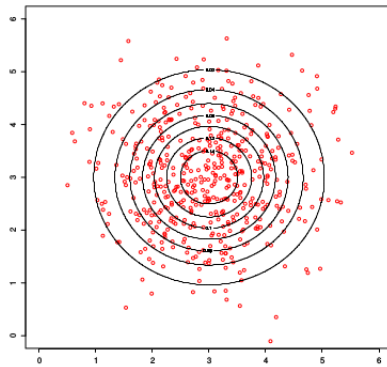
Bayesian decision theory

The Gaussian Distribution

- The quantity $d(x, \mu) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$ is called the **Mahalanobis distance**
- What is behind the choice of a multivariate Gaussian for a class?
→ examples from the class are noisy versions of a *prototype*:
 - Prototype: modeled by the mean vector
 - Noise: modeled by the covariance matrix
- Very important: the number of parameters is $\frac{d(d+1)}{2} + d$

Bayesian decision theory

The Gaussian Distribution



$$\mu = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

- The surfaces of equal probability $d(\mathbf{x}, \boldsymbol{\mu}) = ct$ are **hyperellipsoids**
- The principal directions or components (PC) of the hyperellipsoids are given by the **eigenvectors** \mathbf{u}_i of Σ , which satisfy $\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i, \lambda_i > 0$
- The lengths of the hyperellipsoids along these axes are proportional to $\sqrt{\lambda_i}$, the **singular values** associated with \mathbf{u}_i (note Σ is p.d. and therefore all $\lambda_i > 0$)

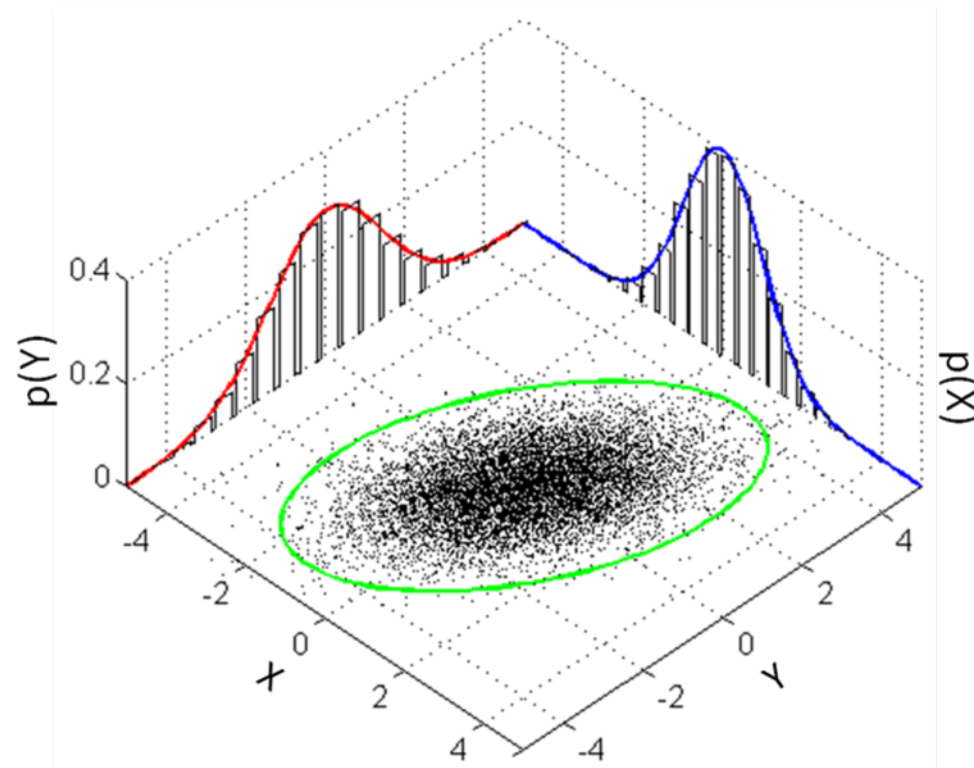
Bayesian decision theory

Properties of the Gaussian Distribution

- **Simplified forms:** if the X_i are statistically independent, then
$$p(\mathbf{x}) = \prod_{i=1}^d p(x_i),$$
 Σ is diagonal and the PCs are axis-aligned
(**marginal densities** [integrating out some variables] and **conditional densities** [setting some variables to fixed values] are also normal)
- **Analytical properties**, e.g. any moment $\mathbb{E}[X^p]$ can be expressed as a function of μ and Σ
- **Central limit theorem**, the mean of d i.i.d. random variables tends to a normal distribution, in the limit of infinite d
- **Linear transformation invariance**, the distribution of a linear transformation of the coordinate system remains normal

Bayesian decision theory

The Gaussian Distribution



Observations from a bivariate ($d = 2$) normal distribution, a contour ellipsoid, the two marginal distributions, and their histograms (images from the Wikipedia)