Machine Learning

MIRI Master

Lluís A. Belanche

belanche@cs.upc.edu

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

Universitat Politècnica de Catalunya

Spring Semester 2016-2017

LECTURE 4: Theoretical issues (II): classification

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

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

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

Bayesian decision theory

Introduction: Bayes' formula

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

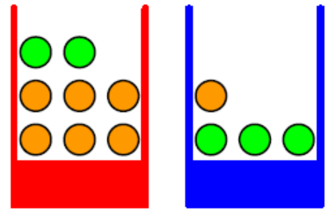
$$P(a) = \sum_{i=1}^{m} P(a, b_i) = \sum_{i=1}^{m} P(a|b_i)P(b_i)$$

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

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

(**posterior** probabilities)

Bayesian decision theory



Example 1 *The red box contains 6 oranges and 2 apples, the blue box contains 1 orange and 3 apples. Suppose we pick the red box 40 % of the time and the blue box 60 % of the time.*

1. *What is the overall probability that we pick an apple?*
2. *Given that we have chosen an orange, what is the probability that the box we chose was the blue one?*

(from Bishop's *Pattern Recognition and Machine Learning* book)

Bayesian decision theory

Let us introduce random variables B for box and F for fruit:

- $B = r$ (for red) and $B = b$ (for blue)

- $F = o$ (for orange) and $F = a$ (for apple)

The **prior** probabilities of selecting the red or blue boxes are

$$P(B = r) = \frac{4}{10} \qquad\qquad P(B = b) = \frac{6}{10}$$

Bayesian decision theory

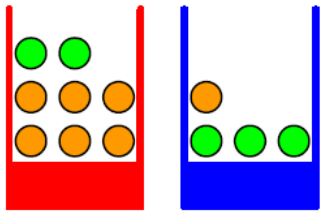
Now for the **conditional** probabilities:

$$P(F = a|B = r) = \frac{1}{4}$$

$$P(F = o|B = r) = \frac{3}{4}$$

$$P(F = a|B = b) = \frac{3}{4}$$

$$P(F = o|B = b) = \frac{1}{4}$$



Bayesian decision theory

What is the overall (**unconditional**) probability that we pick an apple?

$$\begin{aligned} P(F=a) &= P(F=a|B=r)P(B=r)+P(F=a|B=b)P(B=b) \\ &= \frac{1}{4}\cdot\frac{4}{10}+\frac{3}{4}\cdot\frac{6}{10}=\frac{11}{20} \end{aligned}$$

Therefore $P(F=o)=1-\frac{11}{20}=\frac{9}{20}$.

Although there are more oranges in total, picking an apple is more likely!

Bayesian decision theory

Given that we have chosen an orange, what is the **posterior** probability that the box we chose was the blue one?

$$P(B=b|F=o) = \frac{P(F=o|B=b)P(B=b)}{P(F=o)} = \frac{1}{4} \cdot \frac{6}{10} \cdot \frac{20}{9} = \frac{1}{3}$$

$$P(B=r|F=o) = \frac{P(F=o|B=r)P(B=r)}{P(F=o)} = \frac{3}{4} \cdot \frac{4}{10} \cdot \frac{20}{9} = \frac{2}{3}$$

Note that $P(B=b|F=o) + P(B=r|F=o) = 1$, as they should, because conditional distributions are distributions.

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.

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

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 $\{y_1, \ldots, y_m\}$.

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

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

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:

$$\boxed{\text{if } P(\omega_1) > P(\omega_2) \text{ then class of object is } \omega_1 \text{ else is } \omega_2}$$

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

$$P_{\mathrm{e}}(\text{rule1}) = \text{mín}\{P(\omega_1), P(\omega_2)\}$$

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 $\{x_1, \ldots, x_d\}$ 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:

$$\boxed{\text{if } P(\omega_1|x) > P(\omega_2|x) \text{ then class of object is } \omega_1 \text{ else class is } \omega_2}$$

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

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

Decision rules

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

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

Proof.

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

$$= \sum_{i=1}^{d} \text{mín}\{P(x_i)P(\omega_1|x_i), P(x_i)P(\omega_2|x_i)\} \qquad (\textit{Bayes formula})$$

$$= \sum_{i=1}^{d} \text{mín}\{P(x_i|\omega_1)P(\omega_1), P(x_i|\omega_2)P(\omega_2)\} \qquad (\textit{iterated lemma})$$

$$\leq \text{mín}\{\sum_{i=1}^{d} P(x_i|\omega_1)P(\omega_1), \sum_{i=1}^{d} P(x_i|\omega_2)P(\omega_2)\}$$

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

$$= \text{mín}\{P(\omega_1), P(\omega_2)\}$$

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 2 *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 $\{yellow, white\}$.*

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

$P(yellow|\omega_1) = \frac{1}{5}, P(white|\omega_1) = \frac{4}{5};$ $P(yellow|\omega_2) = \frac{2}{3}, P(white|\omega_2) = \frac{1}{3}$

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

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

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

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

$$\Longrightarrow P_{\text{e}} = \frac{23}{45} \cdot \frac{3}{23} + \frac{22}{45} \cdot \frac{5}{11} = \frac{13}{45} < \frac{1}{3} = \text{mín}\left\{\frac{1}{3}, \frac{2}{3}\right\}$$

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. $\mathcal{P} = \{x \in \mathbb{R} | \; p(x) > 0\}$.

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

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

Proof.

$\int_\mathcal{P} \text{mín}\{P(\omega_1|x), P(\omega_2|x)\}p(x)\, dx$

$= \int_\mathcal{P} \text{mín}\{p(x)P(\omega_1|x), p(x)P(\omega_2|x)\}\, dx$ (*Bayes formula*)

$= \int_\mathcal{P} \text{mín}\{p(x|\omega_1)P(\omega_1), p(x|\omega_2)P(\omega_2)\}\, dx$ (*standard result for integrals*)

$\leq \text{mín}\{\int_\mathcal{P} p(x|\omega_1)P(\omega_1)\, dx, \int_\mathcal{P} p(x|\omega_2)P(\omega_2)\, dx\}$

$= \text{mín}\{P(\omega_1) \int_\mathcal{P} p(x|\omega_1)\, dx, P(\omega_2) \int_\mathcal{P} p(x|\omega_2)\, dx\}$

$= \text{mín}\{P(\omega_1), P(\omega_2)\}$

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

Bayesian decision theory

Example 3 *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:*

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

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

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

$$\Longrightarrow P_{\mathrm{e}} = P(\omega_2) \int_0^{2/3} p(x|\omega_2)\, dx + P(\omega_1) \int_{2/3}^{2} p(x|\omega_1)\, dx = \frac{2}{9} < \frac{1}{3} = \mathrm{mín} \left\{\frac{1}{3}, \frac{2}{3}\right\}$$

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

Bayesian decision theory

The Bayes classifier

The Bayes classifier can be extended in two ways:

1. Consider a vector $X = (X_1, \ldots, X_d)^T$ of continuous r.v. with pdf $p(\boldsymbol{x}) = p(x_1, \ldots, x_d)$ that measures d continuous features

2. Consider a finite number of classes Ω, a discrete r.v. with values $\omega_1, \ldots, \omega_K$, that represent the possible classes ($K \geq 2$)

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

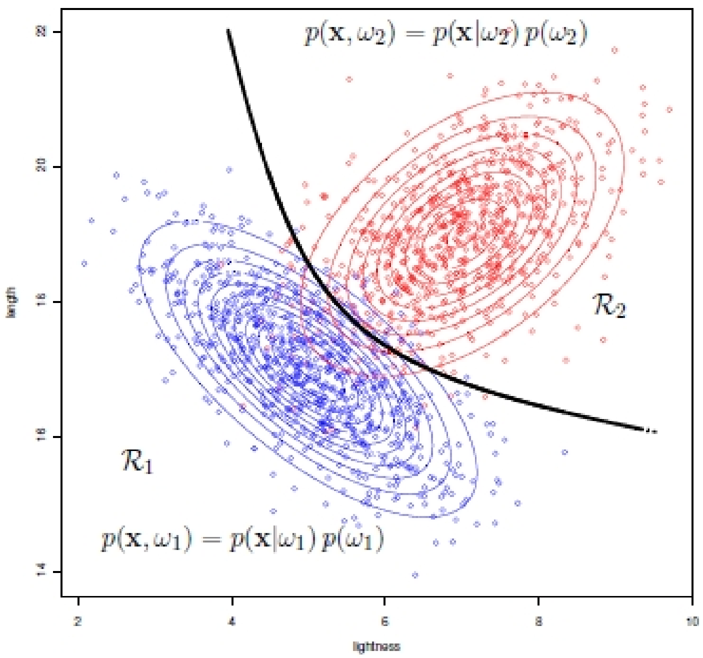
The new Bayes rule says:

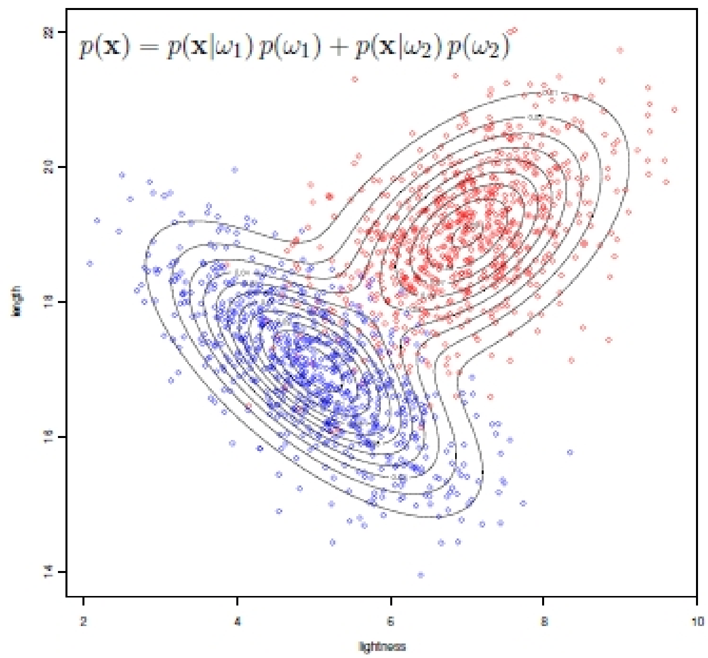
$$\boxed{\text{the class } \widehat{w}(\boldsymbol{x}) \text{ of object } \boldsymbol{x} \text{ is } \omega_k \text{ when } k = \operatorname*{arg\,max}_{i=1,\ldots,K} P(\omega_i|\boldsymbol{x})}$$

The sets $\mathcal{R}_k = \{\boldsymbol{x} / \widehat{w}(\boldsymbol{x}) = k\}$ are called **regions** (and depend on the specific classifier)

Bayesian decision theory

The fish factory





The Bayes rule says:

$$\boxed{\text{if } P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x}) \text{ then } \widehat{w}(\boldsymbol{x}) = \omega_1 \text{ else } \widehat{w}(\boldsymbol{x}) = \omega_2}$$

Bayesian decision theory

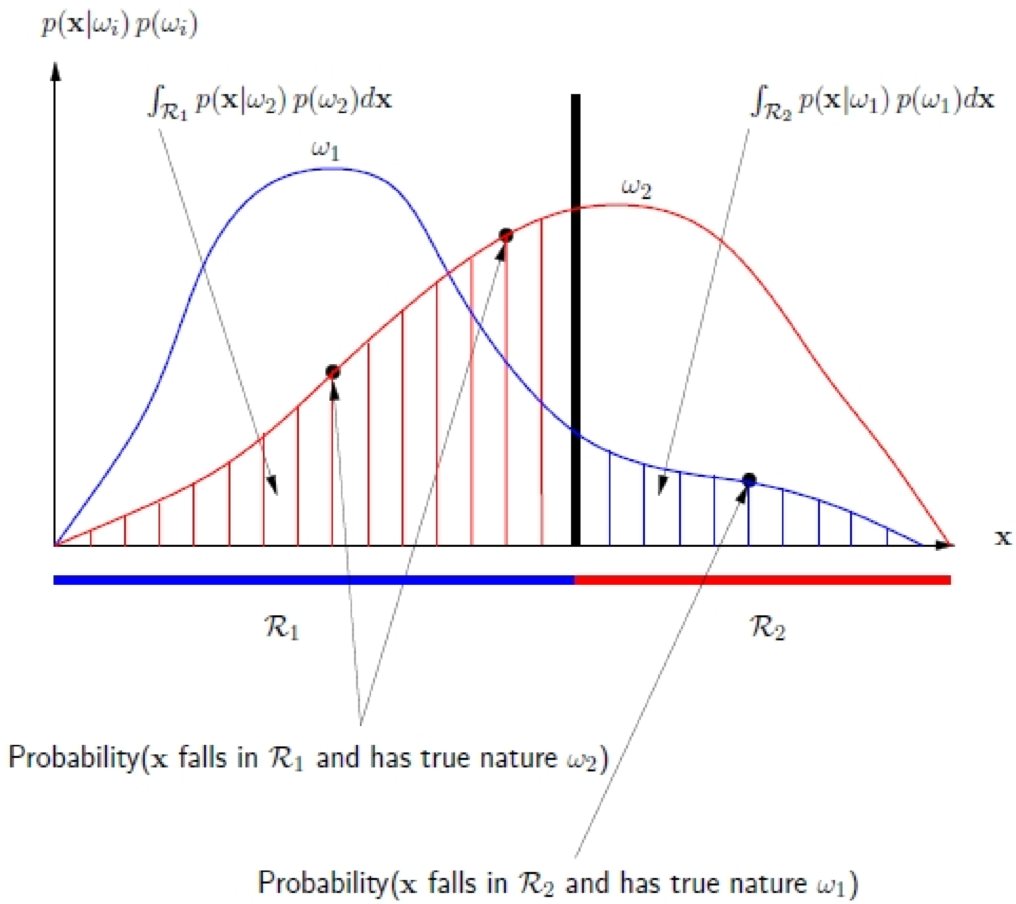
Illustration of the optimal classifier (two-class case)

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

$$\begin{aligned}
P_{\mathrm{e}} &= P(\boldsymbol{x} \in \mathcal{R}_2, \omega_1) + P(\boldsymbol{x} \in \mathcal{R}_1, \omega_2) \\
&= P(\boldsymbol{x} \in \mathcal{R}_2 | \omega_1) P(\omega_1) + P(\boldsymbol{x} \in \mathcal{R}_1 | \omega_2) P(\omega_2) \\
&= \int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1) P(\omega_1)\, d\boldsymbol{x} + \int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2) P(\omega_2)\, d\boldsymbol{x} \\
&\geq \int_{\mathcal{P}} \text{mín} \left\{ p(\boldsymbol{x}|\omega_1) P(\omega_1), p(\boldsymbol{x}|\omega_2) P(\omega_2) \right\} d\boldsymbol{x} \\
&= P_{\mathrm{e}}(\text{Bayes})
\end{aligned}$$

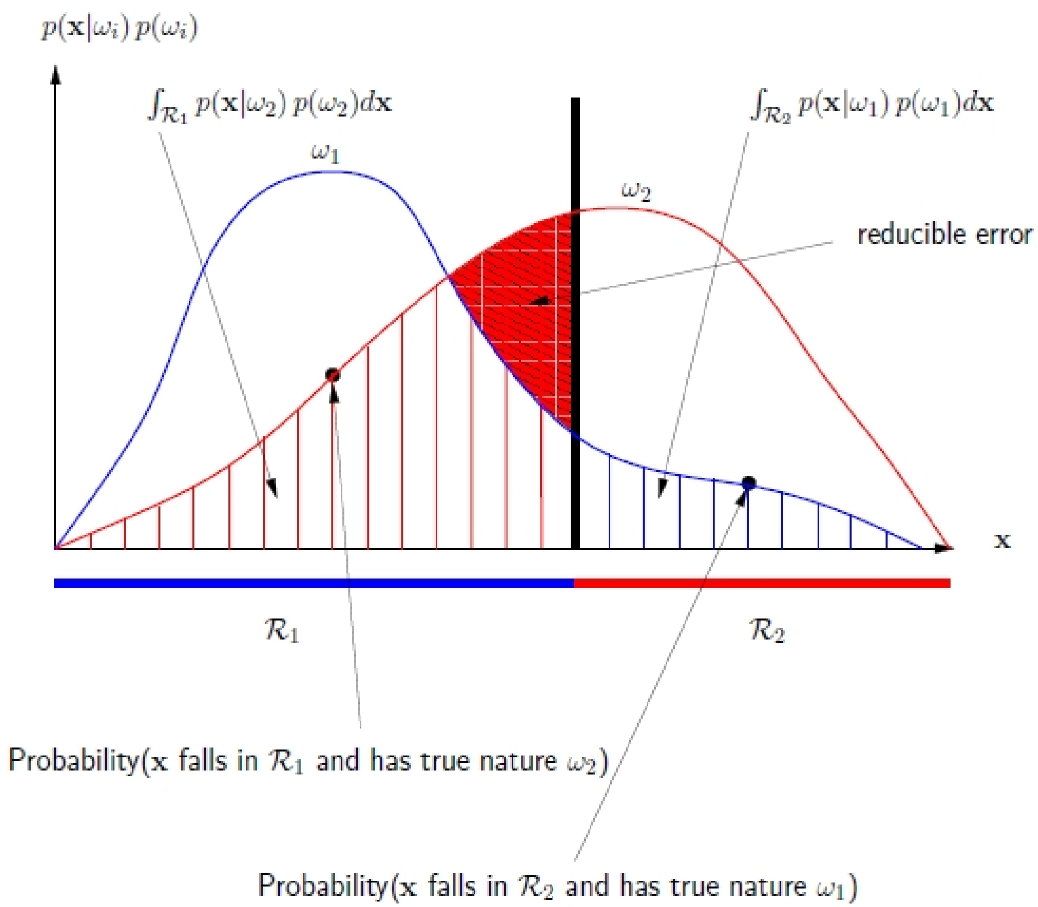
Bayesian decision theory

Graphical illustration



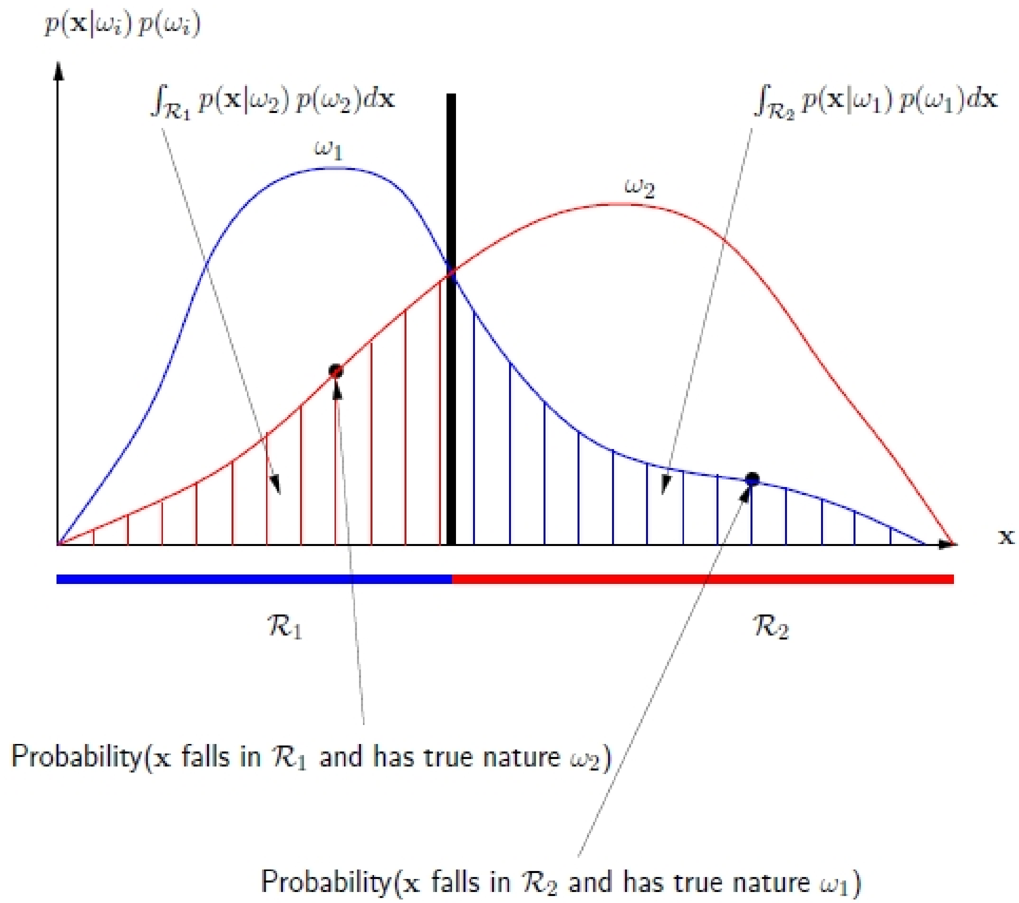
Bayesian decision theory

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

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

else do not classify

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

Bayesian decision theory

The Bayes classifier

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

(note this is a simplified setting in which the loss does not depend on $\boldsymbol{x}$)

Example 4 *Let $m = K + 1$ and let a_i stand for "classify $\boldsymbol{x}$ into class ω_i" for $1 \leq i \leq K$; let a_{K+1} stand for "do not classify $\boldsymbol{x}$". A possible set of losses is:*

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

... 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 $\boldsymbol{x}$, define the **conditional risk** of an action as:

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

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

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

Bayesian decision theory

The notion of risk

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

$$\hat{a}(\boldsymbol{x}) = \operatorname*{arg\,min}_{1 \le j \le m} r(a_j|\boldsymbol{x})$$

(you may recognize it as the Bayes rule!)

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

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

The notion of risk

Example 5 *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:

$$\begin{array}{l} a_1 - \textit{ classify as } \omega_1, \\ a_2 - \textit{ classify as } \omega_2, \\ a_3 - \textit{ do not classify} \end{array}$$

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

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

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

The conditional risks are:

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

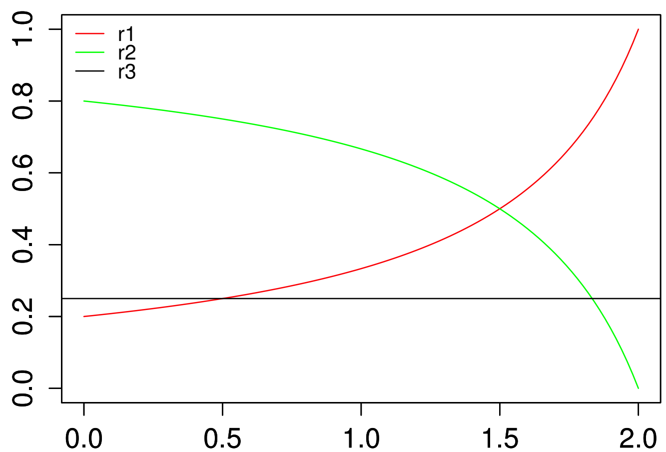
$r_2(x) \equiv r(a_2|x) = 1 \cdot P(\omega_1|x) + 0 \cdot P(\omega_2|x) = \frac{4-2x}{5-2x}$

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

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

$\frac{1}{2} \leq x \leq \frac{11}{16} \Rightarrow$ take action $a_3 \Rightarrow$ do not classify

$\frac{11}{16} \leq x \leq 2 \Rightarrow$ take action $a_2 \Rightarrow$ choose ω_2

Bayesian decision theory

The notion of risk

- The total Bayes risk is:

$$R = \int_0^{\frac{1}{2}} r_1(x)p(x)\,dx + \int_{\frac{1}{2}}^{\frac{11}{16}} r_3(x)p(x)\,dx + \int_{\frac{11}{16}}^{2} r_2(x)p(x)\,dx$$
$$= \frac{1}{12} + \frac{4}{27} + \frac{1}{216} = \frac{1377}{5832} \approx 0{,}236$$

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

$$\int_{\frac{1}{2}}^{\frac{11}{16}} \frac{5-2x}{6}\,dx = \frac{59}{108} \approx 0{,}546$$

Bayesian decision theory

The likelihood-ratio test (LRT)

Consider the simple two-class case: a_1- classify as ω_1, a_2- classify as ω_2.

Given a feature vector $\boldsymbol{x}$, we take action a_1 when $r(a_1|\boldsymbol{x}) < r(a_2|\boldsymbol{x})$:

$$l_{11}P(\omega_1|\boldsymbol{x}) + l_{12}P(\omega_2|\boldsymbol{x}) < l_{21}P(\omega_1|\boldsymbol{x}) + l_{22}P(\omega_2|\boldsymbol{x})$$

For $\boldsymbol{x} \in \mathcal{P}$, applying Bayes' formula and grouping terms:

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

Assuming (rather naturally) that $l_{21} > l_{11}$ and that $l_{12} > l_{22}$, then:

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)};\ \ \lambda = \frac{(l_{12} - l_{22})P(\omega_2)}{(l_{21} - l_{11})P(\omega_1)}$$

The test $\Lambda(\boldsymbol{x}) > \lambda$ (choose a_1) or $\Lambda(\boldsymbol{x}) < \lambda$ (choose a_2) is called the **LRT**.

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 $\boldsymbol{x}$ into ω_i. Then:

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

Bayesian decision theory

Discriminant functions

- Functions of the form $g_k : \mathcal{P} \to \mathbb{R}$ are a useful tool to build an abstract classifier

- An object $\boldsymbol{x}$ is assigned to class ω_i when $g_i(\boldsymbol{x})$ is the highest among the values $g_1(\boldsymbol{x}), \ldots, g_K(\boldsymbol{x})$. Examples:
  - $g_k(\boldsymbol{x}) = P(\omega_k|\boldsymbol{x})$
  - $g_k(\boldsymbol{x}) = P(\omega_k)p(\boldsymbol{x}|\omega_k)$
  - $g_k(\boldsymbol{x}) = -r(a_k|\boldsymbol{x})$

- 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 $g(\boldsymbol{x}) := g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$
  2. assign $\boldsymbol{x}$ to class ω_1 if $g(\boldsymbol{x}) > 0$ and to class ω_2 if $g(\boldsymbol{x}) < 0$

Bayesian decision theory

The Gaussian Distribution

A continuous r.v. X is normally distributed when its pdf is:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

where μ is the *mean* and σ^2 is the *variance*:

- $\mathbb{E}[X] = \int_{\mathbb{R}} xp(x)\,dx = \mu$
- $\mathbb{E}[(X-\mu)^2] = \int_{\mathbb{R}} (x-\mu)^2 p(x)\,dx = \sigma^2$

Notation: It is convenient to write $p(x) = N(x;\mu,\sigma^2)$ or $X \sim N(x;\mu,\sigma^2)$

Bayesian decision theory

The Gaussian Distribution

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

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

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

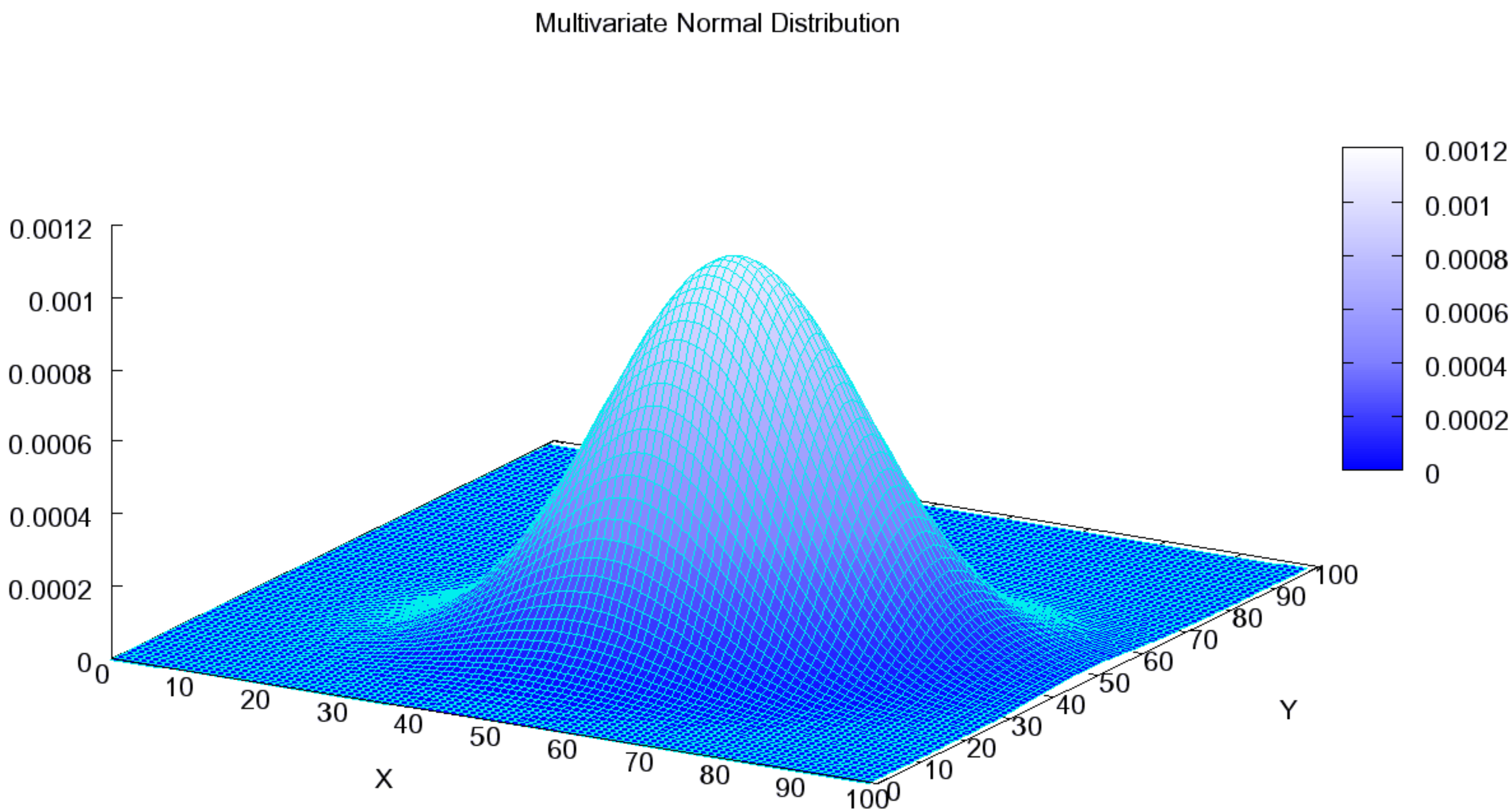
$$\boxed{\text{if } X \sim N(\boldsymbol{\mu}, \Sigma)\text{, then } X_i, X_j \text{ are statistically independent} \iff \text{CoVar}[X_i, X_j] = 0}$$

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

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

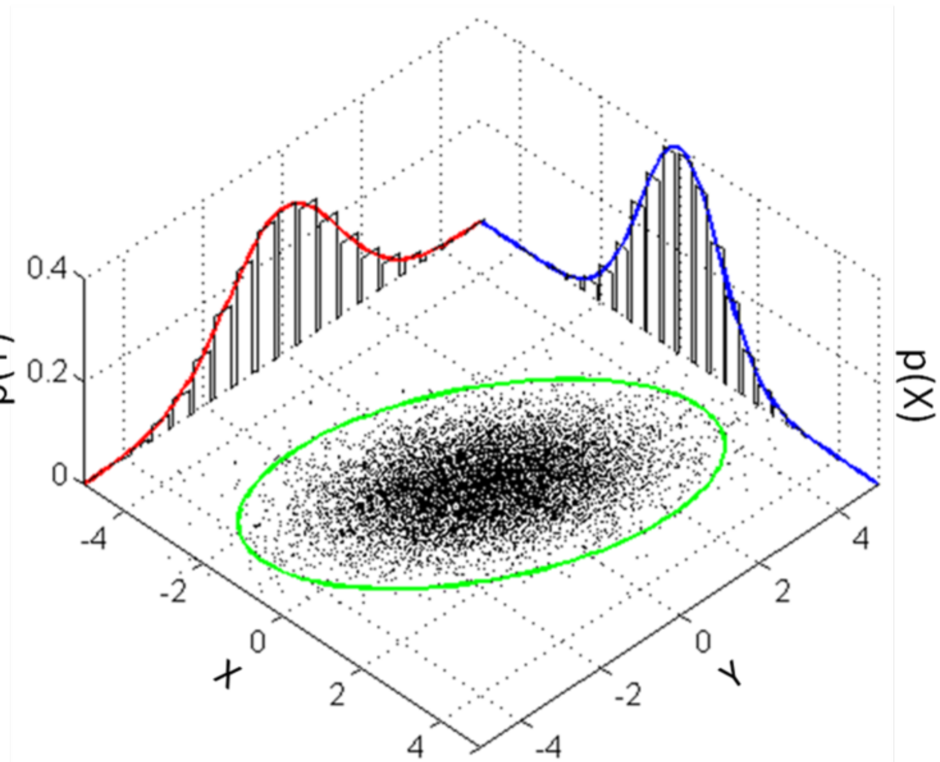
Bayesian decision theory

The Gaussian Distribution



Bayesian decision theory

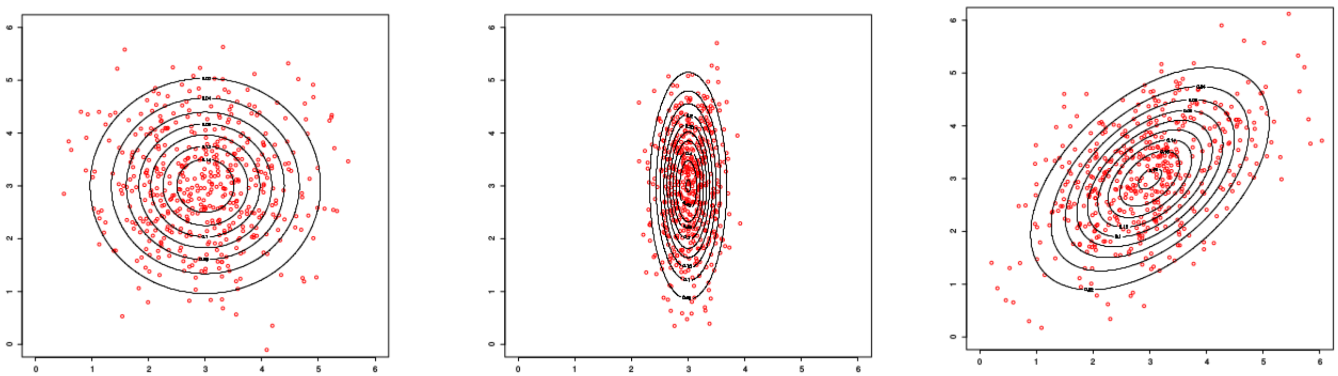
The Gaussian Distribution



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

Bayesian decision theory

The Gaussian Distribution



$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 0.1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

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

Bayesian decision theory

The Gaussian Distribution

- The quantity $d(\boldsymbol{x}, \boldsymbol{\mu}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}$ is called the **Mahalanobis distance**

- What is behind the choice of a multivariate Gaussian for a class?

  $\longrightarrow$ 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

Properties of the Gaussian Distribution

- **Simplified forms**: if the X_i are statistically independent, then $p(\boldsymbol{x}) = \prod_{i=1}^{d} p(x_i)$, 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 $\boldsymbol{\mu}$ 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

Machine Learning

Syllabus

1. Introduction to Machine Learning

2. Theoretical issues (I): regression

3. Linear regression and beyond

4. Theoretical issues (II): classification

5. Generative classifiers

6. Discriminative classifiers

7. Clustering

8. Learning with kernels (I): The SVM

9. Learning with kernels (II): Kernel functions

10. Learning with kernels (III): Other kernel methods

11. Artificial neural networks (I): the MLP

12. Artificial neural networks (II): the RBF

13. Ensemble methods: Random Forests

14. Advanced topics and frontiers