KTH Computer Science
and Communication

2E1395 - Pattern Recognition
Solutions to Introduction to Pattern Recognition, Chapter 2: Bayesian pattern classification

Preface

This document[1] is a solution manual for selected exercises from "Introduction to Pattern Recognition" by Arne Leijon. The notation followed in the text book will be fully respected here. A short review of the issue discussed in the corresponding chapter of the text book is given here as a reference. For a complete proof of these results and for the problem text refer to the text book.

problem definition and optimization

This chapter of the text book generalizes the problem of classification introduced in the previous chapter. Extensions are:

- more than two categories are considered
- more than one signal features can be employed
- the performance criterion is generalized

The new scheme of the classifier is depicted in fig. 1 taken from the text book. All the elements

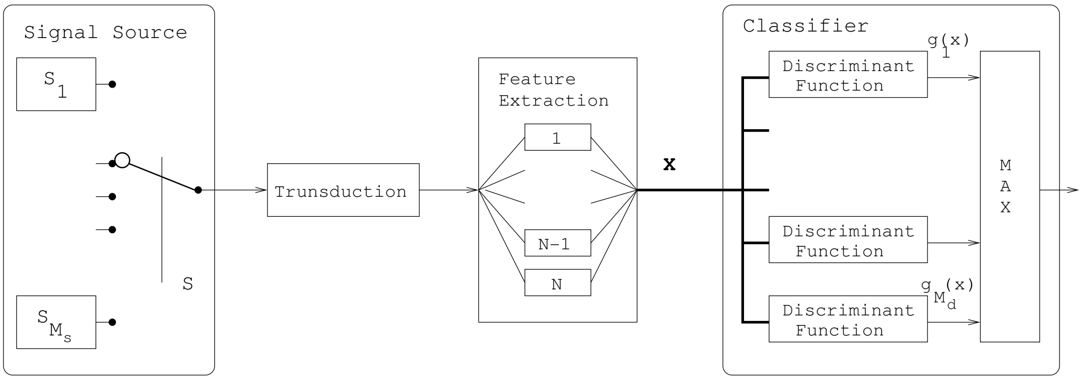

Figure 1. General signal classification

[1] This work comes under the terms of the Creative Commons © BY-SA 2.0 license

of this classification system are described statistically: the *signal state* can take any value from the set $\{j = 1 \dots M_s\}$ with *a priori* probability $P_S(j)$. The *observation feature vector* $\mathbf{x}$ is the outcome of a random vector $\mathbf{X}$ whose distribution depends on the state S and can be written as $f_{\mathbf{X}|S}(\mathbf{x}|j)$ for each of the M_s possible states. The *decision rule* $D(\mathbf{x})$ makes use of the information given by the *a posteriori* probability $P_{S|\mathbf{X}}(j|\mathbf{x})$, obtained with the *Bayes Rule* as:

$$P_{S|\mathbf{X}}(j|\mathbf{x}) = \frac{f_{\mathbf{X}|S}(\mathbf{x}|j)P_S(j)}{\sum_{j=1}^{M_s} f_{\mathbf{X}|S}(\mathbf{x}|j)P_S(j)}$$

to perform some actions for any incoming observation vector $\mathbf{x}$. This decision mechanism is the result of an optimization process aimed at fulfilling a *performance criterion.* This criterion is defined by a *cost function* $L(D = i\,, S = j)$ that describes the loss the system is subjected to when it takes the decision $D = i$, being the source in the state $S = j$. Since all decisions are taken with regard to the observation vector $\mathbf{x}$, and this is only statistically related to the “true” state S of the source, we can predict (statistically) a *Conditional Expected Loss* or *Conditional Risk*:

$$R(D = i|\mathbf{x}) = \sum_{j=1}^{M_s} L(D = i\,, S = j)P_{S|\mathbf{X}}(j|\mathbf{x})$$

The *performance criterion* is hence the one that leads to the minimum risk and is called *Bayes Minimum-Risk decision rule*:

$$D(\mathbf{x}) = \arg\min_i R(D = i|\mathbf{x})$$

The last rule is proved to minimize the total expected loss $Q = E\left[R(D(\mathbf{X})|\mathbf{X})\right]$ over all possible outcomes of the random vector $\mathbf{X}$

Special cases

- if the decision is to guess the state of the source, and the loss function is

  $$L(D = i\,, S = j) = \begin{cases} 0, & \text{if } i = j \\ 1, & \text{otherwise} \end{cases}$$

  then the optimal decision rule introduced before can be simplified to the *Maximum a Posteriori decision rule* (MAP):

  $$D(\mathbf{x}) = \arg\max_i f_{\mathbf{X}|S}(\mathbf{x}|i)P_S(i)$$

- if the previous conditions are verified and the *a priori* probabilities are all the same ($P_S(j) = \frac{1}{M_s}$, for all j), then the resulting decision rule is called *Maximum Likelihood Decision Rule* (ML):

  $$D(\mathbf{x}) = \arg\max_i f_{\mathbf{X}|S}(\mathbf{x}|i)$$

In general any decision rule can be expressed in the form:

$$D(\mathbf{x}) = \arg\max_{i=1\dots M_d} g_i(\mathbf{x})$$

and the $g_i(\mathbf{x})$ are called *discriminant functions.*

Exercise 2.1

We observe two sequences and we know that one is generated by a human being and the other by a random-number generator of a computer.

$$x = \{1\ 2\ 3\ 4\ 5; 2\ 4\ 5\ 1\ 3\}$$

There are two possible states of the source:

$$S = \{1, 2\} = \{[h, c], [c, h]\}$$

Where c stands for computer and h for human being. The *a priori* probability of the states are equally distributed:

$$P_S(1) = P_S(2) = \frac{1}{2}$$

To continue the solution of this problem we have to formulate some assumptions:

- in the absence of other information it is reasonable to assume that the machine generates uniformly distributed numbers, and that any sequence of the kind considered in this example has the same probability of being generated:

$$P(\{1\ 2\ 3\ 4\ 5\}|c) = P(\{2\ 4\ 5\ 1\ 3\}|c) = q$$

- common sense experience (and perhaps psychological arguments) would suggest that the probability that a human being generates the sequence $\{1\ 2\ 3\ 4\ 5\}$ is higher than that of generating the sequence $\{2\ 4\ 5\ 1\ 3\}$. In symbols:

$$P(\{1\ 2\ 3\ 4\ 5\}|h) = p_1;\ P(\{2\ 4\ 5\ 1\ 3\}|h) = p_2;\ p_1 > p_2$$

Combining the events, and assuming that they are independent we can write:

$$P_{X|S}(x|1) = P\left(\{1\ 2\ 3\ 4\ 5; 2\ 4\ 5\ 1\ 3\}\,|[h, c]\right) = p_1 q$$

$$P_{X|S}(x|2) = P\left(\{1\ 2\ 3\ 4\ 5; 2\ 4\ 5\ 1\ 3\}\,|[c, h]\right) = q p_2$$

Applying Bayes' rule:

$$\begin{aligned} P_{S|X}(j|x) &= \frac{P_S(j)P_{X|S}(x|j)}{P_S(1)P_{X|S}(x|1) + P_S(2)P_{X|S}(x|2)} \\ &= \frac{\frac{1}{2}qp_j}{\frac{1}{2}q(p_1 + p_2)} = \frac{p_j}{p_1 + p_2} \end{aligned}$$

that can be read as the probability of the state j given the observation x. The optimal MAP guess about the source is equivalent to the maximum likelihood optimal guess:

$$S_{\text{opt}} = \arg\max_j P_{S|X}(j|x) = \arg\max_j \frac{p_j}{p_1 + p_2} = \arg\max_j p_j$$

According to our assumptions on the values of p_1 and p_2 the optimal guess is $S = 1$: the human being has most probably generated the sequence $\{1\ 2\ 3\ 4\ 5\}$ while the machine the sequence $\{2\ 4\ 5\ 1\ 3\}$.

Exercise 2.2

a) The minimum error probability criterion is achieved by considering the loss function:

$$L(D = i\, , S = j) = \left\{ \begin{array}{l} 1, i \neq j \\ 0, i = j \end{array} \right.$$

Since the *a priori* probabilities of the state are uniformly distributed, we are in the *Maximum likelihood* case: the decision rule is

$$D(x) = \arg\max_i f_{X|S}(x|i)$$

where

$$f_{X|S}(x|i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}$$

To simplify the decision rule I chose to maximize a monotone increasing function of the argument instead of the argument itself, for example taking the logarithm:

$$g_i(x) \propto \ln f_{X|S}(x|i) \propto -\frac{1}{2}\frac{(x-\mu_i)^2}{\sigma^2} \propto -(x-\mu_i)^2$$

where we simplify all the constant terms that don't affect the maximization process. Since the decision mechanism checks whether $g_1(x)$ is greater or smaller than $g_2(x)$, which are monotone functions of the argument x, this can be implemented by a simple threshold x_t with $g_1(x_t) = g_2(x_t)$. Substituting:

$$(x_t - \mu_1)^2 = (x_t - \mu_2)^2 \iff$$

$$x_t = \frac{\mu_1 + \mu_2}{2}$$

This result was predictable when considering that two Gaussian distributions with the same

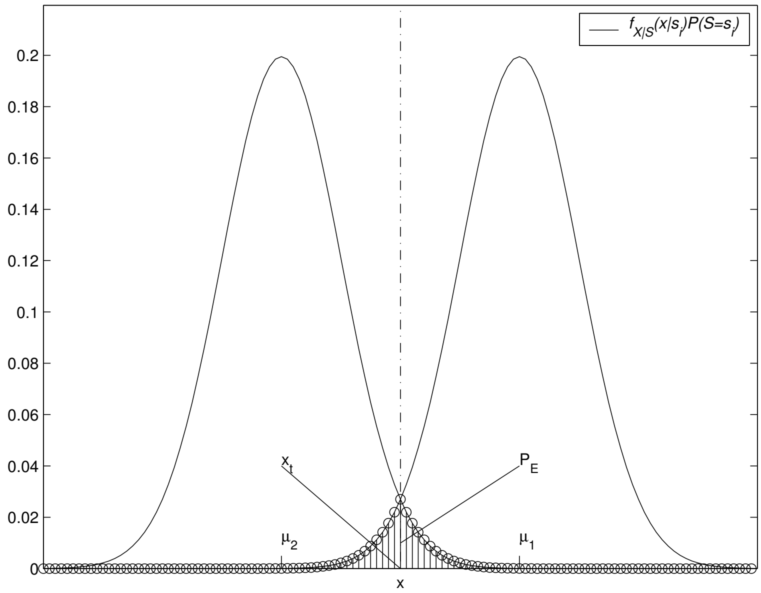

Figure 2.

variance intersect in the median point between their mean values, and that (as pointed out more that once in chapter 1) the optimal threshold with regard to the minimum error probability corresponds to this intersection point.

b) As previously explained (see chapter 1 and fig. 2), if we assume $\mu_1 > \mu_2$ as in this case, the total probability of error is given by:

$$P_E = P_S(1)\int_{-\infty}^{x_t} f_{X|S}(x|1)dx + P_S(2)\int_{x_t}^{+\infty} f_{X|S}(x|2)dx$$

Substituting the given values, and given the symmetry:

$$\begin{aligned} P_E &= \frac{1}{2}\int_{-\infty}^{0} N(2,1)dx + \frac{1}{2}\int_{0}^{+\infty} N(-2,1)dx \\ &= 2\frac{1}{2}\int_{-\infty}^{0} N(2,1)dx = [1-\Phi(2)] = 0.023 \end{aligned}$$

For the numerical result refer to BETA[2] pag. 405.

Exercise 2.3

We have a signal source with N possible outcomes $j = 1, N$, governed by the known probabilities $P_S(j)$. There are $N+1$ possible decisions ($D = j; j = 1, N$ if the state was $S = j$ and $D = N+1$ "no decision"). The cost function is given in the exercise text as:

$$L(D{=}i, S{=}j) = \begin{cases} 0, & i = j \qquad\quad j = 1...N \\ r, & i = N+1, \quad j = 1...N \\ c, & \text{otherwise} \end{cases}$$

that sets the cost to 0 if the decision was correct, to c if it was taken, but incorrect and to r if it was rejected.

a) The expected cost is by definition $R(D{=}i|x) = \sum_j L(D{=}i\,, S{=}j)P_{S|X}(j|x)$. To compute this we consider two different cases:

1) the decision is taken ($i \neq N+1$);

$$R(D{=}i|x) = c\sum_{j\neq i} P_{S|X}(j|x) = c(1 - P_{S|X}(i|x))$$

The last equality is true because $\sum_j P_{S|X}(j|x) = 1$. In this case we know that the minimum expected cost is achieved with the following decision function:

$$D(x) = \arg\max_i [cP_{S|X}(i|x)]$$

2) the decision is not taken ($i = N+1$).

$$R(D{=}N+1|x) = r\sum_{j=1}^{N} P_{S|X}(j|x) = r$$

and, since $i = N+1$,

$$D(x) = \text{“no decision”}$$

[2] Beta, mathematics handbook, Studentlitteratur

The last thing to check is which is the best choice between the first and second case for each x. The decision will not be rejected if $\forall i \neq N+1$:

$$R(D{=}i|x) \leq R(D{=}N+1|x) \iff$$

$$c[1 - P_{S|X}(i|x)] \leq r \iff$$

$$P_{S|X}(i|x) \geq 1 - \frac{r}{c}, \quad \forall i$$

This way we have proved that the decision function $D(x)$ proposed in the example is optimal.

b) If $r = 0$ then rejecting a decision is free of cost, if $c \to \infty$ then a wrong decision would be enormously costly. In both cases it's never worth risking an error $\Rightarrow d(x)$ will always reject the decision. From a mathematical point of view, the condition to accept a decision (no rejection) becomes

$$P_{S|X}(i|x) \geq 1$$

that is never verified unless the observation x can only be generated by the source state $S = i$ (the equality holds), and there is no doubt on the decision to take.

c) If $r > c$, rejecting a decision will always be more expensive than trying one $\Rightarrow$ no decision will be rejected, from the mathematical point of view, the condition is that the probability of the state given the observation be grater than a negative number, which is always verified by probabilities:

$$P_{S|X}(i|x) \geq -\epsilon; \quad \epsilon > 0$$

d) If $i = 1, N$ then the discriminant function correspond to the one in point a) which we know to be optimal. We have to prove that the choice of the decision $N+1$ leads to the same condition as in point a). Decision $N+1$ is chosen ($\arg\max(.) = N+1$) if and only if $\forall i = 1, N$,

$$\begin{aligned} g_{N+1}(x) = (1 - \frac{r}{c}) \sum_{j=1}^{N} f_{X|S}(x|j) P_S(j) \quad &> \quad g_i(x) = f_{X|S}(x|i) P_S(i) \iff \\ \frac{f_{X|S}(x|i) P_S(i)}{\sum_{j=1}^{N} f_{X|S}(x|j) P_S(j)} \quad &< \quad 1 - \frac{r}{c}, \quad \forall i = 1, N \end{aligned}$$

applying Bayes rule,

$$P_{S|X}(i|x) < 1 - \frac{r}{c}, \ \ \forall i = 1, N$$

as we wanted to prove.

e) The three functions $g_i(x)$ are:

$$\begin{aligned} g_1(x) &= P_S(1) f_{X|S}(x|1) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}} \\ g_2(x) &= P_S(2) f_{X|S}(x|2) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+1)^2}{2}} \\ g_3(x) &= (1 - \frac{r}{c})[g_1(x) + g_2(x)] = \frac{3}{4}[g_1(x) + g_2(x)] \end{aligned}$$

These functions are plotted in fig. 3.

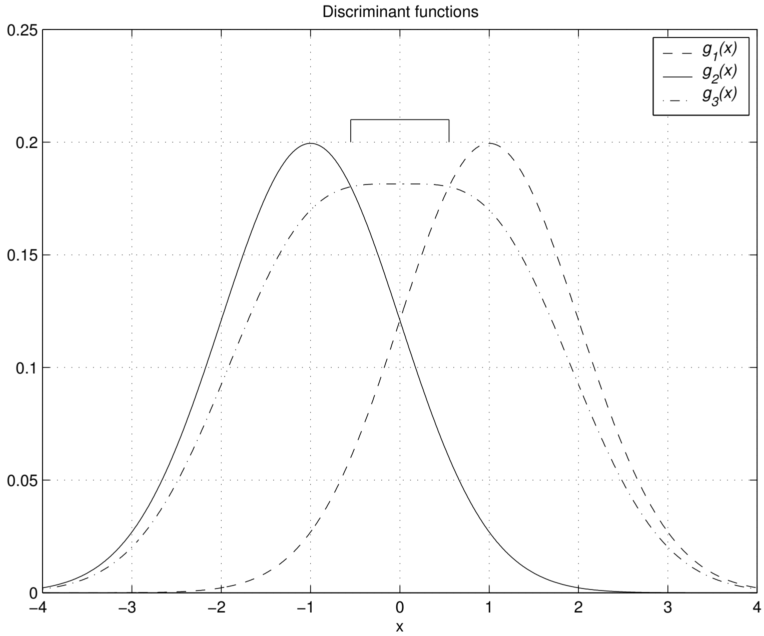

Figure 3.

f) The decision $D(x) = 3$ is taken if and only if $g_3(x) > \max[g_1(x), g_2(x)]$, which is verified in the region indicated in the figure by a ⊓ sign. The total probability of rejection is

$$\begin{aligned} P_D(3) &= P_{X|S}(-x_0 < x < x_0|1)P_S(1) + P_{X|S}(-x_0 < x < x_0|2)P_S(2) \\ &= \frac{1}{2}\int_{-x_0}^{x_0} N(1,1)dx + \frac{1}{2}\int_{-x_0}^{x_0} N(-1,1)dx \end{aligned}$$

Since the problem is fully symmetric the two terms are equal and

$$\begin{aligned} P_D(3) &= \int_{-x_0}^{x_0} N(-1,1)dx \\ &= \Phi(x_0+1) - \Phi(-x_0+1) \end{aligned}$$

Last thing to do is to find the value of x_0 i.e. the value at which $g_3(x) = g_1(x)$:

$$\begin{aligned} g_3(x) &= g_1(x) \iff \\ (1-\frac{r}{c})[\frac{1}{2}N(1,1) + \frac{1}{2}N(-1,1)] &= \frac{1}{2}N(1,1) \\ (1-\frac{r}{c})\frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2+1}{2}}[e^{-x}+e^{x}] &= \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2+1}{2}}e^{x} \\ e^{2x} = \frac{c-r}{r} &\iff x = \log\sqrt{\frac{c-r}{r}} \end{aligned}$$

With the values specified by the problem $x_0 = \ln\sqrt{3}$. The total probability of rejection is then

$$\begin{aligned} P_R &= P_D(3) = \Phi(\ln\sqrt{3}+1) - \Phi(-\ln\sqrt{3}+1) \\ &= \Phi(1.549) - \Phi(0.45) \simeq 0.27 \end{aligned}$$

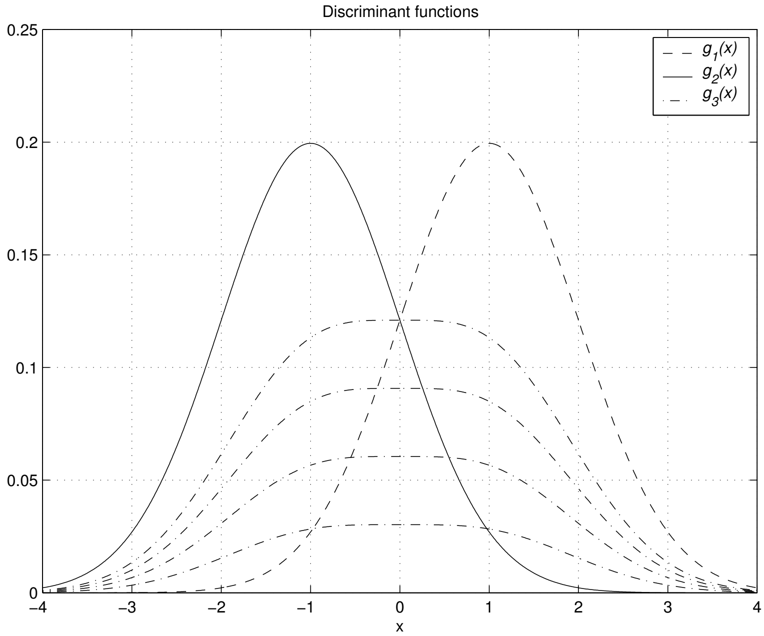

Figure 4. Rejection is never considered

Where the function Φ is tabled in BETA[3] pag. 405.

The decision g_3 (rejection) is never chosen if $g_3(x) < \max[g_1(x), g_2(x)], \quad \forall x \in R$. This is guaranteed if $g_3(0) < g_1(0)$ as is clear looking at fig 4. Since $g_3(0) = (1 - \frac{r}{c})[g_1(0) + g_2(0)] = 2\ (1 - \frac{r}{c})\ g_1(0)$ for the symmetry, then rejection is never considered if

$$r > \frac{c}{2}$$

Intersection of Two Gaussian Distributions The problem of finding for which x, $P_{XS}(x, 0) <> P_{XS}(x, 1)$ is common in the exercises seen so far. This corresponds to finding where $P_S(0)P_{X|S}(x, 0) <> P_S(1)P_{X|S}(x, 1)$. In case of Gaussian distributions $N(\mu_i, \sigma_i)$ if we set $p_i = P_S(i)$ with $i = 0, 1$ the intersection points are:

$$x_{1,2} = \frac{\sigma_1^2\mu_0 - \sigma_0^2\mu_1 \pm \sigma_0\sigma_1\sqrt{(\mu_0 - \mu_1)^2 + 2(\sigma_1^2 - \sigma_0^2)\ln\left(\frac{p_0\sigma_1}{p_1\sigma_0}\right)}}{\sigma_1^2 - \sigma_0^2}$$

In the special case in which $\sigma_0 = \sigma_1 = \sigma$ that is the most interesting in our case (same noise that affects both observations), there is at most one finite solution, and a single threshold on the value of x is a solution to the problem described before:

$$x_1 = \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2}{(\mu_0 - \mu_1)} \ln\left(\frac{p_1}{p_0}\right)$$

If the probability of the source is equal ($p_0 = p_1 = 1/2$) then

$$x_1 = \frac{\mu_1 + \mu_0}{2}$$

[3] Beta, mathematics handbook, Studentlitteratur

... or if the means are opposite to each other ($\mu_0 = -\mu$ and $\mu_1 = \mu$)

$$x_1 = \frac{\sigma^2}{2\mu} \ln \left(\frac{p_1}{p_0} \right)$$