

# Théorie Bayésienne de la décision

Stéphane Canu

scanu@insa-rouen.fr

<http://asi.insa-rouen.fr/~scanu/>

6 janvier 2003

## 1 Le problème de détection

Un problème typique de décision se pose lors de l'observation d'un patient par un médecin. Le patient est malade ou non. Le praticien lui peut décider de la soigner, de ne pas le soigner, d'effectuer des examens complémentaires, ou de l'adresser à un collègue spacialiste parceque décidément ce qu'il observe ne ressemble à rien de ce qu'il connaît. On voit sur cet exemple qu'il convient de distinguer les situations possibles (le patient est malade ou non) des actions envisagées par le praticien (rien faire, soigner, examens complémentaires ou recours à un collègue). En général nous formulerons le problème de détection de la manière suivante :

- on considère deux sources  $S_1$  et  $S_2$ .
- une observation  $x$ , réalisation d'une variable aléatoire continue  $X$ ,
- quatre actions (décisions) : ( $a_1$ ) décider que c'est la source  $S_1$  qui a émise  $x$ , ( $a_2$ ) décider que c'est la source  $S_2$ , ( $a_3$ ) décider que c'est l'une ou l'autre (rejet d'ambiguïté), ( $a_4$ ) décider que c'est aucune des deux (rejet de distance).

La théorie statistique de la décision consiste à considérer que les sources  $S$  et les observations  $X$  sont des variables aléatoires :

- la source  $S$  qui émet le signal  $X$  est tirée au hasard suivant une loi de bernouilli<sup>1</sup>.  $S$  est donc une variable aléatoire discrète avec :  $\mathbb{P}(S = S_1)$  et donc  $\mathbb{P}(S = S_2) = 1 - \mathbb{P}(S = S_1)$ .
- sachant  $S_1$ , la variable aléatoire  $X$  est distribuée suivant une loi normale  $\mathcal{N}(\mu_1, \sigma_1)$
- s'il est émis par la source  $S_2$ , le signal  $X$  est distribué suivant une loi normale  $\mathcal{N}(\mu_2, \sigma_2)$

Afin de construire une règle de décision, nous avons besoin de connaître les couts associés à chaque couple (source-action) :

source/action	$a_1$	$a_2$	$a_3$	$a_4$
$s_1$	0	1	0,4	0,75
$s_2$	1	0	0,2	0,75

Tableau des couts  $l_{ij}$

Ensuite on calcule le risque associé à chaque décision, et on choisi l'action qui présente le plus petit risque. Le risque dépend de la source qui à émis le signal  $x$ . En considérant que les couts  $l_{ij}$  sont des variables aléatoires dépendant, pour une action donnée des diffétents sources, le risque peut s'interpréter comme le cout

<sup>1</sup>on code 0 pour la source  $S_1$  et 1 pour la source  $S_2$

moyen  $R(a, S) = \mathbb{E}(l_{s,a})$ . Il s'écrit de la manière suivante :

$$\begin{aligned} R(a_1, x) &= l_{11}\mathbb{P}(S_1|\mathbf{x}) + l_{21}\mathbb{P}(S_2|\mathbf{x}) \\ R(a_2, x) &= l_{12}\mathbb{P}(S_1|\mathbf{x}) + l_{22}\mathbb{P}(S_2|\mathbf{x}) \\ R(a_3, x) &= l_{13}\mathbb{P}(S_1|\mathbf{x}) + l_{23}\mathbb{P}(S_2|\mathbf{x}) \\ R(a_4, x) &= l_{14}\mathbb{P}(S_1|\mathbf{x}) + l_{24}\mathbb{P}(S_2|\mathbf{x}) \end{aligned}$$

soit, en posant  $p = \mathbb{P}(S_1|\mathbf{x})$  et  $1 - p = \mathbb{P}(S_2|\mathbf{x})$  :

$$\begin{aligned} R(a_1, x) &= 1 - p \\ R(a_2, x) &= p \\ R(a_3, x) &= 0,4p + 0,2(1 - p) = 0,2 + 0,2p \\ R(a_4, x) &= 0,75 \end{aligned}$$

Les risques des différentes actions sont représentés figure 1 en fonction de  $p$ .

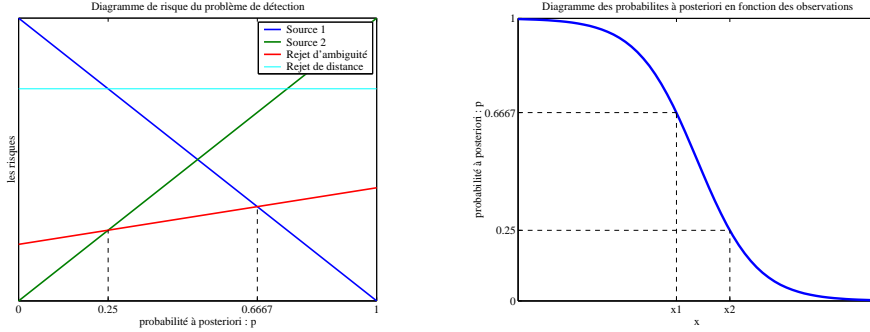


FIG. 1 – Illustration des risques associés aux différentes actions possibles et des probabilité a posteriori associés.

On obtient alors la règle de décision suivante :

$$D(x) = \begin{cases} a_1 & \text{si } p > \frac{2}{3} \\ a_2 & \text{si } p < \frac{1}{4} \\ a_3 & \text{sinon} \end{cases}$$

Dans ce cadre, le rejet de distance  $a_4$  n'est jamais décidé.

Une manière de résoudre ce problème consiste à décider  $a_4$  (rejet de distance) lorsque la densité de l'observation  $x$  est inférieure à un seuil  $\beta$  ( $f(x) < \beta$ ). Le règle de décision complète dépend de  $p$  et de  $f(x)$ . Elle s'écrit :

$$D(x) = \begin{cases} a_4 & \text{si } f(x) < \beta \\ a_1 & \text{sinon et si } p > \frac{2}{3} \\ a_2 & \text{sinon et si } p < \frac{1}{4} \\ a_3 & \text{sinon} \end{cases}$$

Pour réécrire la règle de décision en fonction de  $x$  il faut intégrer cette règle avec la valeur des probabilités a posteriori (figure 1). On obtient alors la règle suivante :

$$D(x) = \begin{cases} a_4 & \text{si } f(x) < \beta \\ a_1 & \text{sinon et si } x < x_1 \\ a_2 & \text{sinon et si } x > x_2 \\ a_3 & \text{sinon} \end{cases}$$

En conclusion, la règle de décision s'écrit en deux parties. Les couts  $l_{ij}$  associés aux couples sources-actions permettent d'établir la dépendance entre la fonction

de décision et la probabilité a posteriori. Cette dernière s'obtient à partir des loi a priori et des vraisemblance conditionnelles ou évidences du problème. Donc c'est à partir de  $\mathbb{P}$ ,  $f_i$  et  $l_{ij}$  que l'on peut déduire  $D$ . Souvent la difficulté vient( de la complexité des évidences  $f_i$ .

## 2 Cadre théorique

Le problème de détection se formule à partir des quantités suivantes :

- un ensemble de classes (sources)  $\mathcal{S} = \{s_1, s_2, \dots, s_K\}$
- un ensemble d'actions  $\mathcal{A} = \{a_1, a_2, \dots, a_L\}$
- un vecteur forme  $x \in \mathbb{R}^d$
- fonction cout

$$\begin{aligned} l : \mathcal{A} \times \mathcal{S} &\longrightarrow \mathbb{R} \\ (a, s) &\longmapsto l(a, s) \end{aligned}$$

Le problème de détection consiste, après avoir observé  $\mathbf{x}$ , à décider la meilleure action. Nous allons maintenant reformuler ce problème dans un cadre statistique.

Pour ce faire, on distingue le cas où le vecteur d'observation est une variable aléatoire continue et le cas où c'est une variable aléatoire discrète. Considérons le cas continu.

- la loi de  $S$  est appelée loi a priori et notée  $\mathbb{P}(S_i)$ . On considère que, sans autre information, la source est tirée au hasard suivant une loi connue.
- les lois conditionnelles de  $X$  sachant  $S$  sont appelées les vraisemblances, dont les densité sont notées  $f(\mathbf{x}|S_i)$  ou  $f_i(\mathbf{x})$ ,
- on recherche les lois de  $A$  conditionnellement à  $X$  que l'on appelle les lois a posteriori et que l'on note  $\mathbb{P}(s_j|\mathbf{x})$ .

Le problème qui nous intéresse pour l'instant est le suivant : étant donné les probabilités a priori  $\mathbb{P}(S_i)$ , les vraisemblances  $f(\mathbf{x}|S_i)$  et les couts  $l_{ij}$ , comment déterminer la meilleure règle de décision. On recherche donc une fonction appelée règle de décision :

$$\begin{aligned} D : \mathbb{R}^d &\longrightarrow \mathcal{A} \\ \mathbf{x} &\longmapsto D(\mathbf{x}) \end{aligned}$$

Pour se faire, il faut définir le risque conditionnel d'une action après avoir observé  $\mathbf{x}$  :

$$R(a_i, \mathbf{x}) = \sum_{j=1}^K l(a_i, s_j) \mathbb{P}(s_j|\mathbf{x})$$

Ce risque est tout a fait analogue à un cout moyen.

Risque moyen associé à une règle de décision  $D$  (moyenne des couts par rapport à toutes les sources et de nouveau moyenne par rapport à toutes les occurrences des observations  $\mathbf{x}$ ) :

$$R(D) = \mathbb{E}(R(D(X), X)) = \int R(D(\mathbf{x}), \mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

Le problème de Bayes consiste à trouver la règle de risque minimum. La solution de ce problème, la règle qui donne le plus petit risque moyen est appelée règle de Bayes et est définie par :

$$D_b(x) = \underset{D}{\operatorname{argmin}} R(D)$$

C'est aussi la règle qui minimise les risques conditionnels :

$$D_b(x) = \underset{i \in \{1, L\}}{\operatorname{argmin}} R(a_i, \mathbf{x})$$

Démonstration : l'espérance est minimisée si l'on minimise le risque conditionnel  $R(D(x), x)$  pour tous les  $\mathbf{x}$ . Ce qui revient, à  $\mathbf{x}$  fixé à choisir l'action de risque conditionnel minimum.

Le risque associé  $R_b = R(D_b)$  est appelé risque de Bayes et représente le plus petit risque atteignable.

### 3 Discriminaton à deux classes : le problème de détection

On considère maintenant le problème de détection à deux classes ( $K = L = 2$ ). Les risques associés à chacune des décisions sont donc :

$$\begin{aligned} R(a_1, \mathbf{x}) &= l_{11}\mathbb{P}(S_1|\mathbf{x}) + l_{12}\mathbb{P}(S_2|\mathbf{x}) \\ R(a_2, \mathbf{x}) &= l_{21}\mathbb{P}(S_1|\mathbf{x}) + l_{22}\mathbb{P}(S_2|\mathbf{x}) \end{aligned}$$

D'après la règle de Bayes, on décide  $a_1$  lorsque :

$$\begin{aligned} &l_{11}\mathbb{P}(S_1|\mathbf{x}) + l_{12}\mathbb{P}(S_2|\mathbf{x}) < l_{21}\mathbb{P}(S_1|\mathbf{x}) + l_{22}\mathbb{P}(S_2|\mathbf{x}) \\ \Leftrightarrow &(l_{11} - l_{21})\mathbb{P}(S_1|\mathbf{x}) < (l_{12} - l_{22})\mathbb{P}(S_2|\mathbf{x}) \\ \Leftrightarrow &(l_{11} - l_{21})f(\mathbf{x}|S_1)\mathbb{P}(S_1) < (l_{12} - l_{22})f(\mathbf{x}|S_2)\mathbb{P}(S_2) \end{aligned}$$

soit, en faisant l'hypothèse raisonnable que  $l_{11} < l_{21}$  :

$$\frac{f(\mathbf{x}|S_1)}{f(\mathbf{x}|S_2)} > \frac{(l_{12} - l_{22})\mathbb{P}(S_2)}{(l_{11} - l_{21})\mathbb{P}(S_1)}$$

Cette formulation présente l'avantage de faire apparaître le rapport de vraisemblance qui est une statistique habituelle des décisions.

Pratiquement on est souvent amené à considérer le logarithme de ce rapport et donc à décider  $S_1$  si :

$$\mathcal{L}(x) = \log\left(\frac{f(\mathbf{x}|S_1)}{f(\mathbf{x}|S_2)}\right) > k$$

avec

$$k = \log\left(\frac{(l_{12} - l_{22})\mathbb{P}(S_2)}{(l_{11} - l_{21})\mathbb{P}(S_1)}\right)$$

La règle de décision associée est alors :

$$D(\mathbf{x}) = \begin{cases} a_1 & \text{si } \mathcal{L}(x) \geq k \\ a_2 & \text{si } \mathcal{L}(x) < k \end{cases}$$

### 4 Discriminaton à deux classes avec rejet

On distingue deux types de rejets. Soit une observation est ambiguë, non loin de la frontière de décision. Elle une classe ou une autre. Soit elle se situe loin de toute observation précédente, elle est alors ni une classe, ni une autre. Dans le premier cas, on parle de rejet d'ambiguïté, dans le second de rejet de distance. Si le premier cas s'accorde bien avec le théorème bayésien, il n'en n'est pas de même avec le second.

#### 4.1 Rejet d'ambiguïté

On définit une action  $a_3$  qui consiste à ne pas décider entre les sources. Dans ce cas les risques sont les suivants :

$$\begin{aligned} R(a_1, \mathbf{x}) &= l_{11}\mathbb{P}(S_1|\mathbf{x}) + l_{12}\mathbb{P}(S_2|\mathbf{x}) \\ R(a_2, \mathbf{x}) &= l_{21}\mathbb{P}(S_1|\mathbf{x}) + l_{22}\mathbb{P}(S_2|\mathbf{x}) \\ R(a_3, \mathbf{x}) &= l_{31}\mathbb{P}(S_1|\mathbf{x}) + l_{32}\mathbb{P}(S_2|\mathbf{x}) \end{aligned}$$

Supposons que les coûts  $l_{11} = l_{22} = 0$ ,  $l_{21} = l_{12} = 1$  et  $l_{31} = l_{32} = \alpha$ . Les risques sont alors :

$$\begin{aligned} R(a_1, \mathbf{x}) &= \mathbb{P}(S_2|\mathbf{x}) = 1 - \mathbb{P}(S_1|\mathbf{x}) \\ R(a_2, \mathbf{x}) &= \mathbb{P}(S_1|\mathbf{x}) = 1 - \mathbb{P}(S_2|\mathbf{x}) \\ R(a_3, \mathbf{x}) &= \alpha \end{aligned}$$

La règle de décision associée est la suivante :

$$D(\mathbf{x}) = \begin{cases} a_1 & \text{si } \mathbb{P}(S_1|\mathbf{x}) \geq \mathbb{P}(S_2|\mathbf{x}) \text{ et } \mathbb{P}(S_1|\mathbf{x}) > 1 - \alpha \\ a_2 & \text{si } \mathbb{P}(S_1|\mathbf{x}) < \mathbb{P}(S_2|\mathbf{x}) \text{ et } \mathbb{P}(S_2|\mathbf{x}) > 1 - \alpha \\ a_3 & \text{sinon} \end{cases}$$

Si  $\alpha$  est trop grand, le rejet d'ambiguïté  $a_3$  n'est jamais décidé. En effet, puisque  $\mathbb{P}(S_1|\mathbf{x}) + \mathbb{P}(S_2|\mathbf{x}) = 1$ , la plus grande des deux probabilités est nécessairement supérieure à  $1/2$ . Donc si  $\alpha > 1/2$ ,  $a_3$  n'est jamais décidé. Lorsque  $\alpha$  varie entre  $1/2$  et  $0$ , le taux de rejet varie entre  $0$  et  $100\%$ . La règle de décision se simplifie ainsi :

$$D(\mathbf{x}) = \begin{cases} a_1 & \text{si } \mathbb{P}(S_1|\mathbf{x}) > 1 - \alpha \\ a_2 & \text{si } \mathbb{P}(S_2|\mathbf{x}) > 1 - \alpha \\ a_3 & \text{sinon} \end{cases}$$

Cette règle de décision peut se reformuler en faisant apparaître le rapport de vraisemblance de la manière suivante :

$$\begin{aligned} \mathbb{P}(S_1|\mathbf{x}) > 1 - \alpha &\Leftrightarrow \frac{f(\mathbf{x}|S_1)\mathbb{P}(S_1)}{f(\mathbf{x})} > 1 - \alpha \\ &\Leftrightarrow f(\mathbf{x}|S_1)\mathbb{P}(S_1) > (1 - \alpha)(f(\mathbf{x}|S_1)\mathbb{P}(S_1) + f(\mathbf{x}|S_2)\mathbb{P}(S_2)) \\ &\Leftrightarrow f(\mathbf{x}|S_1)(\mathbb{P}(S_1) - (1 - \alpha)\mathbb{P}(S_1)) > (1 - \alpha)f(\mathbf{x}|S_2)\mathbb{P}(S_2) \\ &\Leftrightarrow \frac{f(\mathbf{x}|S_1)}{f(\mathbf{x}|S_2)} > \frac{(1 - \alpha)(1 - \mathbb{P}(S_1))}{\alpha\mathbb{P}(S_1)} \end{aligned}$$

et donc si  $\mathcal{L}(\mathbf{x}) = \log\left(\frac{f(\mathbf{x}|S_1)}{f(\mathbf{x}|S_2)}\right)$  la règle de décision s'écrit :

$$D(\mathbf{x}) = \begin{cases} a_1 & \text{si } \mathcal{L}(\mathbf{x}) > k_1 \\ a_2 & \text{si } \mathcal{L}(\mathbf{x}) < k_2 \\ a_3 & \text{sinon} \end{cases}$$

avec  $k_1 = \log\left(\frac{(1-\alpha)(1-\mathbb{P}(S_1))}{\alpha\mathbb{P}(S_1)}\right)$  et  $k_2 = \log\left(\frac{\alpha\mathbb{P}(S_1)}{(1-\alpha)(1-\mathbb{P}(S_1))}\right)$  (démonstration laissée en exercice).

## 4.2 Rejet de distance hors du cadre bayésien

Hors du cadre bayésien, on peut décider qu'une observation  $\mathbf{x}$  est à rejeter en distance s'il est peu vraisemblable qu'elle appartienne à l'une des deux sources identifiées. Ce peut être justement l'apparition d'une nouvelle source ou, en diagnostic, l'apparition d'un nouveau mode de fonctionnement. Un test statistique permettant de décider si une observation a été générée par une loi se construit à l'aide d'un seuil  $\beta$ . L'appartenance de  $\mathbf{x}$  à la distribution de densité  $f$  sera rejetée lorsque  $f(\mathbf{x}) < \beta$ . On obtient alors la règle de décision suivante :

$$D(\mathbf{x}) = \begin{cases} a_4 & \text{si } f(\mathbf{x}) < \beta \\ a_1 & \text{sinon et si } \mathcal{L}(\mathbf{x}) > k_1 \\ a_2 & \text{sinon et si } \mathcal{L}(\mathbf{x}) < k_2 \\ a_3 & \text{sinon} \end{cases}$$

### 4.3 Rejet de distance dans le cadre bayésien

Pour rester dans le cadre bayésien, il faut se donner une nouvelle source  $S_3$  qui correspond à tout ce qui n'est ni  $S_1$  ni  $S_2$ . Les risques sont les suivants :

$$\begin{aligned} R(a_1, \mathbf{x}) &= l_{11}\mathbb{P}(S_1|\mathbf{x}) + l_{12}\mathbb{P}(S_2|\mathbf{x}) + l_{13}\mathbb{P}(S_3|\mathbf{x}) \\ R(a_2, \mathbf{x}) &= l_{21}\mathbb{P}(S_1|\mathbf{x}) + l_{22}\mathbb{P}(S_2|\mathbf{x}) + l_{23}\mathbb{P}(S_3|\mathbf{x}) \\ R(a_3, \mathbf{x}) &= l_{31}\mathbb{P}(S_1|\mathbf{x}) + l_{32}\mathbb{P}(S_2|\mathbf{x}) + l_{33}\mathbb{P}(S_3|\mathbf{x}) \\ R(a_4, \mathbf{x}) &= l_{41}\mathbb{P}(S_1|\mathbf{x}) + l_{42}\mathbb{P}(S_2|\mathbf{x}) + l_{43}\mathbb{P}(S_3|\mathbf{x}) \end{aligned}$$

Comme précédemment, avec  $\alpha$  comme cout de rejet et les couts 0/1 pour les autres classes on obtient :

$$\begin{aligned} R(a_1, \mathbf{x}) &= 1 - \mathbb{P}(S_1|\mathbf{x}) \\ R(a_2, \mathbf{x}) &= 1 - \mathbb{P}(S_2|\mathbf{x}) \\ R(a_3, \mathbf{x}) &= \alpha \\ R(a_4, \mathbf{x}) &= 1 - \mathbb{P}(S_3|\mathbf{x}) \end{aligned}$$

Si  $\mathbb{P}(\mathbf{x}|S_3)$  suit une loi uniforme sur le domaine des possibles, cette règle de décision revient à la précédente.

## 5 Stratégies de décision

### 5.1 Règle du maximum a posteriori

Une manière typique d'aborder le problème en faisant l'impasse sur des couts difficiles à définir consiste à travailler avec des couts 0/1. Dans ce cas, on considère autant d'actions que de classes et :

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

Dans ce cas :

$$R(\mathbf{a}_i, \mathbf{x}) = \sum_{j=1, j \neq i}^K \mathbb{P}(s_j|\mathbf{x}) = 1 - \mathbb{P}(s_i|\mathbf{x})$$

La décision de cout minimale (la règle de Bayes) est donc aussi celle qui maximise la probabilité a posteriori des classes. C'est la règle dite du maximum a posteriori.

Dans le cas de deux classes, pour toute forme  $\mathbf{x}$ , et en considérant les sources codées 0 ou 1 suivant une loi de Bernoulli, la probabilité *a posteriori* d'une source est donnée par la fonction suivante :

$$r(\mathbf{x}) = \mathbb{P}(S = 1|\mathbf{X} = \mathbf{x}) = \mathbb{E}(S|\mathbf{X} = \mathbf{x})$$

La règle de Bayes s'écrit alors :

$$D_b(\mathbf{x}) = \begin{cases} 0 & \text{si } r(\mathbf{x}) > \frac{1}{2} \\ 1 & \text{sinon} \end{cases} \quad (1)$$

2.1 par exemple) Puisque  $\mathbb{P}(S_1|\mathbf{x}) + \mathbb{P}(S_2|\mathbf{x}) = 1$  la plus grande des deux est nécessairement supérieure à  $\frac{1}{2}$ .

Souvent, pour des raisons pratiques, les classes sont codées sur  $\{-1, 1\}$  au lieu de  $\{0, 1\}$ . Dans ce cas, la règle de Bayes devient :  $D_b(\mathbf{x}) = \text{signe}(f(\mathbf{x}))$  avec  $f(\mathbf{x}) = 2r(\mathbf{x}) - 1$ .

## 5.2 Critère « minmax »

Comment faire lorsque les probabilités a priori  $p$  nous sont inconnues ? Une technique permettant de faire face à cette situation consiste à envisager de choisir la règle de décision qui donnera le meilleur résultat *dans le pire des cas*. La stratégie de choix de la règle de décision est alors la suivante : faire apparaître le risque comme une fonction de  $p$  puis choisir celui dont le maximum est le plus petit, d'où le nom de cette stratégie : le minmax.

Le risque moyen s'écrit alors comme une fonction affine de  $p$  :

$$\begin{aligned} R(D) &= R(D = S_1) + R(D = S_2) \\ &= \int_{\mathbf{x} \in R_1} l_{11} \mathbb{P}(S_1 | \mathbf{x}) + l_{12} \mathbb{P}(S_2 | \mathbf{x}) d\mathbf{x} + \int_{\mathbf{x} \in R_2} l_{21} \mathbb{P}(S_1 | \mathbf{x}) + l_{22} \mathbb{P}(S_2 | \mathbf{x}) d\mathbf{x} \\ &= Ap + B \end{aligned}$$

avec

$$\begin{aligned} A &= l_{11} - l_{22} - (l_{21} - l_{11}) \int_{\mathbf{x} \in R_2} f(\mathbf{x} | S_1) d\mathbf{x} - (l_{12} - l_{22}) \int_{\mathbf{x} \in R_1} f(\mathbf{x} | S_2) d\mathbf{x} \\ B &= l_{22} + (l_{21} - l_{22}) \int_{\mathbf{x} \in R_2} f(\mathbf{x} | S_2) d\mathbf{x} \end{aligned}$$

La stratégie minimax consiste à annuler  $A$  (le maximum étant atteint lorsque  $\frac{\partial R(D)}{\partial p} = 0$ ). On obtient alors le risque suivant :

$$R_M = l_{22} + (l_{21} - l_{22}) \int_{\mathbf{x} \in R_2} f(\mathbf{x} | S_2) d\mathbf{x}$$

Cette stratégie est pénalisante et couteuse, mais en contrepartie elle garantit le risque, et l'on sait malheureusement que, suivant la loi de Murphy, « si le pire peut arriver, il arrivera ». Une stratégie alternative consiste à considérer  $p$  comme une variable aléatoire et à le munir d'une loi à priori : c'est le point de vue dit bayésien.

## 5.3 Le critère de Neyman-Pearson

Si l'on ne connaît pas non plus les couts, le problème se reformule comme un test d'hypothèse.

## 6 Probabilité d'erreur d'une règle de décision

Dans le cas de deux classes et de deux actions, la fonction de décision réalise une partition de l'espace des formes. Appelons  $\mathcal{R}_1(D)$  la région dans laquelle la règle de décision  $D$  choisit la classe 1.  $\mathcal{R}_2(D)$  est alors la région dans laquelle la règle de décision  $D$  choisit la classe 2. Il y a deux manières de se tromper : soit on décide la classe 1 alors que  $\mathbf{x}$  a été émise par la source 2, soit inversement on décide la classe 2 alors que  $\mathbf{x}$  a été émise par la source 1. En considérant que  $S$  est une variable aléatoire de Bernoulli codant la source sur 0/1 et que les actions sont elles aussi codées sur 0/1, la probabilité d'erreur de la fonction de décision  $D$  s'écrit alors :

$$\begin{aligned} \mathbb{P}(e|x) &= \mathbb{P}(D(\mathbf{x}) \neq S) \\ &= \mathbb{P}(\mathbf{x} \in \mathcal{R}_1 \text{ et } S_2) + \mathbb{P}(\mathbf{x} \in \mathcal{R}_2 \text{ et } S_1) \\ &= \int_{\mathcal{R}_1} 1 - \mathbb{P}(S_1 | \mathbf{x}) f(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{R}_2} \mathbb{P}(S_1 | \mathbf{x}) f(\mathbf{x}) d\mathbf{x} \\ &= (1 - p) \int_{\mathcal{R}_1} f(\mathbf{x}) d\mathbf{x} + p \int_{\mathcal{R}_2} f(\mathbf{x}) d\mathbf{x} \\ &= R(D) \end{aligned}$$

Dans ce cas, minimiser le risque c'est aussi minimiser la probabilité de la règle de décision. Cela, nous le verrons, nous ouvre d'autres possibilités quant au choix de la méthode pour estimer  $D$ .

## 7 Le cas multiclasse

Dans le cas de plus de deux classes, il est plus pratique de traiter chaque classe séparément en associant à chacune une *fonction de discrimination* ou fonction discriminante, définie comme le risque conditionnel de chaque décision.

## 8 Le cas Gaussien

La densité d'une loi gaussienne s'écrit de la manière générale sous la forme :

$$f_1(\mathbf{x}) = \frac{1}{2\pi\Sigma} \exp^{-\frac{1}{2}(\mathbf{x}-\mu_1)^\top \Sigma_1^{-1}(\mathbf{x}-\mu_1)}$$

de même la loi de  $X$  sachant qu'elle provient de la source  $S_2$  est distribuée suivant la loi normale  $\mathcal{N}(\mu_2, \Sigma_2)$ .

### 8.1 Deux classes de même variance

Commençons par le cas le plus simple. Lorsque  $\mathbf{x}$  est monodimensionnel on a :

$$\begin{aligned} \mathcal{L}(x) &= \frac{-\|x - \mu_1\|^2}{2\sigma^2} - \frac{-\|x - \mu_2\|^2}{2\sigma^2} \\ &= \frac{-x^2 + 2x\mu_1 - \mu_1^2 + x^2 - 2x\mu_2 + \mu_2^2}{2\sigma^2} \\ &= \frac{2x(\mu_1 - \mu_2) - \mu_1^2 + \mu_2^2}{2\sigma^2} \end{aligned}$$

et donc

$$\begin{aligned} \mathcal{L}(x) < k &\Leftrightarrow 2x(\mu_1 - \mu_2) - \mu_1^2 + \mu_2^2 < 2k\sigma^2 \\ &\Leftrightarrow x(\mu_1 - \mu_2) < k' \end{aligned}$$

avec  $k' = k\sigma^2 - \frac{1}{2}(\mu_1^2 - \mu_2^2)$

Dans le cas multidimensionnel avec  $\Sigma = \sigma^2 I$  où  $I$  est la matrice identité, les calculs sont exactement les mêmes. La frontière de décision est dans ce cas l'hyperplan d'équation :

$$(\mu_1 - \mu_2)^\top \mathbf{x} > k'$$

Si l'on considère maintenant le cas quelconque la frontière de décision est toujours un hyperplan. Il s'écrit :

$$\underbrace{((\mu_1 - \mu_2)\Sigma^{-1})^\top}_{\mathbf{w}^\top} \mathbf{x} > k'$$

### 8.2 Deux classes de variances différentes

Dans ce cas :

$$\begin{aligned} \mathcal{L}(x) &= -\frac{1}{2}(\mathbf{x} - \mu_1)^\top \Sigma_1^{-1}(\mathbf{x} - \mu_1) + \frac{1}{2}(\mathbf{x} - \mu_2)^\top \Sigma_2^{-1}(\mathbf{x} - \mu_2) \\ &= \frac{1}{2}\mathbf{x}^\top (\Sigma_2^{-1} - \Sigma_1^{-1})\mathbf{x} + (\Sigma_2^{-1}\mu_2 + \Sigma_1^{-1}\mu_1)^\top \mathbf{x} + \frac{1}{2}(\mu_1^\top \Sigma_1^{-1}\mu_1 - \mu_2^\top \Sigma_2^{-1}\mu_2) \\ &= \frac{1}{2}\mathbf{x}^\top H\mathbf{x} + \mathbf{w}^\top \mathbf{x} + k \end{aligned}$$

avec  $H$  une matrice,  $\mathbf{w}$  un vecteur et  $k$  une constante. La frontière de décision est maintenant une fonction quadratique (cercle, ellipse, hyperbole ou parabole).

### 8.3 $K$ classes de même variance

Il est alors plus utile de considérer une fonction de discrimination pour chaque classe plutôt que les rapports de vraisemblance. Pour chacune des  $K$  classes (sources) et pour des couts 0/1, on a la fonction de discrimination suivante pour  $c = 1, K$  :

$$g_c(\mathbf{x}) = \underbrace{(\mu_c \Sigma^{-1})^\top}_{\mathbf{w}^\top} \mathbf{x} - \frac{1}{2} \mu_c^\top \Sigma^{-1} \mu_c + \log \mathbb{P}(S_c)$$

Les frontières de décisions sont alors linéaires par morceau.

## 9 Conclusion

Si l'on connaît  $f$ ,  $\mathbb{P}$  et  $l$ , tout va bien. On peut résumer la théorie de la décision de bayes dans le tableau suivant :

cadre général	cadre bayésien	2 classes et cout 0/1
fonction de discrimination $g_c(\mathbf{x})$	risque conditionnel $R(\alpha_c, \mathbf{x})$	probabilité a posteriori $\mathbb{P}(S_1   \mathbf{x})$
rapport de vraisemblance		$\log \frac{\mathbb{P}(S_1   \mathbf{x})}{\mathbb{P}(S_2   \mathbf{x})}$
règle de décision	$D(\mathbf{x}) = \min_c R(\alpha_c, \mathbf{x})$	$D(\mathbf{x}) = \max(\mathbb{P}(S_1   \mathbf{x}), \mathbb{P}(S_2   \mathbf{x}))$
$D(\mathbf{x}) = c^*$ avec $g_{c^*}(\mathbf{x}) \geq g_c(\mathbf{x})$		$D(\mathbf{x}) = \text{signe} \left( \log \frac{\mathbb{P}(S_1   \mathbf{x})}{\mathbb{P}(S_2   \mathbf{x})} \right)$

On peut alors élaborer les trois stratégies d'estimation suivantes :

- estimer  $f$ ,  $\mathbb{P}$  à partir des données dont on dispose puis en déduire la règle de décision,
- dans le cas de couts 0/1 et de deux classes, estimer la loi a posteriori, et appliquer la règle de Bayes (on décide par rapport à la valeur 1/2),
- estimer directement la règle de décision, souvent comme étant le signe d'une fonction de discrimination  $g$ .

Ces trois stratégies ne sont pas si antagonistes qu'il y paraît. Il existe des passerelles permettant de passer de l'une à l'autre de ces stratégies comme le critère softmax. Audela du point de vue statistique, il existe d'autres manières de formuler le problème, comme par exemple la théorie de l'information, la théorie de l'évidence ou la logique floue. Loin d'être antagonistes, ces cadres sont même parfois complémentaires.

## Exercices

1. Une chaîne de production fonctionne correctement ou non. Pour surveiller cette chaîne on procède à un examen préliminaire à l'issue duquel on peut décider soit de ne rien faire, soit d'arrêter la chaîne soit de procéder à des examens complémentaires approfondis. Reformuler ce problème comme un problème de décision bayésienne à deux classes. Dans ce cas à quoi pourrait correspondre le rejet de distance ?
2. on souhaite réaliser un système permettant de distinguer les caractères manuscrits 1 et 7. formuler ce problème comme un problème de décision bayésien.
3. écrire la règle du MAP dans le cas de fonctions de vraisemblance exponentielles de paramètres  $\lambda_1$  et  $\lambda_2$ ,

4. quel est le risque minmax dans ce cas,
5. écrire la règle du MAP dans le cas de fonctions de vraisemblance Laplace,
6. montrez que le point de vue bayésien avec une loi a priori uniforme revient à la stratégie minmax.