Eléments de Théorie Bayésienne de la Décision

Machine Learning

Cédric RICHARD
Université Côte d'Azur

Théorie bayésienne de la décision

Préalable

Objectif : La théorie bayésienne de la décision est une approche statistique fondamentale pour la résolution de problème de détection et classification.

Coûts : Cette approche repose sur la recherche d'une décision optimum définie à partir des lois de probabilité régissant les hypothèses testées et de coûts associés. Elle vise à proposer la décision de coût minimum.

Exemple : détection de cible par un radar, diagnostic médical

Hypothèses : Le problème est posé en termes probabilistes. Toutes les lois nécessaires sont supposées connues.

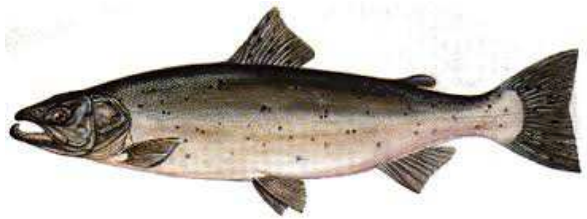
Théorie bayésienne de la décision

Exemple

▷ On considère le problème introductif de classification de poissons selon les classes "saumon" et "bar".

▷ On définit une variable aléatoire ω telles que :

$$\begin{cases} \omega = \omega_1 & \text{pour un saumon} \\ \omega = \omega_2 & \text{pour un bar} \end{cases}$$



Saumon

Bar

Théorie bayésienne de la décision

Probabilités *a priori*

▷ Les probabilités *a priori* définissent les probabilités de chaque classe avant le recueil des données.

On note $P(\omega = \omega_i)$ ou plus simplement $P(\omega_i)$

Pour C classes en compétition, on a : $\sum_{i=1}^{C} P(\omega_i) = 1$

▷ Pour l'exemple précédent, elles définissent les probabilités de voir, soit un saumon, soit un bar, sur le convoyeur du bateau usine

Exemple : $P(\omega_1) = \frac{3}{5} \qquad P(\omega_2) = \frac{2}{5}$

▷ Les probabilités *a priori* peuvent varier selon la situation.

— Si l'on observe autant de saumons que de bars sur le convoyeur, on pourra fixer $P(\omega_1) = P(\omega_2) = \frac{1}{2}$

— Ces probabilités peuvent évoluer selon la saison.

Théorie bayésienne de la décision

Prise de décision à partir des probabilités *a priori*

▷ Une règle de décision prescrit l'action à entreprendre étant donné une observation.

Exemple : attribuer le poisson observé $\boldsymbol{x}$ à la classe ω_2 des bars.

▷ Si les seules informations disponibles (hypothèses) sont :
- les probabilités *a priori* de chaque classe
- les conséquences des mauvaises décisions "choisir ω_1 pour ω_2" et "choisir ω_2 pour ω_1" sont équivalentes

Quelle règle de décision adopter ?

▷ Choisir ω_1 si $P(\omega_1) > P(\omega_2)$, sinon ω_2
- Règle raisonnable mais consistant à choisir **toujours la même classe de poisson**, quelle que soit l'observation $\boldsymbol{x}$.
- Sous les 2 hypothèses ci-dessus, aucune règle ne fait mieux. Voir ci-après.
- Mais sinon, que faire pour d'autres hypothèses ? Voir ci-après.

Théorie bayésienne de la décision

Paramètres et espace des paramètres

▷ Un **paramètre** (feature) est une variable/caractéristique observable de $\boldsymbol{x}$.

▷ **L'espace des paramètres** (feature space) est l'ensemble dans lequel les paramètres sont échantillonnés.
Exemple : pour les attributs des poissons
— longueur
— largeur
— couleur
— position de la nageoire dorsale

▷ Pour simplifier, on supposera que les paramètres sont des variables continues.

▷ On note : $\boldsymbol{x} \in \mathbb{R}^d$, où d est le nombre de caractéristiques.

Théorie bayésienne de la décision

Densités conditionnelles

▷ La **densité de probabilité conditionnelle** à la classe ω_i est la densité de probabilité de $\boldsymbol{x}$ étant donné l'appartenance de l'individu à la classe ω_i :

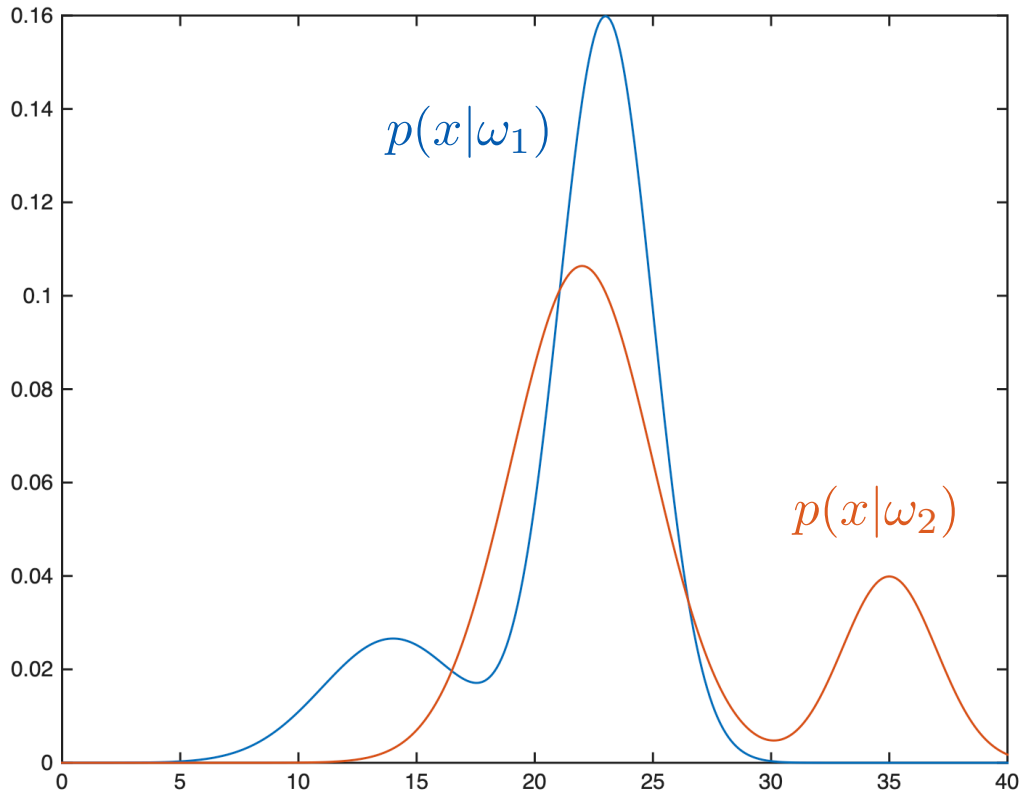
$$p(\boldsymbol{x}|\omega_i)$$

Cette fonction est également appelée *vraisemblance*.

Exemple : attributs des poissons selon le type

Théorie bayésienne de la décision

Densités conditionnelles



Théorie bayésienne de la décision

Formule de Bayes et probabilité *a posteriori*

▷ La **probabilité *a posteriori*** désigne la probabilité d'une classe ω_i étant donné une observation $\boldsymbol{x}$:

$$P(\omega_i|\boldsymbol{x})$$

▷ Par la formule de Bayes

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

on aboutit à :

$$\begin{aligned} P(\omega_i|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\omega_i)\, P(\omega_i)}{p(\boldsymbol{x})} \\ &= \frac{p(\boldsymbol{x}|\omega_i)\, P(\omega_i)}{\sum_j p(\boldsymbol{x}|\omega_j)\, P(\omega_j)} \end{aligned}$$

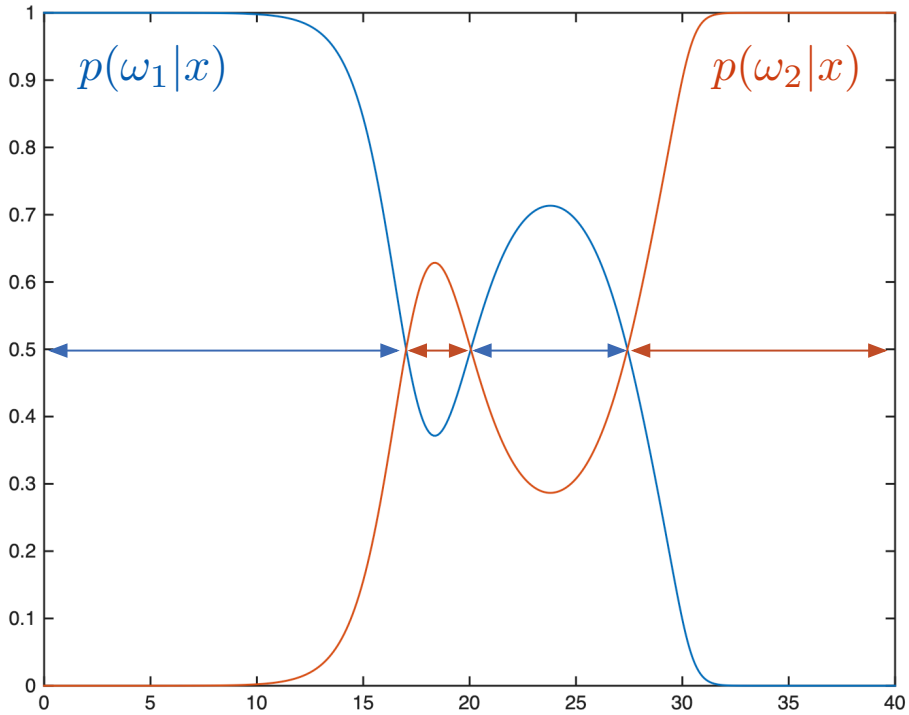
▷ Puisque $p(\boldsymbol{x})$ apparait comme un terme de normalisation, indépendant de la classe, on note que :

$$P(\omega_i|\boldsymbol{x}) \propto p(\boldsymbol{x}|\omega_i)\, P(\omega_i)$$

Théorie bayésienne de la décision

Probabilité *a posteriori*

Exemple du transparent 7 avec $P(\omega_1) = \frac{3}{5} \quad P(\omega_2) = \frac{2}{5}$

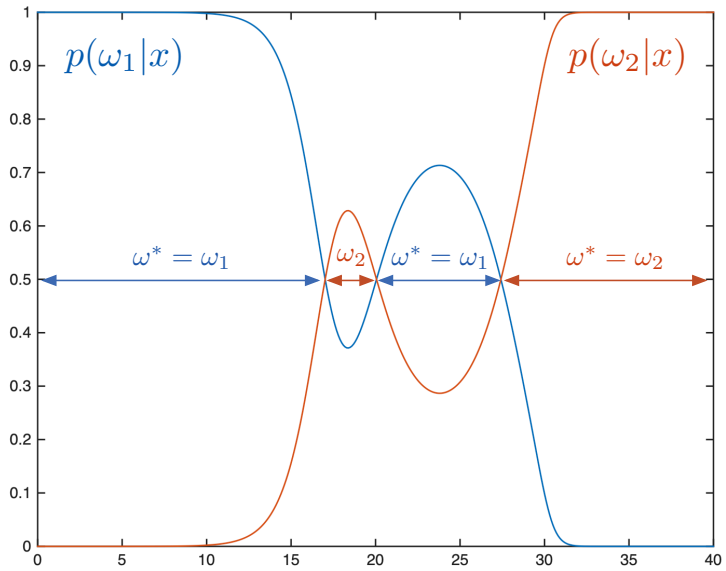


Théorie bayésienne de la décision

Probabilité *a posteriori*

Etant donné une observation $\boldsymbol{x}$, il est pertinent de prendre la décision à partir de la probabilité *a posteriori* :

$$\boxed{\omega^* = \arg\max_{\omega_i} P(\omega_i|\boldsymbol{x})}$$



Théorie bayésienne de la décision

Règle du maximum de probabilité *a posteriori*

On considère un problème de classification à 2 classes (ω_1 et ω_2).

La règle de décision du maximum de probabilité *a posteriori* s'écrit :

Si $P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$ choisir ω_1, sinon choisir ω_2

que l'on peut écrire plus simplement :

$$\boxed{P(\omega_1|\boldsymbol{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P(\omega_2|\boldsymbol{x})}$$

Théorie bayésienne de la décision

Règle du maximum de probabilité *a posteriori*

En utilisant la formule de Bayes, on aboutit à :

$$\frac{p(\boldsymbol{x}|\omega_1)P(\omega_1)}{p(\boldsymbol{x})} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{p(\boldsymbol{x}|\omega_2)P(\omega_2)}{p(\boldsymbol{x})}$$

c'est-à-dire :

$$\boxed{\Lambda(\boldsymbol{x}) \triangleq \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}}$$

Le terme $\Lambda(\boldsymbol{x}) \triangleq \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)}$ est appelé *rapport de vraisemblance*, et $\lambda_0 = \frac{P(\omega_2)}{P(\omega_1)}$ le *seuil*.

Ce test nécessite de connaitre les probabilités *a priori* $P(\omega_1)$ et $P(\omega_2)$, ainsi que les densités de probabilité conditionnelles $p(\boldsymbol{x}|\omega_1)$ et $p(\boldsymbol{x}|\omega_2)$.

Théorie bayésienne de la décision

Exemple

On considère un problème de catégorisation à 2 classes, dont les densités de probabilité conditionnelles sont définies par une loi normale scalaire :

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

avec $\mu_1 = 4$, $\mu_2 = 10$ et $\sigma_1^2 = \sigma_2^2 = 1$.

On suppose que : $P(\omega_1) = P(\omega_2) = \frac{1}{2}$.

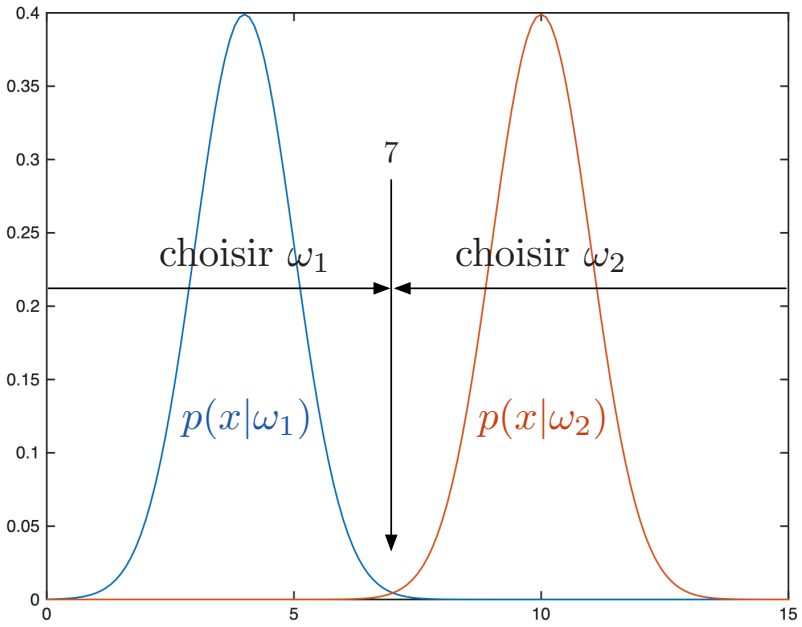
Montrer que la règle du maximum de probabilité *a posteriori* s'écrit :

$$x \underset{\omega_1}{\overset{\omega_2}{\gtrless}} 7$$

Théorie bayésienne de la décision

Exemple

$$x \underset{\omega_1}{\overset{\omega_2}{\gtrless}} 7$$



Théorie bayésienne de la décision

Exemple

Déterminer la règle de décision lorsque $P(\omega_1) = 2P(\omega_2)$

Théorie bayésienne de la décision

Probabilité d'erreur

On considère la règle de décision (problème à 2 classes) :

$$\boxed{P(\omega_1|\boldsymbol{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P(\omega_2|\boldsymbol{x})}$$

Étant donné une observation $\boldsymbol{x}$, on commet une erreur si :

- ▷ on décide ω_1 alors que $\omega = \omega_2$
- ▷ on décide ω_2 alors que $\omega = \omega_1$

En conséquence :

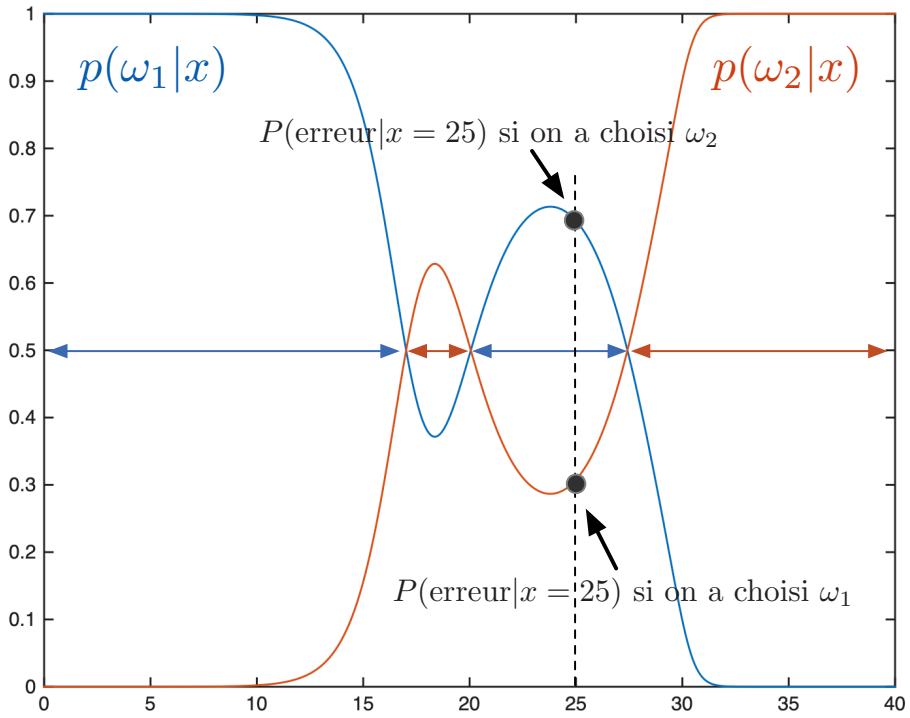
$$P(\text{erreur}|\boldsymbol{x}) = \begin{cases} P(\omega_1|\boldsymbol{x}) \text{ si on a décidé } \omega_2 \\ P(\omega_2|\boldsymbol{x}) \text{ si on a décidé } \omega_1 \end{cases}$$

$$P(\text{erreur}) = \int P(\text{erreur}|\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

Théorie bayésienne de la décision

Probabilité d'erreur

Exemple du transparent 7 :



Théorie bayésienne de la décision

Probabilité d'erreur

Afin de minimiser

$$P(\text{erreur}) = \int P(\text{erreur}|\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

avec

$$P(\text{erreur}|\boldsymbol{x}) = \begin{cases} P(\omega_1|\boldsymbol{x}) \text{ si on a décidé } \omega_2 \\ P(\omega_2|\boldsymbol{x}) \text{ si on a décidé } \omega_1 \end{cases}$$

Étant donné $\boldsymbol{x}$, il faut donc :

- ▷ choisir ω_1 si $P(\omega_2|\boldsymbol{x}) < P(\omega_1|\boldsymbol{x})$ car alors $P(\text{erreur}|\boldsymbol{x}) = P(\omega_2|\boldsymbol{x})$
- ▷ choisir ω_2 si $P(\omega_1|\boldsymbol{x}) < P(\omega_2|\boldsymbol{x})$ car alors $P(\text{erreur}|\boldsymbol{x}) = P(\omega_1|\boldsymbol{x})$

On retrouve la règle du maximum de probabilité *a posteriori*

La règle du maximum de probabilité *a posteriori* minimise $P(\text{erreur})$

Théorie bayésienne de la décision

Calcul de la probabilité d'erreur

Par le théorème des probabilités totales :

$$P(\text{erreur}) = P(\text{erreur}|\omega_1)P(\omega_1) + P(\text{erreur}|\omega_2)P(\omega_2)$$

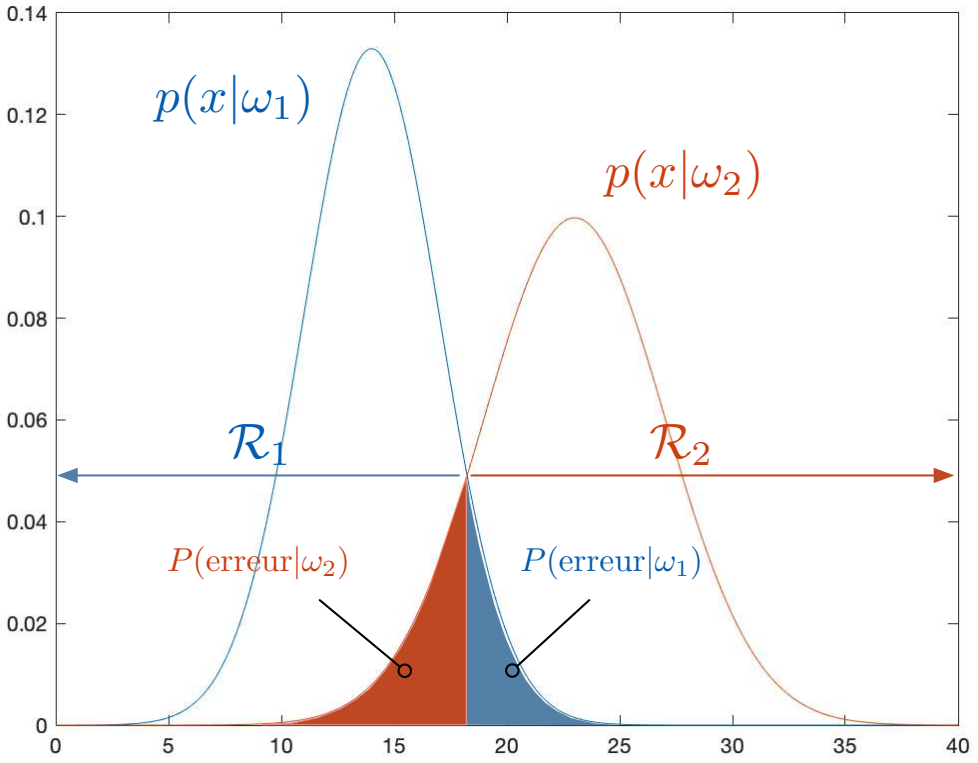
avec

$$P(\text{erreur}|\omega_i) = P(\text{choisir } \omega_j|\omega_i) = \int_{\boldsymbol{x}\in\mathcal{R}_j} p(\boldsymbol{x}|\omega_i)\, d\boldsymbol{x}, \quad i \neq j$$

où $\mathcal{R}_j$ désigne la région de décision de $\boldsymbol{x}$ correspondant à la décision ω_j

Théorie bayésienne de la décision

Calcul de la probabilité d'erreur



Théorie bayésienne de la décision

Calcul de la probabilité d'erreur

Exemple du transparent 13 :

$$\begin{aligned} P(\text{erreur}) &= \frac{1}{2}\int_{x\geq 7} p(x|\omega_1)\,dx + \frac{1}{2}\int_{x<7} p(x|\omega_2)\,dx \\ &= \frac{1}{2\sqrt{2\pi}}\int_{x\geq 7} \exp\left(-\frac{(x-4)^2}{2}\right)dx + \frac{1}{2\sqrt{2\pi}}\int_{x<7} \exp\left(-\frac{(x-10)^2}{2}\right)dx \end{aligned}$$

Soit $\Phi(\lambda)$ la fonction de répartition de la loi normale centrée réduite :

$$\Phi(\lambda) = \int_{-\infty}^{\lambda} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)du$$

Par les changements de variables $z = x - 4$ et $z' = x - 10$, on en déduit :

$$\begin{aligned} P(\text{erreur}) &= \frac{1}{2}\left[1 - \Phi(3)\right] + \frac{1}{2}\Phi(-3) \\ &= \frac{1}{2}\left[1 - \Phi(3)\right] + \frac{1}{2}\left[1 - \Phi(3)\right] \approx 1.35 \cdot 10^{-3} \end{aligned}$$

Théorie bayésienne de la décision

Règle du maximum de probabilité *a posteriori*

$$\boxed{\Lambda(\boldsymbol{x}) \triangleq \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}}$$

Si $P(\omega_1) = P(\omega_2)$ la règle devient :

$$p(\boldsymbol{x}|\omega_1) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} p(\boldsymbol{x}|\omega_2)$$

Si $p(\boldsymbol{x}|\omega_1) = p(\boldsymbol{x}|\omega_2)$ la règle devient (en réponse au transparent 4, point 3) :

$$P(\omega_1) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P(\omega_2)$$

Théorie bayésienne de la décision

Généralisation à C classes

La règle du maximum *a posteriori* se généralise à C classes $\omega_1, \ldots, \omega_C$

$$\boxed{\omega^* = \arg\max_{\omega_i} P(\omega_i|\boldsymbol{x})}$$

Elle minimise la probabilité d'erreur, définie par :

$$P(\text{erreur}) = \int P(\text{erreur}|\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

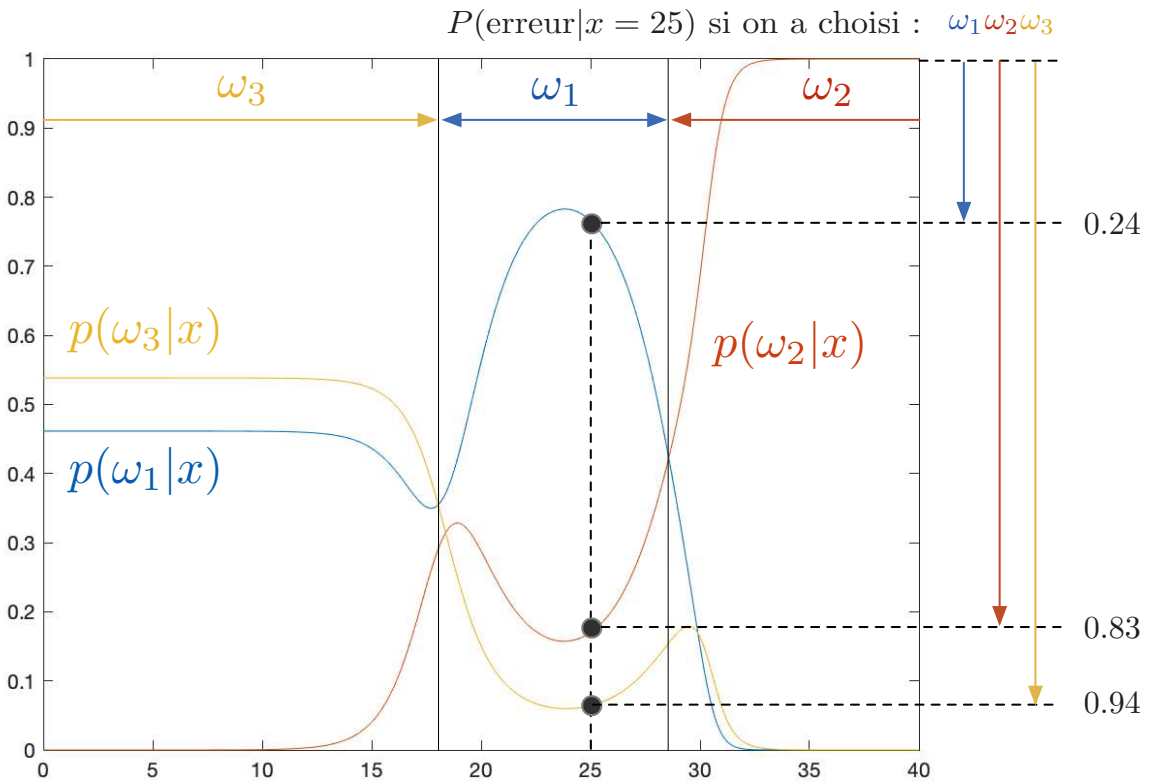
avec

$$P(\text{erreur}|\boldsymbol{x}) = \begin{cases} 1 - P(\omega_1|\boldsymbol{x}) \text{ si on a décidé } \omega_1 \\ \ldots \\ 1 - P(\omega_C|\boldsymbol{x}) \text{ si on a décidé } \omega_C \end{cases}$$

En effet, étant donné $\boldsymbol{x}$, décider ω^* tel que $\omega^* = \arg\min_{\omega_i} (1 - P(\omega_i|\boldsymbol{x}))$ permet de minimiser $P(\text{erreur}|\boldsymbol{x})$.

Théorie bayésienne de la décision

Généralisation à C classes



Théorie bayésienne de la décision

Calcul de la probabilité d'erreur

Par le théorème des probabilités totales :

$$P(\text{erreur}) = \sum_{i=1}^{C} P(\text{erreur}|\omega_i)P(\omega_i)$$

avec

$$P(\text{erreur}|\omega_i) = \sum_{j \neq i} P(\text{choisir } \omega_j|\omega_i) = 1 - \int_{\boldsymbol{x} \in \mathcal{R}_i} p(\boldsymbol{x}|\omega_i)\, d\boldsymbol{x}$$

où $\mathcal{R}_i$ désigne la région de décision de $\boldsymbol{x}$ correspondant à la décision ω_i

Théorie bayésienne de la décision

Critère de Bayes

▷ La règle précédente accorde la même importance à toutes les erreurs

"choisir ω_i alors que ω_j est la décision exacte"

▷ Pour certaines applications, les erreurs n'ont pas la même gravité :

"ne pas détecter un cancer alors que le patient en a un"

versus

"détecter un cancer alors que le patient n'en a pas"

▷ Le critère de Bayes permet de fixer le coût de chaque mauvaise décision

▷ On considère C classes $\{\omega_1, \ldots, \omega_C\}$

▷ On considère D décisions possibles $\{\delta_1, \ldots, \delta_D\}$

▷ Soit $\lambda(\delta_i, \omega_j)$ le coût de la décision δ_i tandis que la classe est ω_j

Théorie bayésienne de la décision

Critère de Bayes

▷ Exemple :

ω_1 : le patient a une tumeur maligne

ω_2 : le patient a une tumeur bénigne

δ_1 : le patient subit des examens complémentaires car diagnostiqué malade

δ_2 : le patient est renvoyé chez lui car diagnostiqué sain

ici on a $\lambda(\delta_2, \omega_1) \gg \lambda(\delta_1, \omega_2)$ sans ambiguïté

Théorie bayésienne de la décision

Critère de Bayes

▷ On définit un coût $R(\delta_i|\boldsymbol{x})$ pour chacune des décisions possibles δ_i en fonction des classes d'appartenance possibles ω_j d'une observation $\boldsymbol{x}$

$$R(\delta_i|\boldsymbol{x}) = \sum_{j=1}^{C} \lambda(\delta_i, \omega_j) P(\omega_j|\boldsymbol{x})$$

Notation : on pourra écrire $\lambda_{ij} = \lambda(\delta_i, \omega_j)$ par souci de concision

Théorie bayésienne de la décision

Critère de Bayes

▷ Soit $\delta(\boldsymbol{x})$ une fonction de décision qui, à toute observation $\boldsymbol{x}$ associe une décision parmi $\{\delta_1, \ldots, \delta_D\}$, c'est-à-dire :

$$\begin{aligned} \mathbb{R}^d &\to \{\delta_1, \ldots, \delta_D\} \\ \boldsymbol{x} &\mapsto \delta(\boldsymbol{x}) \end{aligned}$$

▷ On définit le coût moyen d'une règle de décision $\delta(\boldsymbol{x})$ pour tout $\boldsymbol{x}$ par :

$$\boxed{R = \int R\big(\delta(\boldsymbol{x})|\boldsymbol{x}\big)\, p(\boldsymbol{x})\, d\boldsymbol{x}}$$

La règle de décision recherchée $\delta^*(\cdot)$ est celle minimisant R

Théorie bayésienne de la décision

Critère de Bayes

▷ La règle de Bayes minimise le coût moyen R

▷ En effet, étant donné une observation $\boldsymbol{x}$, elle consiste à choisir la décision δ^* minimisant $R(\delta_i|\boldsymbol{x})$ parmi toutes les décisions possibles δ_i

$$\begin{aligned}\delta^* &= \arg\min_{\delta_i} R(\delta_i|\boldsymbol{x}) \\ &= \arg\min_{\delta_i} \sum_{j=1}^{C} \lambda(\delta_i, \omega_j) P(\omega_j|\boldsymbol{x})\end{aligned}$$

Théorie bayésienne de la décision

Critère de Bayes : coûts 0-1

▷ Cas particulier des coûts 0-1 :

$$\lambda(\delta_i, \omega_j) = \begin{cases} 0 \text{ si } i = j \\ 1 \text{ si } i \neq j \end{cases} \qquad i, j = 1, 2, \ldots, C$$

Coût nul pour une bonne décision, mêmes coûts pour les mauvaises décisions

▷ Pour les coût 0-1, on a :

$$\begin{aligned} R(\delta_i|\boldsymbol{x}) &= \sum_{j \neq i} P(\omega_j|\boldsymbol{x}) \\ &= 1 - P(\omega_i|\boldsymbol{x}) \end{aligned}$$

▷ Choisir la décision δ_i pour laquelle $R(\delta_i|\boldsymbol{x})$ est minimum revient à choisir celle pour laquelle $P(\omega_i|\boldsymbol{x})$ est maximum

Pour les coûts 0-1, on retrouve à la règle du maximum de probabilité *a posteriori*

Théorie bayésienne de la décision

Critère de Bayes : cas de 2 classes

▷ On considère deux classes $\{\omega_1, \omega_2\}$ et deux décisions possibles $\{\delta_1, \delta_2\}$, où la bonne décision est δ_i lorsque la vraie classe de $\boldsymbol{x}$ est ω_i

▷ On a :

$$R(\delta_1|\boldsymbol{x}) = \lambda_{11} P(\omega_1|\boldsymbol{x}) + \lambda_{12} P(\omega_2|\boldsymbol{x})$$
$$R(\delta_2|\boldsymbol{x}) = \lambda_{21} P(\omega_1|\boldsymbol{x}) + \lambda_{22} P(\omega_2|\boldsymbol{x})$$

▷ La règle de Bayes s'écrit :

$$R(\delta_2|\boldsymbol{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} R(\delta_1|\boldsymbol{x})$$

Théorie bayésienne de la décision

Critère de Bayes : cas de 2 classes

▷ En appliquant la règle de Bayes, on aboutit à :

$$\lambda_{21} p(\boldsymbol{x}|\omega_1) P(\omega_1) + \lambda_{22} p(\boldsymbol{x}|\omega_2) P(\omega_2) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \lambda_{11} p(\boldsymbol{x}|\omega_1) P(\omega_1) + \lambda_{12} p(\boldsymbol{x}|\omega_2) P(\omega_2)$$

▷ En supposant que $\lambda_{21} > \lambda_{11}$ et $\lambda_{12} > \lambda_{22}$, c'est-à-dire que le coût d'une mauvaise décision est supérieur à celui d'une bonne décision, on obtient :

$$\boxed{\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}}$$

Théorie bayésienne de la décision

Critère de Bayes : cas de 2 classes

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

- ▷ La règle de Bayes revient à comparer le rapport de vraisemblance à un seuil
- ▷ Dans le cas de coût 0-1, la règle de Bayes se ramène à la règle du maximum de probabilité *a posteriori* :

$$\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}$$

Théorie bayésienne de la décision

Exemple

On considère un problème de catégorisation à 2 classes, dont les densités de probabilité conditionnelles sont définies par une loi normale scalaire :

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

avec $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1^2 = 3$ et $\sigma_2^2 = 1$.

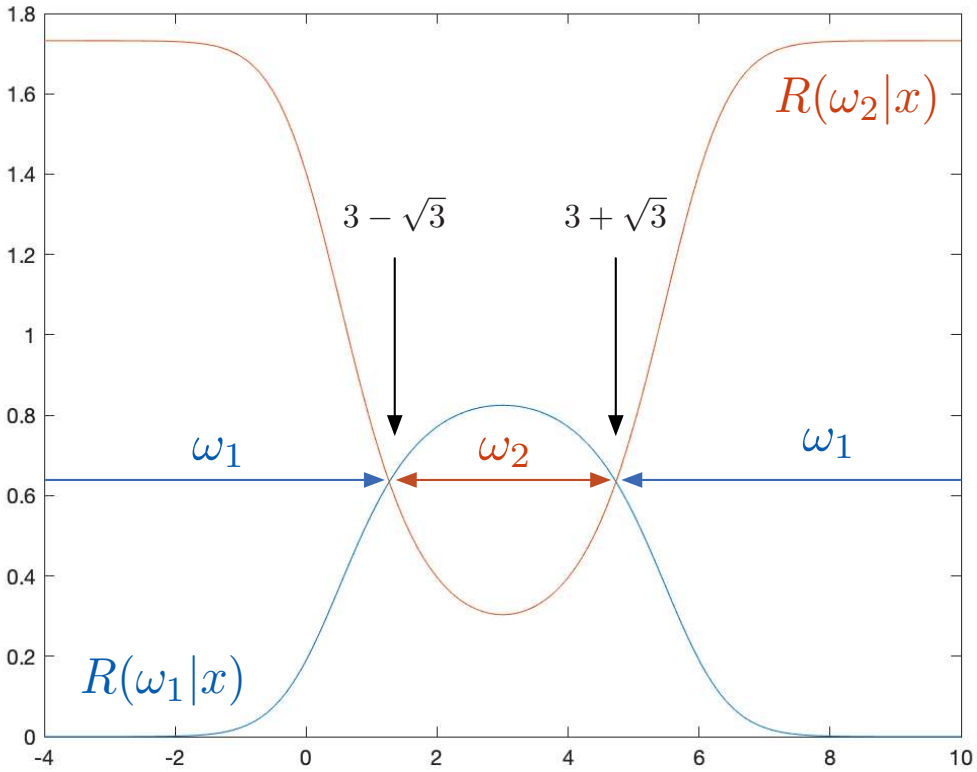
On suppose que : $P(\omega_1) = P(\omega_2) = \frac{1}{2}$, $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = 1$ et $\lambda_{21} = \sqrt{3}$

Montrer que la règle de Bayes s'écrit :

$$2x^2 - 12x + 12 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$

Théorie bayésienne de la décision

Exemple



Théorie bayésienne de la décision

Récapitulatif

Règle de Bayes :

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

Règle du maximum de probabilité *a posteriori* :

$$\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}$$

Règle du maximum de vraisemblance :

$$\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$

Fonctions discriminantes

- Les fonctions discriminantes constituent une façon pratique pour représenter des classifieurs
- Soit $g_i(\boldsymbol{x})$ une fonction discriminante pour la classe ω_i. Le classifieur assigne une observation $\boldsymbol{x}$ à la classe ω_i si

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad \forall j \neq i$$

  ou, de manière équivalente

$$i = \arg\max_j g_j(\boldsymbol{x})$$

- Exemples :
  - règles de Bayes : $g_i(\boldsymbol{x}) = -R(\delta_i|\boldsymbol{x}) = -\sum_{j=1}^{C} \lambda_{ij} P(\omega_j|\boldsymbol{x})$
  - règle du maximum *a posteriori* : $g_i(\boldsymbol{x}) = P(\omega_i|\boldsymbol{x})$

Fonctions discriminantes

Unicité

▷ Le choix de fonctions discriminantes n'est pas unique. Soit $f(\cdot)$ une fonction strictement croissante. En effet, on a :

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad \forall j \neq i \iff f\big(g_i(\boldsymbol{x})\big) > f\big(g_j(\boldsymbol{x})\big) \quad \forall j \neq i$$

▷ Le choix d'une formulation particulière peut faciliter la compréhension ou l'implémentation.

▷ Exemples : les fonctions discriminantes suivantes sont toutes de probabilité d'erreur minimum

— $g_i(\boldsymbol{x}) = P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{\sum_j p(\boldsymbol{x}|\omega_j)P(\omega_j)}$

— $g_i(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_i)P(\omega_i)$

— $g_i(\boldsymbol{x}) = \log p(\boldsymbol{x}|\omega_i) + \log P(\omega_i)$

Régions de décision

Visualisation

▷ Toute règle de décision divise l'espace des observations en région de décision

▷ On note $\mathcal{R}_i$ la région de décision correspondant à la décision ω_i, définie par :

$$\text{si } g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \ \forall j \neq i, \text{ alors } \boldsymbol{x} \in \mathcal{R}_i$$

▷ Les frontières de décision séparent les régions de décision.

▷ La frontière de décision entre $\mathcal{R}_i$ et $\mathcal{R}_j$ vérifient $g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$

Fonctions discriminantes

Cas de 2 classes

- Pour un problème à 2 classes, une seule fonction discriminante est nécessaire :

$$g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$$

- La règle de décision correspondante est définie par :

$$g(\boldsymbol{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$

- La frontière de décision est la surface définie par $g(\boldsymbol{x}) = 0$

Discriminants dans le cas de lois normales

Loi normale mono-dimensionnelle

▷ La fonction de densité de la loi normale, ou de Gauss, mono-dimensionnelle est définie par :

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

▷ La moyenne μ est l'espérance de x :

$$\mu = E[x] = \int_{\mathbb{R}} x\, p(x)\, dx$$

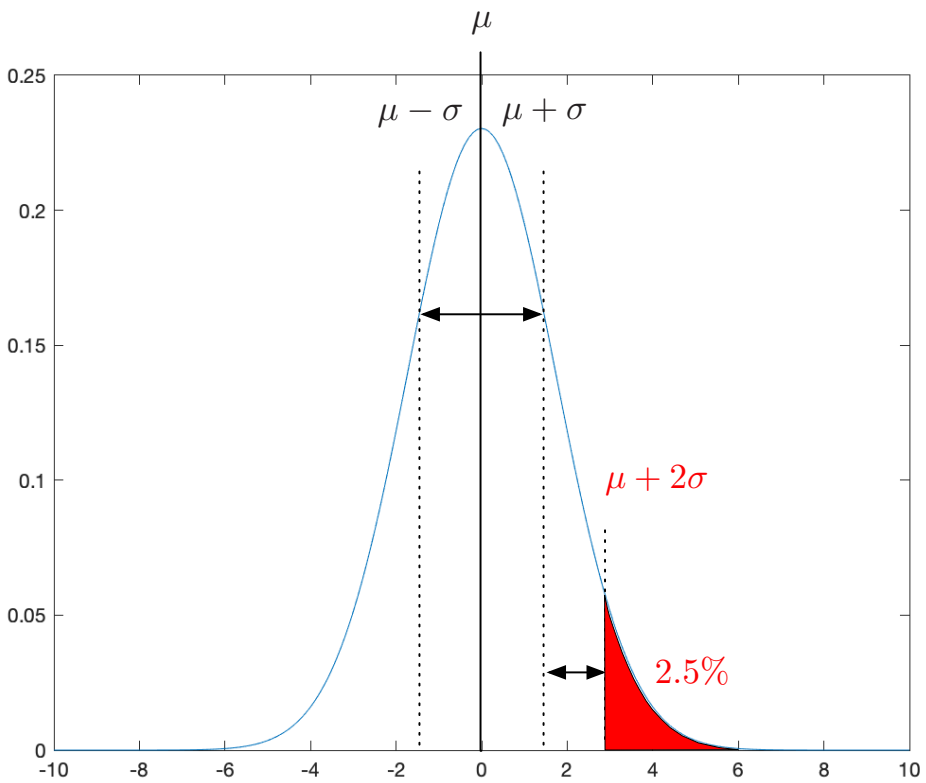
▷ La variance σ^2 est l'écart quadratique moyen de x :

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

▷ On note $x \sim \mathcal{N}(\mu, \sigma^2)$

Discriminants dans le cas de lois normales

Loi normale mono-dimensionnelle



Discriminants dans le cas de lois normales

Loi normale multi-dimensionnelle

▷ La fonction de densité de la loi normale, ou de Gauss, multidimensionnelle dans $\mathbb{R}^d$ est définie par :

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

▷ La moyenne $\boldsymbol{\mu}$ est l'espérance de $\boldsymbol{x}$:

$$\boldsymbol{\mu} = E[\boldsymbol{x}] = \int_{\mathbb{R}^d} \boldsymbol{x}\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

▷ La matrice de variance-covariance $\boldsymbol{\Sigma}$ de $\boldsymbol{x}$ est :

$$\boldsymbol{\Sigma} = E[(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})] = \int_{\mathbb{R}} (\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

▷ On note $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Discriminants dans le cas de lois normales

Matrice de variance-covariance

$$\boldsymbol{\Sigma} = E[(\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{x} - \boldsymbol{\mu})] = \int_{\mathbb{R}} (\boldsymbol{x} - \boldsymbol{\mu})^\top (\boldsymbol{x} - \boldsymbol{\mu})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

- ▷ La matrice $\boldsymbol{\Sigma}$ est symétrique, semi-définie positive :

$$\boldsymbol{u}^\top \boldsymbol{\Sigma}\, \boldsymbol{u} \geq 0 \qquad \forall \boldsymbol{u} \in \mathbb{R}^d$$

- ▷ Chaque terme diagonal σ_{ii} représente la variance de la variable x_i
- ▷ Chaque terme non-diagonal σ_{ij} est la covariance des variables x_i et x_j
- ▷ Dans le cas d'une loi normale, si $\sigma_{ij} = 0$, alors les variables x_i et x_j sont statistiquement indépendantes

Discriminants dans le cas de lois normales

Distance de Mahalanobis

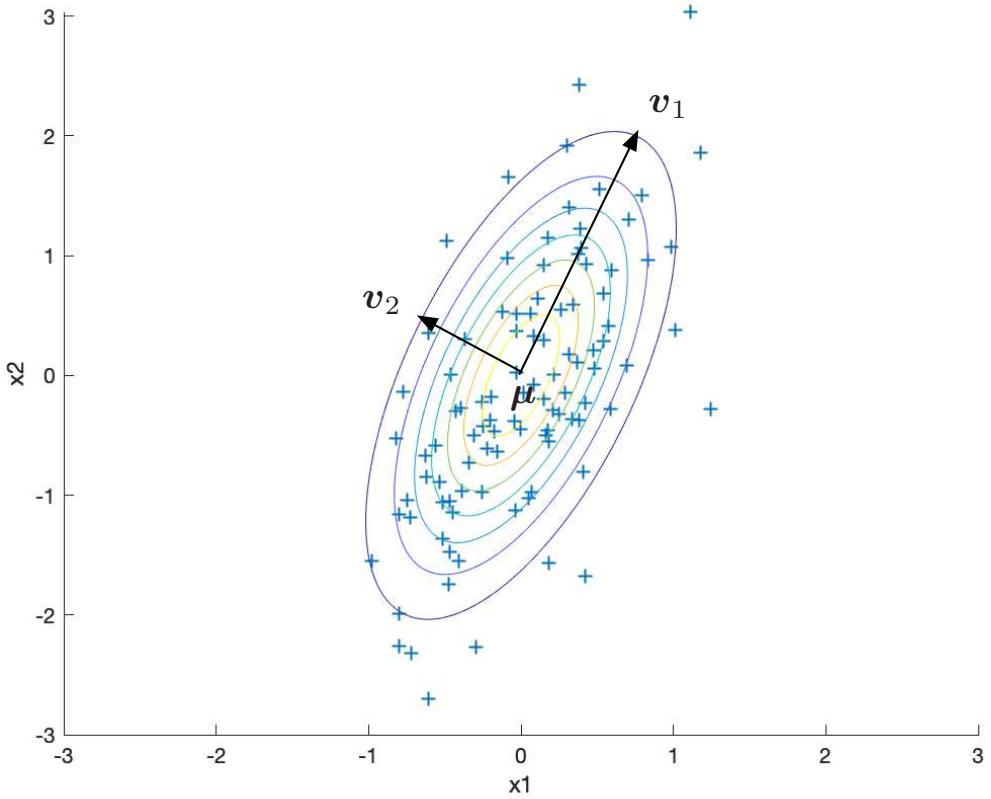
- La forme de la fonction densité de probabilité $p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ est définie par la matrice de covariance $\boldsymbol{\Sigma}$
- Les lignes d'iso-densité, i.e., $p(\boldsymbol{x}) = \text{Cte}$, sont des ellipsoïdes. Les points s'y trouvant sont à une même **distance de Mahalanobis** du point moyen $\boldsymbol{\mu}$

$$\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} \triangleq (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

- Les vecteurs propres de $\boldsymbol{\Sigma}$ sont les axes principaux des ellipsoïdes, et les valeurs propres les longueurs de ces demi-axes

Discriminants dans le cas de lois normales

Distance de Mahalanobis



Discriminants dans le cas de lois normales

Cas général

- ▷ On considère un problème de classification où les C classes ω_i en compétition suivent chacune une loi normale $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ avec une probabilité $P(\omega_i)$
- ▷ Le critère choisi pour élaborer les fonctions discriminantes est celui de la probabilité d'erreur minimum
- ▷ On a vu que les fonctions discriminantes sont de la forme $g_i(\boldsymbol{x}) = P(\omega_i|\boldsymbol{x})$, ou toutes fonctions strictement croissantes de celles-ci (voir précédemment)
- ▷ Compte tenu de la forme exponentielle de la loi normale, on choisit

$$g_i(\boldsymbol{x}) = \log p(\boldsymbol{x}|\omega_i) + \log P(\omega_i)$$

On obtient

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

Discriminants dans le cas de lois normales

Cas $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$

- ▷ Comment sont les frontières de décision lorsque $p(\boldsymbol{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 \boldsymbol{I})$, i.e., les composantes de $\boldsymbol{x}$ sont indépendantes et de même variance σ^2 ?

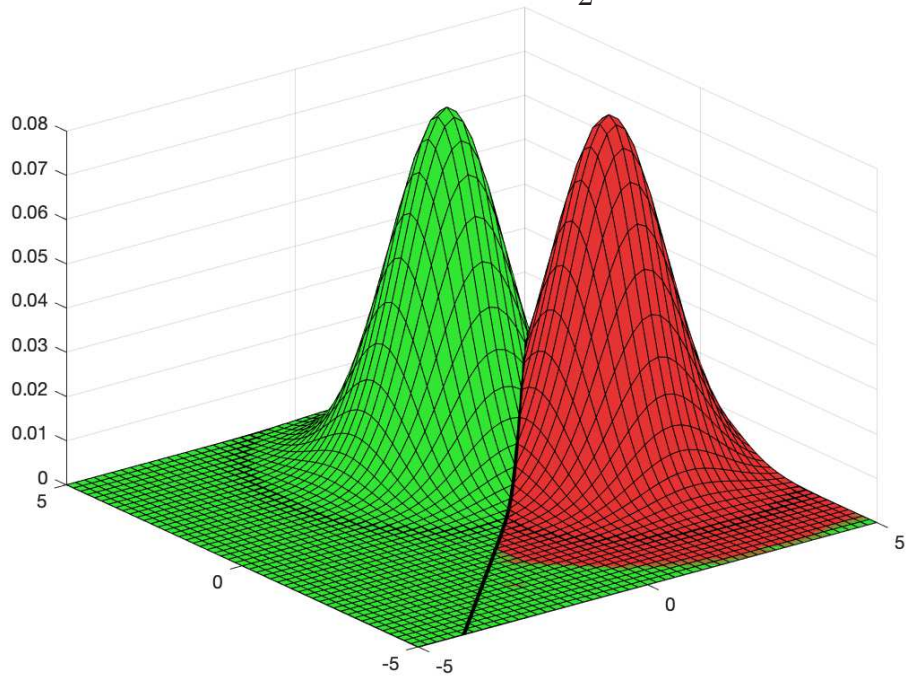
- ▷ Il s'agit d'hyperplans comme montré dans la suite

Discriminants dans le cas de lois normales

Cas $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}$$

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

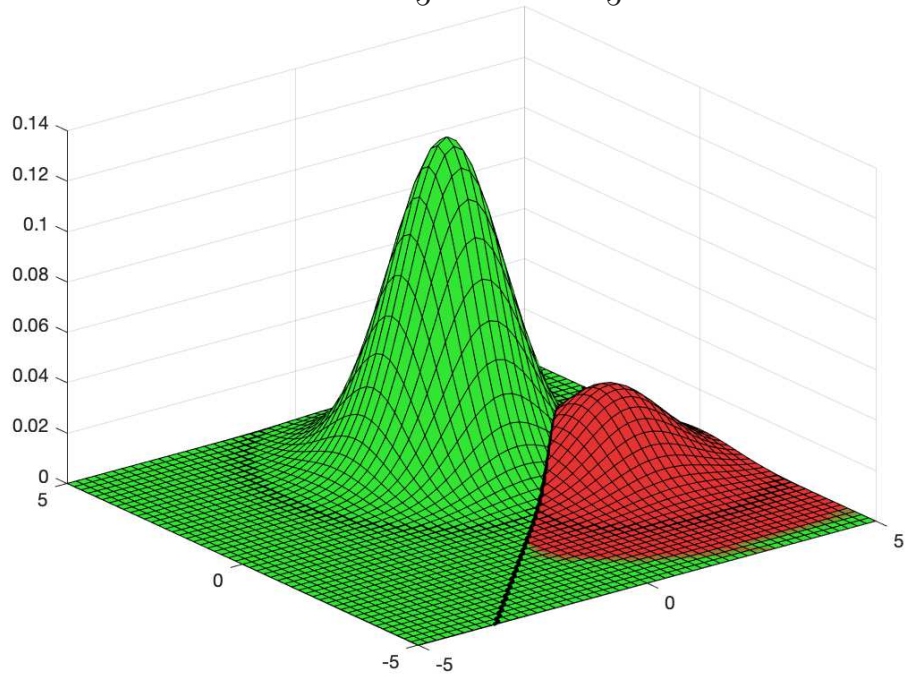


Discriminants dans le cas de lois normales

Cas $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}$$

$$P(\omega_1) = \frac{1}{5} \quad P(\omega_2) = \frac{4}{5}$$



Discriminants dans le cas de lois normales

Cas $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$

▷ Les fonctions discriminantes sont de la forme

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

où les termes ne dépendant pas de l'indice i de la classe ω_i ont été supprimés

▷ La frontière entre les régions de décision $\mathcal{R}_i$ et $\mathcal{R}_j$, définie par les points $\boldsymbol{x}$ satisfaisant $g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$, a pour équation

$$\frac{\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} - \log P(\omega_i) = \frac{\|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2}{2\sigma^2} - \log P(\omega_j)$$

Il s'agit d'un hyperplan orthogonal au vecteur $\boldsymbol{\mu}_i - \boldsymbol{\mu_j}$

Discriminants dans le cas de lois normales

Cas $\boldsymbol{\Sigma}_i = \sigma^2 \boldsymbol{I}$

▷ En développant l'équation $g_i(\boldsymbol{x}) = g_j(\boldsymbol{x})$, on trouve en effet :

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{x} = \frac{1}{2}\left(\|\boldsymbol{\mu}_i\|^2 - \|\boldsymbol{\mu}_j\|^2\right) + \sigma^2 \log\left(\frac{P(\omega_j)}{P(\omega_i)}\right)$$

ce qui conduit à

$$\boldsymbol{w}^\top(\boldsymbol{x} - \boldsymbol{x}_0) = 0$$

$$\boldsymbol{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

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

▷ Si $P(\omega_i) = P(\omega_j)$, l'hyperplan passe par le milieu du segment liant $\boldsymbol{\mu}_i$ à $\boldsymbol{\mu}_j$

Discriminants dans le cas de lois normales

Cas général

▷ Comment sont les frontières de décision lorsque $p(\boldsymbol{x}|\omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$?

▷ Les fonctions discriminantes sont de la forme

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

$$\boldsymbol{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$$

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

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log P(\omega_i)$$

▷ Les frontières entre 2 régions de décision sont donc des coniques

$$\boldsymbol{x}^\top(\boldsymbol{W}_i - \boldsymbol{W}_j)\,\boldsymbol{x} + (\boldsymbol{w}_i - \boldsymbol{w}_j)^\top \boldsymbol{x} + (w_{i0} - w_{j0}) = 0$$

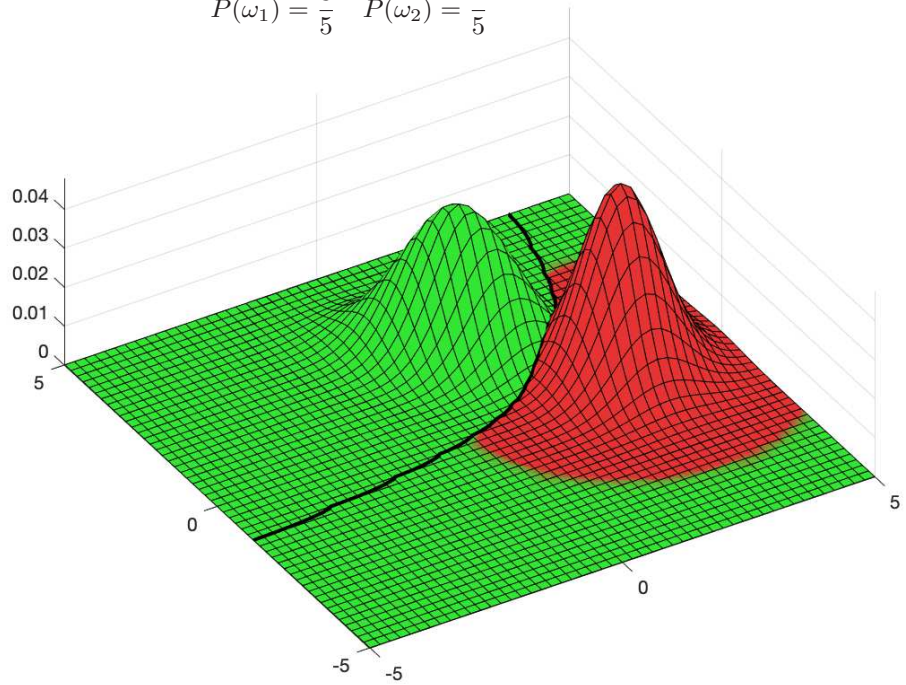
Discriminants dans le cas de lois normales

Cas général

$\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$

$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

$P(\omega_1) = \frac{3}{5} \quad P(\omega_2) = \frac{2}{5}$



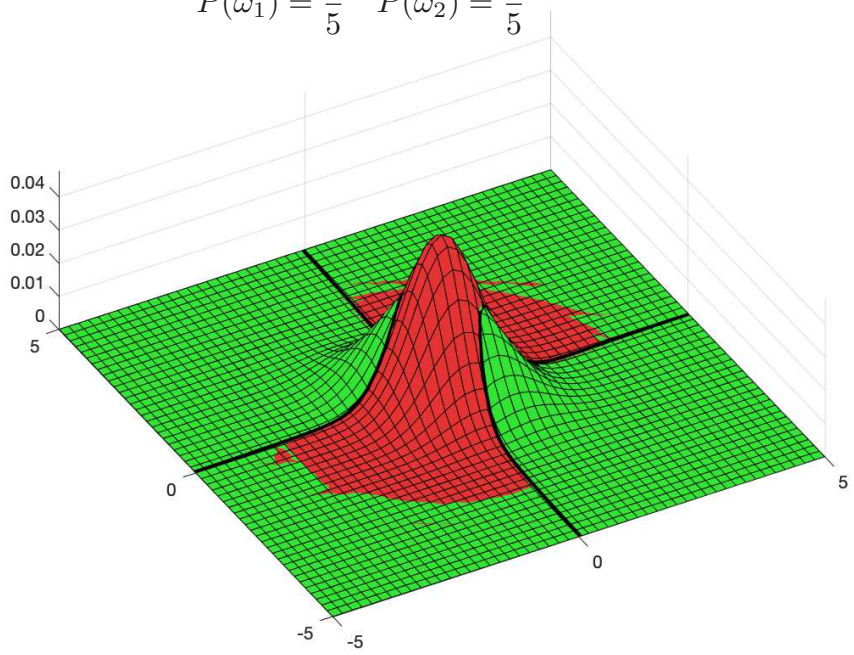
Discriminants dans le cas de lois normales

Cas général

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$P(\omega_1) = \frac{3}{5} \quad P(\omega_2) = \frac{2}{5}$$



Discriminants dans le cas de lois normales

Cas général

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

$$P(\omega_1) = \frac{1}{5} \quad P(\omega_2) = \frac{4}{5}$$