2.3 Méthode probabiliste

2.3.1 Définitions

Avec la méthode probabiliste, on considère que le vecteur de caractéristique $\boldsymbol{x} = (x_1, ..., x_N)$ a une densité de probabilité (ddp) qui dépend de sa classe.

Exemple d'une ddp pour un vecteur x ?

$$P(\boldsymbol{x}) = \frac{1}{\left(\sqrt{2\pi\,|\boldsymbol{\Gamma}|}\right)^N} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T \boldsymbol{\Gamma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right]$$

Dans la suite de ce document, on utilise les notations suivantes

$$\int_{x_1} ... \int_{x_N} g(\boldsymbol{x})P(\boldsymbol{x})dx_1...dx_N = \int g(\boldsymbol{x})P(\boldsymbol{x})d\boldsymbol{x}$$

On a en particulier

$$\begin{array}{l} \int P(\boldsymbol{x})d\boldsymbol{x} = 1 \\ \int \boldsymbol{x}P(\boldsymbol{x})d\boldsymbol{x} = \boldsymbol{\mu} \\ \int \left(\boldsymbol{x}-\boldsymbol{\mu}\right)\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^T P(\boldsymbol{x})d\boldsymbol{x} = \boldsymbol{\Gamma} \end{array}$$

où $\boldsymbol{\mu}$ est le vecteur moyen et $\boldsymbol{\Gamma}$ est la matrice de covariance de $\boldsymbol{X}$.

Rappel. Un discriminateur est une application

$$\begin{array}{ccc} \mathbb{R}^N & \mapsto & \{\omega_1, \omega_2, ..., \omega_C\} \\ \boldsymbol{x} & \mapsto & \widehat{d}(\boldsymbol{x}) \end{array}$$

Avec la méthode probabiliste, comme au moment où le discriminateur doit effectuer la discrimination, celui-ci ne connaît pas la classe ω, celle-ci est supposée être une variable aléatoire discrète.

On suppose que $\forall \boldsymbol{x}$, il existe une unique classe associée à $\boldsymbol{x}$. Ainsi

$$P(\omega_i \text{ et } \omega_j|\boldsymbol{x}) = 0 \quad \text{pour i}\neq j \quad \text{non ambiguïté}$$

$$\sum_{c=1}^{C} P(\omega_c|\boldsymbol{x}) = 1 \qquad\qquad \text{exhaustivité}$$

Il y 2 propriétés importantes à connaître et savoir manipuler :

1. Loi de Bayes

$$\boxed{P(\boldsymbol{x},\omega_c) = P(\boldsymbol{x}|\omega_c)P(\omega_c) = P(\omega_c|\boldsymbol{x})P(\boldsymbol{x})}$$

2. Calcul des marginales (ou théorème des probabilités totales)

$$\boxed{P(\boldsymbol{x}) = \sum_{c=1}^{C} P(\boldsymbol{x},\omega_c) = \sum_{c=1}^{C} P(\boldsymbol{x}|\omega_c)P(\omega_c)}$$

et

$$\boxed{P(\omega_c) = \int P(\boldsymbol{x},\omega_c)d\boldsymbol{x} = \int P(\omega_c|\boldsymbol{x})P(\boldsymbol{x})d\boldsymbol{x}}$$

Quel est le sens des variables ci-dessous ?

- $P(\omega_c)$
- $P(\boldsymbol{x}|\omega_c)$
- $P(\omega_c|\boldsymbol{x})$
- $P(\boldsymbol{x})$

Pour chaque classe, on peut calculer la moyenne et la matrice de covariance intra-classe

$$\boxed{\begin{array}{l} \boldsymbol{\mu}_c = \int \boldsymbol{x}P(\boldsymbol{x}|\omega_c)d\boldsymbol{x} \\ \boldsymbol{\Gamma}_c = \int \left(\boldsymbol{x}-\boldsymbol{\mu}_c\right)\left(\boldsymbol{x}-\boldsymbol{\mu}_c\right)^T P(\boldsymbol{x}|\omega_c)d\boldsymbol{x} \end{array}}$$

à la condition que ces intégrales existent.

2.3.2 Application de la théorie de la décision dans le cas à deux classes

Quel est le critère à utiliser pour caractériser les performances d'un discriminateur ?

$$P_{err} = P(\text{faire une erreur}) = \int_{\mathbb{R}^N} P(\boldsymbol{x}, \text{faire une erreur}) d\boldsymbol{x}$$

Dans un problème à deux classes, on a 2 erreurs possibles

réalité	Probabilité	décision
$\boldsymbol{x} \in \omega_1$	$P(\boldsymbol{x}, \omega_1)$	$\boldsymbol{x} \mapsto \omega_2$
$\boldsymbol{x} \in \omega_2$	$P(\boldsymbol{x}, \omega_2)$	$\boldsymbol{x} \mapsto \omega_1$

Ainsi

$$P_{err} = \int_{\mathcal{R}_1} P(\boldsymbol{x}, \omega_2) d\boldsymbol{x} + \int_{\mathcal{R}_2} P(\boldsymbol{x}, \omega_1) d\boldsymbol{x}$$

où $\mathcal{R}_1$ et $\mathcal{R}_2$ correspondent à la partition liée à notre discriminateur.

Exemple. Voir Figure 10.

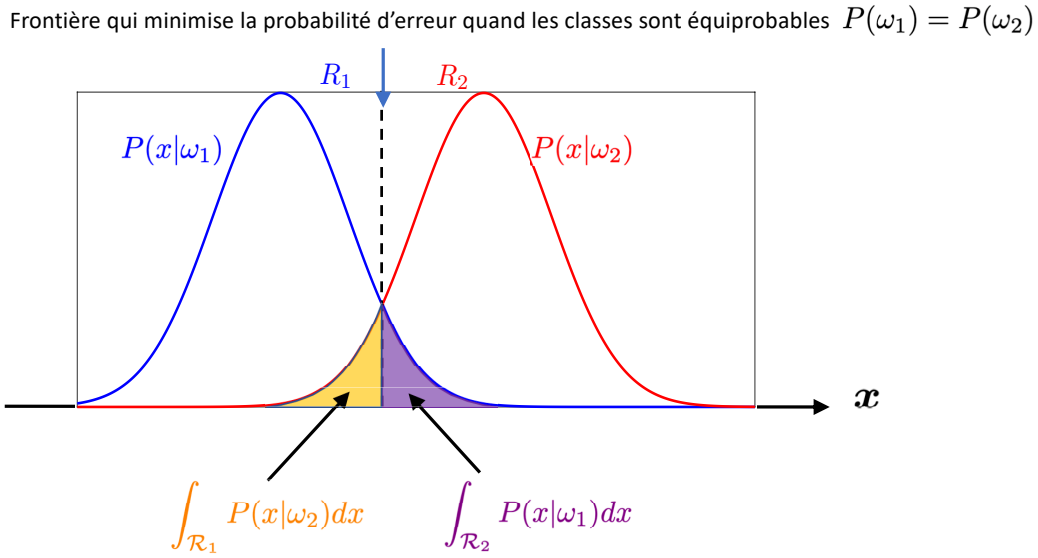

FIGURE 10 – Illustration dans le cas avec deux classes gaussiennes équiprobables.

On peut montrer que la probabilité d'erreur minimale est donnée par

$$P_{err}^{opt} = \int \min\left(P(\boldsymbol{x}, \omega_1), P(\boldsymbol{x}, \omega_2)\right) d\boldsymbol{x}$$

Ceci s'illustre en 1D sur la Figure 10 dans le cas équiprobable.

Ainsi la règle de décision qui minimise la probabilité d'erreur P_{err} est

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

ce qui est équivalent à

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

mais ce qui est aussi équivalent à

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

Que signifie le fait que cette règle de décision est optimale quand on choisit comme critère probabilité d'erreur ?

Que faire si on connaît $P(\boldsymbol{x}|\omega_i)$, mais on ne connaît pas $P(\omega_i)$?

2.3.3 Cas gaussien avec 2 classes équiprobables et paramètres connus

Si on suppose

$$P(\boldsymbol{x}|\omega_c) = \frac{1}{\sqrt{2\pi}^N \sqrt{|\boldsymbol{\Gamma}_c|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^T \boldsymbol{\Gamma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)\right]$$

Alors le discriminateur qui minimise la probabilité d'erreur

$$\ln\left(\frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_2)}\right) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$

peut s'écrire

$$\frac{1}{2}\ln\frac{|\boldsymbol{\Gamma}_2|}{|\boldsymbol{\Gamma}_1|} + \frac{1}{2}(Q_2(\boldsymbol{x}) - Q_1(\boldsymbol{x})) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$

où

$$Q_c(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu}_c)^T \boldsymbol{\Gamma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c)$$

Que représente $Q_c(\boldsymbol{x})$?

La frontière qui sépare les deux régions est une fonction quadratique. Elle peut décrire plusieurs formes :

— 1 ellipse avec une équation du type

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} = 1$$

— 1 hyperbole avec une équation du type

$$\frac{x_1^2}{a_1^2} - \frac{x_2^2}{a_2^2} = 1$$

— 1 parabole avec une équation du type

$$x_2^2 = a x_1$$

Dans le cas particulier où $\boldsymbol{\Gamma}_1 = \boldsymbol{\Gamma}_2 = \boldsymbol{\Gamma}$, le test se simplifie en

$$Q_2(\boldsymbol{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} Q_1(\boldsymbol{x})$$

ainsi

$$(\boldsymbol{x}-\boldsymbol{\mu}_2)^T \boldsymbol{\Gamma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} (\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Gamma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)$$

En développant, on obtient

$$-\boldsymbol{\mu}_2^T\boldsymbol{\Gamma}^{-1}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_2 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} -\boldsymbol{\mu}_1^T\boldsymbol{\Gamma}^{-1}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_1$$

et finalement

$$\boldsymbol{\mu}_2^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}_1 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 2\left(\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_1^T\right)\boldsymbol{\Gamma}^{-1}\boldsymbol{x}$$

De quel type de frontière s'agit-il ?

Quels sont les risques de cette approche ?

Remarque. Il est possible de chercher une frontière **linéaire**, sans avoir besoin de supposer que les classes sont gaussiennes. Pour cela, on peut repartir de la règle de décision

$$\ln\left(\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})}\right) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$

Si on suppose qu'il existe une frontière linéaire sur la statistique $\ln\left(\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})}\right)$, alors il existe w_0 et $\boldsymbol{w}$ tels que

$$\ln\left(\frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})}\right) = w_0 + \boldsymbol{w}^T\boldsymbol{x}$$

Or, comme

$$P(\omega_1|\boldsymbol{x}) + P(\omega_2|\boldsymbol{x}) = 1$$

on a donc

$$\ln\left(\frac{P(\omega_1|\boldsymbol{x})}{1 - P(\omega_1|\boldsymbol{x})}\right) = w_0 + \boldsymbol{w}^T\boldsymbol{x}$$

et donc

$$P(\omega_1|\boldsymbol{x}) = \frac{e^{w_0+\boldsymbol{w}^T\boldsymbol{x}}}{1 + e^{w_0+\boldsymbol{w}^T\boldsymbol{x}}}$$

et

$$P(\omega_2|\boldsymbol{x}) = 1 - P(\omega_1|\boldsymbol{x}) = \frac{1}{1 + e^{w_0+\boldsymbol{w}^T\boldsymbol{x}}}$$

C'est le discriminateur *logistique*.

Un intérêt de cette solution est que pour identifier $\boldsymbol{w}$, on peut utiliser comme critère

$$J(\boldsymbol{w}) = \sum_{\{p|d^{(p)}=1\}} \log \frac{e^{\boldsymbol{w}^T\boldsymbol{x}^{(p)}}}{1 + e^{\boldsymbol{w}^T\boldsymbol{x}^{(p)}}} + \sum_{\{p|d^{(p)}=-1\}} \log \frac{1}{1 + e^{\boldsymbol{w}^T\boldsymbol{x}^{(p)}}}$$

qui revient à maximiser une sorte de vraisemblance sur la base d'apprentissage.

On peut montrer que cette optimisation est bien plus facile à réaliser que celle du gradient par exemple.

Synthèse sur le linéaire. On a plusieurs solutions pour faire du linéaire.

- Hebb, pas de critère mathématique, il s'agit d'une solution plutôt empirique.
- pseudo-inverse (PI), le critère est

$$J = \sum_p \|d^{(p)} - \boldsymbol{w}^T\boldsymbol{x}^{(p)}\|^2$$

- RA et gradient utilisent le même critère, mais ils régularisent l'inversion de matrice.
- le cas gaussien avec $\boldsymbol{\Gamma}_1 = \boldsymbol{\Gamma}_2$,
- le logistique avec le critère ci-dessus.

On a donc 6 discriminateurs qui mènent à une frontière linéaire.

Est-ce que les solutions donnent des résultats identiques ?

2.3.4 Cas gaussien avec 2 classes équiprobables et paramètres inconnus

Si en pratique, on ne connaît pas les paramètres associés à chacune des lois $P(\boldsymbol{x}|\omega_c)$, il faut mettre en œuvre des estimateurs.

Quels estimateurs vous connaissez ?

Dans le cas gaussien où l'on ne suppose pas que les covariances sont égales, si on a N_c vecteur de la classe c dans la base d'apprentissage, alors les estimateurs s'écrivent $\forall c$

$$\widehat{\boldsymbol{\mu}}_c = \frac{1}{N_c}\sum_{p=1}^{N_c} \boldsymbol{x}^{(p)}$$

et

$$\widehat{\boldsymbol{\Gamma}}_c = \frac{1}{N_c}\sum_{p=1}^{N_c} \left(\boldsymbol{x}^{(p)} - \widehat{\boldsymbol{\mu}}_c\right)\left(\boldsymbol{x}^{(p)} - \widehat{\boldsymbol{\mu}}_c\right)^T$$

Si on sait que les covariances sont égales, alors on peut montrer que l'estimateur de la matrice de covariance $\widehat{\boldsymbol{\Gamma}}$ s'écrit alors

$$\widehat{\boldsymbol{\Gamma}} = \frac{1}{2}\left(\widehat{\boldsymbol{\Gamma}}_1 + \widehat{\boldsymbol{\Gamma}}_2\right)$$

Qu'est-ce qu'on perd quand les paramètres sont inconnus ?

Qu'est-ce qu'il convient de faire dans ce cas ?

2.3.5 Généralisation dans le cas à C classes

Dans le cas à C classes, il faut sommer toutes les erreurs possibles

$$P_{err} = \sum_{i=1}^{C} \int_{\mathcal{R}_i} \sum_{j \text{ avec } j \neq i} P(\boldsymbol{x}, \omega_j) d\boldsymbol{x}$$

ce qui peut s'écrire

$$P_{err} = \sum_{i=1}^{C} \int_{\mathcal{R}_i} \sum_{j} P(\boldsymbol{x}, \omega_j) d\boldsymbol{x} - \sum_{i=1}^{C} \int_{\mathcal{R}_i} P(\boldsymbol{x}, \omega_i) d\boldsymbol{x}$$

Or $\sum_j P(\boldsymbol{x}, \omega_j) d\boldsymbol{x} = P(\boldsymbol{x})$ et ainsi

$$P_{err} = \int_{\mathbb{R}} P(\boldsymbol{x}) d\boldsymbol{x} - \sum_{i=1}^{C} \int_{\mathcal{R}_i} P(\boldsymbol{x}, \omega_i) d\boldsymbol{x} = 1 - \sum_{i=1}^{C} \int_{\mathcal{R}_i} P(\boldsymbol{x}, \omega_i) d\boldsymbol{x}$$

Pour minimiser la probabilité d'erreur, il faut donc choisir comme partitionnement $\mathcal{R}_i$

$$\mathcal{R}_i = \{\boldsymbol{x} | P(\boldsymbol{x}, \omega_i) \geq P(\boldsymbol{x}, \omega_j) \quad \forall j\}$$

Si toutes les classes sont équiprobables, alors quelle la règle qui minimise la probabilité d'erreur ?