2 Illustration en science des données : la discrimination

Introduction

En science des données, à partir d'une observation $\boldsymbol{x}$ [1] bruitée par un bruit noté $\boldsymbol{b}$, il peut y avoir plusieurs tâches possibles à réaliser :

- débruitage : $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{b}$. Dans ce cas, on cherche une fonction f telle que $f(\boldsymbol{x}) \simeq \boldsymbol{s}$.
- estimation : $\boldsymbol{x} \sim P_{\boldsymbol{\theta}}$. Dans ce cas, on cherche une fonction f telle que $f(\boldsymbol{x}) \simeq \boldsymbol{\theta}$.
- détection : $H_0 \ : \ \boldsymbol{x} = \boldsymbol{b}$ ou $H_1 \ : \ \boldsymbol{x} = \boldsymbol{s} + \boldsymbol{b}$. Dans ce cas, on cherche une fonction f et un seuil ρ tels que $f(\boldsymbol{x}) \underset{H_1}{\overset{H_0}{\gtrless}} \rho$ permette de discriminer correctement entre l'hypothèse H_0 et H_1.

Ce cours porte sur la discrimination
Cela signifie qu'à partir de la connaissance de $\boldsymbol{x}$, il faut retrouver à quelle classe le vecteur $\boldsymbol{x}$ appartient.
On peut donc représenter un discriminateur comme une application

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

où pour simplifier on a supposé que $\boldsymbol{x} \in \mathbb{R}^N$, où les ω_i sont les différentes classes possibles et $\widehat{d}(\boldsymbol{x})$ est la classe estimée. Par exemple, on trouvera des situations pour lesquelles on a

$$\widehat{d}(\boldsymbol{x}) = \arg \max_{c=1,2,...,C} f(\boldsymbol{x}; c)$$

où f est une fonction qui reste à déterminer et qui dépend de $\boldsymbol{x}$ et de c.

Remarque. En pratique, le vecteur $\boldsymbol{x}$ peut correspondre à toute sorte de données : image, spectre, son audio, ADN, cours de la bourse, vidéo, ...

Exemple. On considère un problème où une photo représente soit un chien, soit un chat. On souhaite faire un système capable de distinguer à partir d'une photo s'il s'agit d'un chien ou d'un chat.

Que signifie résoudre ce problème de discrimination par apprentissage ?
Cela signifie que pour trouver l'application $\widehat{d}$, on utilise une base de données pour laquelle le problème a déjà été résolu. C'est à dire qu'on dispose d'un ensemble

$$\mathcal{B} = \{(\boldsymbol{x}^{(p)}, d^{(p)}); p \in \{1, ..., P\}\}$$

où $d^{(p)}$ représente la classe de $\boldsymbol{x}^{(p)}$ (on parle aussi souvent d'étiquette) et P représente la taille de la base de données.

On appelle $\mathcal{B}$ la base d'apprentissage car celle-ci est utilisée pour **apprendre** la fonction $\widehat{d}$.

Est-il possible d'effectuer de la discrimination en aveugle (i.e. sans base d'apprentissage) ?

Dans ce cours, on s'intéressera uniquement à des algorithmes de discrimination avec base d'apprentissage (on dit aussi classification supervisée).

Quels sont les objectifs de ce cours ?
Donner aux élèves les éléments méthodologiques leur permettant d'analyser les comportements de discriminateurs, et en particulier d'analyser leurs performances.

Pourquoi se focaliser sur l'étude des performances ?
Parce qu'en tant qu'ingénieur généraliste, les élèves seront utilisateur de ces techniques, et pas forcément développeur. Sans être des spécialistes, ils doivent savoir en analyser les performances.

1. Dans la partie 2 du cours, on choisit la convention suivante : les vecteurs sont notés en gras (exemple $\boldsymbol{x}$), les matrices sont notées en gras majuscules (exemple $\boldsymbol{X}$).

2.1 Discrimination linéaire

2.1.1 Définitions

Soit $\boldsymbol{x} \in \mathcal{S} \subset \mathbb{R}^N$, un vecteur de caractéristiques.

Trouver une technique de classification revient à partitionner l'espace $\mathcal{S}$. On considère des partitionnements avec les propriétés ci-dessous :

$$\begin{array}{ll} \cup_n \mathcal{R}_n = \mathcal{S} & \text{exhaustivité} \\ \mathcal{R}_n \cap \mathcal{R}_{n'} = \emptyset & \text{non ambiguïté} \end{array}$$

et on définit la règle de décision suivante :

$$\text{Si } \boldsymbol{x} \in \mathcal{R}_n, \text{ alors on attribue la classe } \omega_n$$

Dans le reste de ce chapitre, on considère le cas à 2 classes :

$$\begin{array}{ll} \mathcal{R}_1 \cup \mathcal{R}_2 = \mathcal{S} & \text{exhaustivité} \\ \mathcal{R}_1 \cap \mathcal{R}_2 = \emptyset & \text{non ambiguïté} \end{array}$$

Définition. Un discriminateur linéaire signifie que l'espace $\mathcal{S}$ est partitionné avec des frontières décrites par des hyper-plans.

Remarque. La discrimination linéaire est souvent le premier outil à utiliser car il est simple à mettre en œuvre et est très facile à interpréter. On commence donc par celui-ci. Ensuite, on verra des méthodes plus compliquées comme l'approche probabiliste ou les réseaux de neurones.

Mathématiquement la règle de décision associée à une frontière linéaire peut être décrite par

$$x_1 w_1 + ... + x_N w_N \underset{\omega_2}{\overset{\omega_1}{\gtrless}} b$$

où $\boldsymbol{x} = (x_1, ..., x_N)^T$ est le vecteur de caractéristiques et

$$x_1 w_1 + ... + x_N w_N = b$$

est l'équation de l'hyper-plan qui joue le rôle de frontière. Le vecteur $\boldsymbol{w} = (w_1, ..., w_N)^T$ est orthogonal à cet hyperplan.

Plutôt que faire apparaître le seuil b, en général on simplifie les notations avec les transformations :

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ ... \\ x_N \end{pmatrix} \mapsto \boldsymbol{x}' = \begin{pmatrix} 1 \\ x_1 \\ ... \\ x_N \end{pmatrix}$$

et

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ ... \\ w_N \end{pmatrix} \mapsto \boldsymbol{w}' = \begin{pmatrix} w_0 \\ w_1 \\ ... \\ w_N \end{pmatrix}$$

où $b = -w_0$. Ainsi, l'équation de la frontière

$$x_1 w_1 + ... + x_n w_n = b$$

qui s'écrit

$$\boldsymbol{w}^T \boldsymbol{x} = b$$

devient simplement

$$(\boldsymbol{w}')^T \boldsymbol{x}' = 0$$

Dans la suite, on ne notera plus les $'$.

Synthèse : à partir d'une base d'apprentissage

$$\boxed{\mathcal{B}_{app} = \{(\boldsymbol{x}^{(p)}, d^{(p)}); p \in \{1, ..., P_{app}\}\}}$$

où P_{app} est la taille de la base d'apprentissage, un discriminateur linéaire est simplement défini par un vecteur $\boldsymbol{w}$ qui est une fonction de $\mathcal{B}_{app}$ et la règle de décision s'écrit simplement

$$\boxed{\boldsymbol{w}^T \boldsymbol{x} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0 \quad \forall \boldsymbol{x}}$$

Dans la suite de ce chapitre, on considère le cas à 2 classes en choisissant la convention suivante :

Si $\boldsymbol{x}^{(p)}$ appartient à la classe ω_1, alors $d^{(p)} = 1$
Si $\boldsymbol{x}^{(p)}$ appartient à la classe ω_2, alors $d^{(p)} = -1$

2.1.2 Méthode de Hebb

Cette méthode établie par M. Hebb en 1949 est très simple

$$\boldsymbol{w}_{\text{Hebb}} = \sum_{p=1}^{P_{app}} d^{(p)} \boldsymbol{x}^{(p)}$$

où la somme est effectuée sur tous les vecteurs de la base d'apprentissage. Ainsi

$$\forall \boldsymbol{x} \quad \widehat{d}_{\text{Hebb}}(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}_{\text{Hebb}}^T \boldsymbol{x})$$

Et ainsi

$$\boldsymbol{w}_{\text{Hebb}}^T \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{w}_{\text{Hebb}} = \sum_{p=1}^{P_{app}} d^{(p)} \boldsymbol{x}^T \boldsymbol{x}^{(p)}$$

Comment interpréter cette relation ?

2.1.3 Méthode de la pseudo-inverse

Plutôt qu'une méthode empirique comme celle de Hebb, on analyse à présent une méthode "mathématique" pour trouver un discriminateur. Avoir une méthode mathématique signifie qu'elle est fondé sur la minimisation d'un critère.

Avec la pseudo-inverse, on utilise comme critère

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

Pourquoi choisir ce critère ?

Si on concatène les vecteurs de la base d'apprentissage pour créer une matrice $\boldsymbol{X}$ de taille $N \times P_{app}$

$$\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(P_{app})}\right)$$

et que l'on définit le vecteur $\boldsymbol{d} = \left(d^{(1)}, ..., d^{(P_{app})}\right)^T$ alors on peut montrer que

$$J = \left(\boldsymbol{d} - \boldsymbol{X}^T \boldsymbol{w}\right)^T \left(\boldsymbol{d} - \boldsymbol{X}^T \boldsymbol{w}\right)$$

En effet, on a

$$\boldsymbol{d} - \boldsymbol{X}^T \boldsymbol{w} = \begin{pmatrix} d^{(1)} - \left(\boldsymbol{x}^{(1)}\right)^T \boldsymbol{w} \\ d^{(2)} - \left(\boldsymbol{x}^{(2)}\right)^T \boldsymbol{w} \\ ... \\ d^{(P_{app})} - \left(\boldsymbol{x}^{(P_{app})}\right)^T \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} d^{(1)} - \boldsymbol{w}^T \boldsymbol{x}^{(1)} \\ d^{(2)} - \boldsymbol{w}^T \boldsymbol{x}^{(2)} \\ ... \\ d^{(P_{app})} - \boldsymbol{w}^T \boldsymbol{x}^{(P_{app})} \end{pmatrix}$$

Vu le critère J, il est clair que l'on cherche le vecteur $\boldsymbol{w}$ tel que

$$\boxed{\boldsymbol{X}^T \boldsymbol{w} = \boldsymbol{d}} \tag{35}$$

$\boldsymbol{X}^T$ est une matrice rectangulaire, il faut calculer $\left(\boldsymbol{X}^T\right)^*$ la pseudo-inverse de $\boldsymbol{X}^T$

$$\boldsymbol{w} = \left(\boldsymbol{X}^T\right)^* \boldsymbol{d}$$

Pour calculer la pseudo-inverse, on peut remarquer qu'en multipliant l'équation (35) à gauche par $\boldsymbol{X}$, on obtient

$$\boldsymbol{X}\,\boldsymbol{X}^T \boldsymbol{w} = \boldsymbol{X}\,\boldsymbol{d}$$

Si $\boldsymbol{X}\,\boldsymbol{X}^T$ est inversible, on trouve ainsi

$$\boxed{\boldsymbol{w}_{PI} = \left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^{-1} \boldsymbol{X}\,\boldsymbol{d}}$$

Ce qui signifie que la pseudo inverse de $\boldsymbol{X}^T$ est dans ce cas

$$\left(\boldsymbol{X}^T\right)^* = \left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^{-1} \boldsymbol{X}$$

Est-ce que $\boldsymbol{X}\,\boldsymbol{X}^T$ est toujours inversible ?
$\boldsymbol{X}$ est une matrice de taille $N \times P_{app}$, la matrice $\boldsymbol{X}\,\boldsymbol{X}^T$ est de taille $N \times N$ et peut s'écrire

$$\boldsymbol{X}\,\boldsymbol{X}^T = \sum_{p=1}^{P_{app}} \boldsymbol{x}^{(p)} \left(\boldsymbol{x}^{(p)}\right)^T$$

De plus, $\boldsymbol{X}\,\boldsymbol{X}^T$ est symétrique, donc elle est décomposable dans une base orthogonale de vecteurs propres

$$\boldsymbol{X}\,\boldsymbol{X}^T = \boldsymbol{U} \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ 0 & . & . & . \\ 0 & . & . & \lambda_N \end{pmatrix} \boldsymbol{U}^T$$

où $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$ et $\boldsymbol{U}\,\boldsymbol{U}^T = \boldsymbol{Id}$. Ainsi, on a

$$\det(\boldsymbol{X}\,\boldsymbol{X}^T) = \prod_{n=1}^{N} \lambda_n$$

A quelle condition sur λ la matrice $\boldsymbol{X}\,\boldsymbol{X}^T$ est inversible ?

De plus, en introduisant les vecteurs colonnes $\boldsymbol{U} = (\boldsymbol{u}_1, ..., \boldsymbol{u}_N)$, on peut aussi écrire

$$\boldsymbol{X}\,\boldsymbol{X}^T = \sum_{n=1}^{N} \lambda_n \boldsymbol{u}_n \boldsymbol{u}_n^T$$

Si $P_{app} < N$, alors comme on a

$$\boldsymbol{X}\,\boldsymbol{X}^T = \sum_{p=1}^{P_{app}} \boldsymbol{x}^{(p)} \left(\boldsymbol{x}^{(p)}\right)^T$$

il est clair qu'on a forcément $\lambda_n = 0$ pour $n > P_{app}$. Donc dans ce cas, $\det(\boldsymbol{X}\,\boldsymbol{X}^T) = 0$ et la matrice est singulière.

Comment faire dans ce cas ?
On considère la décomposition en valeurs singulières

$$\boldsymbol{X} = \boldsymbol{U}\,\boldsymbol{\Delta}_{\delta_{1:r}}\,\boldsymbol{V}^T$$

avec

$$\boldsymbol{\Delta}_{\delta_{1:r}} = \begin{pmatrix} \delta_1 & 0 & ... & 0 & 0 & . & 0 \\ 0 & \delta_2 & ... & 0 & 0 & . & 0 \\ 0 & . & . & 0 & 0 & . & 0 \\ 0 & . & . & \delta_r & 0 & . & 0 \\ 0 & . & . & 0 & 0 & . & 0 \\ . & . & . & . & 0 & . & 0 \\ 0 & . & . & 0 & 0 & . & 0 \end{pmatrix}$$

où r est donc le rang de la matrice.
Ainsi l'équation

$$\boldsymbol{X}\,\boldsymbol{X}^T\boldsymbol{w} = \boldsymbol{X}\,\boldsymbol{d}$$

s'écrit

$$\boldsymbol{U}\,\boldsymbol{\Delta}_{\delta_{1:r}}\boldsymbol{\Delta}_{\delta_{1:r}}^T\boldsymbol{U}^T\boldsymbol{w} = \boldsymbol{U}\,\boldsymbol{\Delta}_{\delta_{1:r}}\boldsymbol{V}^T\boldsymbol{d} \tag{36}$$

On définit

$$\boldsymbol{w} = \boldsymbol{U}\,\boldsymbol{\Delta}_{1/\delta_{1:r}}\boldsymbol{V}^T\boldsymbol{d}$$

Est-ce que ce vecteur $\boldsymbol{w}$ est solution de l'équation (36) ?

En pratique, pour calculer $\boldsymbol{w}$, plutôt que de calculer la décomposition en valeurs singulières de $\boldsymbol{X}$, on préfère souvent utiliser la décomposition en valeurs propres

$$\boldsymbol{X}\,\boldsymbol{X}^T = \boldsymbol{U}\,\boldsymbol{\Delta}_{\lambda_{1:r}}\boldsymbol{U}^T$$

Quel est le lien entre λ_n et δ_n ?

A partir de cette décomposition, il est possible de définir

$$\left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^* = \boldsymbol{U}\,\boldsymbol{\Delta}_{1/\lambda_{1:r}}\boldsymbol{U}^T$$

pour finalement retrouver $\boldsymbol{w}$ avec

$$\boldsymbol{w} = \left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^*\boldsymbol{X}\boldsymbol{d}$$

car en effet, on a :

$$\left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^*\boldsymbol{X}\,\boldsymbol{d} = \boldsymbol{U}\,\boldsymbol{\Delta}_{1/\lambda_{1:r}}\boldsymbol{U}^T\boldsymbol{U}\,\boldsymbol{\Delta}_{\delta_{1:r}}\boldsymbol{V}^T\boldsymbol{d} = \boldsymbol{U}\,\boldsymbol{\Delta}_{1/\delta_{1:r}}\boldsymbol{V}^T\boldsymbol{d}$$

Ainsi dans le cas où $\boldsymbol{X}\,\boldsymbol{X}^T$ n'est pas inversible, la solution s'écrit simplement

$$\boxed{\boldsymbol{w} = \left(\boldsymbol{X}\,\boldsymbol{X}^T\right)^*\boldsymbol{X}\,\boldsymbol{d}}$$

Cette solution sera analysé lors du prochain TP et comparé à celle de M. Hebb.

2.1.4 Analyse des performances (avec barre d'erreur)

On dispose d'une base $\mathcal{B} = \{(x^{(p)}, d^{(p)}); p \in \{1, ..., P\}\}$.

Comment faire pour analyser les performances d'un discriminateur ?
La solution la plus courante consiste à couper la base en deux. On utilise une partie pour réaliser l'apprentissage (i.e. apprendre $\boldsymbol{w}$) et une autre partie pour réaliser l'analyse des performances de ce discriminateur.

Mathématiquement on définit une partition $\{1, ..., P\} = I_{app} \cup I_{gen}$ avec $I_{app} \cap I_{gen} = \emptyset$.

Ainsi, on a une base d'apprentissage

$$\mathcal{B}_{app} = \{(\boldsymbol{x}^{(p)}, d^{(p)}); p \in I_{app}\}$$

pour apprendre à discriminer et une base de généralisation

$$\mathcal{B}_{gen} = \{(\boldsymbol{x}^{(p)}, d^{(p)}); p \in I_{gen}\}$$

pour analyser les performances.

Dans le reste du cours, on note P_{app} la taille de la base d'apprentissage et P_{gen} la taille de la base de généralisation.

Pourquoi on coupe la base en 2 ?

Comment analyser les performances d'un discriminateur ?
Pour n'importe quel vecteur $\boldsymbol{x}^{(p)}$ de la base, on peut définir le score du discriminateur $\widehat{d}$ avec

$$\eta(\boldsymbol{x}^{(p)}) = \begin{cases} 1 \text{ si } \widehat{d}(\boldsymbol{x}^{(p)}) = d^{(p)} \\ 0 \text{ si } \widehat{d}(\boldsymbol{x}^{(p)}) \neq d^{(p)} \end{cases}$$

On peut alors calculer le taux de réussite obtenu sur la base d'apprentissage

$$\tau_{app} = \frac{1}{P_{app}} \sum_{p \in I_{app}} \eta(\boldsymbol{x}^{(p)})$$

On appellera ce taux, le *taux d'apprentissage*.

Néanmoins, ce qui nous intéresse pour caractériser les performances, c'est plutôt le taux de réussite obtenu sur la base de généralisation

$$\tau_g = \frac{1}{P_{gen}} \sum_{p \in I_{gen}} \eta(\boldsymbol{x}^{(p)})$$

que l'on appelle *taux de généralisation*.

Cette fois, la somme est réalisée sur les vecteurs de la base de généralisation, et ceux-ci sont **différents** de ceux de la base d'apprentissage.

Le problème pratique est le suivant : **quand on a deux discriminateurs (par exemple en TP, vous aurez $\tau_g^{(Hebb)}$ et $\tau_g^{(PI)}$), comment savoir si les différences de performances sont significatives ?**

Pour répondre à cette question, il faut aborder des questions préliminaires.

Comment représenter $\eta(\boldsymbol{x}^{(p)})$?

Quelle est la loi suivi par $\eta(\boldsymbol{x}^{(p)})$?

Par conséquent **comment représenter τ_g ?**

Quelle est la loi suivie par τ_g ?

Finalement, on revient sur la question initiale, **quand on a deux discriminateurs (par exemple en TP, on aura $\tau_g^{(Hebb)}$ et $\tau_g^{(PI)}$), comment savoir si les différences de performances sont significatives ?**

Comment calculer les écart-types de τ_g ?

Question préliminaire : **que peut-on dire de la variance de τ_g ?**

$$\text{var}(\tau_g) = \langle \left(\frac{1}{P_{gen}} \sum_p \eta_p - \tau_{gen} \right)^2 \rangle$$

devient

$$\text{var}(\tau_g) = \frac{1}{P_{gen}^2} \langle \left(\sum_p (\eta_p - \tau_{gen}) \right)^2 \rangle$$

En supposant les η_p indépendants, on obtient

$$\text{var}(\tau_g) = \frac{1}{P_{gen}^2} \sum_p \langle (\eta_p - \tau_{gen})^2 \rangle$$

et comme les η_p sont identiquement distribués, on a

$$\text{var}(\tau_g) = \frac{P_{gen}}{P_{gen}^2} \langle (\eta_p - \tau_{gen})^2 \rangle = \frac{1}{P_{gen}} \text{var}(\eta_p)$$

Or pour une loi de Bernouilli, on a $\text{var}(\eta_p) = \tau_{gen}(1 - \tau_{gen})$, donc on obtient

$$\text{var}^{1/2}(\tau_g) = \sqrt{\frac{\tau_{gen}(1 - \tau_{gen})}{P_{gen}}}$$

Remarque. En pratique, on a pas accès à τ_{gen}. **Comment faire pour estimer $\text{var}^{1/2}(\tau_g)$?** (voir TP)