Apprentissage statistique supervisé : premières notions.

Présenté par Guillhem Artis & Clara Genes

École normale supérieure de Rennes

8 novembre 2022

Rappel du plan

Définitions informelles

Définition (apprentissage automatique)

- Observation d'un phénomène,
- construction d'un modèle de ce phénomène,
- prévision et analyse du phénomène grâce au modèle.

Le tout fait sans intervention humaine.

Présentation du problème général.

Entrées : Un échantillon (les données observées sur le phénomène) $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$ où $X_i \in \mathcal{X}$ (souvent $= \mathbb{R}^p$) et $Y_i \in \mathcal{Y}$ (souvent $= \mathbb{R}$).
Les X_i sont les *variables explicatives* et les Y_i sont les *variables à expliquer* (aussi appelées *étiquettes*).
On suppose que $(X_1, Y_1), \ldots, (X_n, Y_n)$ est une suite de variables *iid*.

Présentation du problème général.

Entrées : Un échantillon (les données observées sur le phénomène) $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$ où $X_i \in \mathcal{X}$ (souvent $= \mathbb{R}^p$) et $Y_i \in \mathcal{Y}$ (souvent $= \mathbb{R}$).
Les X_i sont les *variables explicatives* et les Y_i sont les *variables à expliquer* (aussi appelées *étiquettes*).
On suppose que $(X_1, Y_1), \ldots, (X_n, Y_n)$ est une suite de variables *iid*.

- Pourquoi des variables aléatoires ?
- Pourquoi indépendantes ?
- Pourquoi de même loi ?

Présentation du problème général.

Sorties : Une solution du problème de prévision est une application

$$f : \mathcal{X} \to \mathcal{Y}$$

appelée *prédicteur*. C'est une modélisation construite à partir de *l'ensemble d'apprentissage* D_n.

Présentation du problème général.

Sorties : Une solution du problème de prévision est une application

$$f : \mathcal{X} \to \mathcal{Y}$$

appelée *prédicteur*. C'est une modélisation construite à partir de *l'ensemble d'apprentissage* D_n.
Pour une nouvelle observation X_{n+1}, la valeur $f(X_{n+1}) \in \mathcal{Y}$ est un "bon" candidat pour "prévoir" la valeur de l'étiquette Y_{n+1} non observée.

Exemple

Exemple

Pour mieux comprendre :

- Phénomène étudié : marché immobilier,

Exemple

Exemple

Pour mieux comprendre :

- Phénomène étudié : marché immobilier,
- $\mathcal{X}$: l'ensemble des caractéristiques des logements sur Rennes,

Exemple

Exemple

Pour mieux comprendre :

- Phénomène étudié : marché immobilier,
- $\mathcal{X}$: l'ensemble des caractéristiques des logements sur Rennes,
- $X_i \in \mathcal{X}$: un p-uplet de données sur un logement à Rennes $(x_1, x_2, x_3, \dots) = (m^2$, nb de chambres, nb de salles de bains, ...),

Exemple

Exemple

Pour mieux comprendre :

- Phénomène étudié : marché immobilier,
- $\mathcal{X}$: l'ensemble des caractéristiques des logements sur Rennes,
- $X_i \in \mathcal{X}$: un p-uplet de données sur un logement à Rennes $(x_1, x_2, x_3, \dots) = (m^2$, nb de chambres, nb de salles de bains, ...),
- $\mathcal{Y}$: l'ensemble des prix possibles,

Exemple

Exemple

Pour mieux comprendre :

- Phénomène étudié : marché immobilier,
- $\mathcal{X}$: l'ensemble des caractéristiques des logements sur Rennes,
- $X_i \in \mathcal{X}$: un p-uplet de données sur un logement à Rennes $(x_1, x_2, x_3, \dots) = (m^2$, nb de chambres, nb de salles de bains, ...),
- $\mathcal{Y}$: l'ensemble des prix possibles,
- $Y_i \in \mathcal{Y}$: le prix donné au logement X_i.

Exemple

Exemple

Pour mieux comprendre :

- Phénomène étudié : marché immobilier,
- $\mathcal{X}$: l'ensemble des caractéristiques des logements sur Rennes,
- $X_i \in \mathcal{X}$: un p-uplet de données sur un logement à Rennes $(x_1, x_2, x_3, \dots) = (m^2$, nb de chambres, nb de salles de bains, ...),
- $\mathcal{Y}$: l'ensemble des prix possibles,
- $Y_i \in \mathcal{Y}$: le prix donné au logement X_i.
- f : une fonction qui, à partir d'un p-uplet de données sur un logement à Rennes, donne une "bonne" estimation de son prix.

But pricipal

- **Optimiser le modèle** : rendre la valeur $f(X_{n+1})$ la plus proche possible de la vraie valeur qu'aurait Y_{n+1} si elle avait été observée.

But pricipal

- **Optimiser le modèle** : rendre la valeur $f(X_{n+1})$ la plus proche possible de la vraie valeur qu'aurait Y_{n+1} si elle avait été observée.

Définition (fonction de coût, perte)

Soit une fonction mesurable

$$c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

telle que $c(y, y')$ soit d'autant plus petit que y et y' sont similaires et telle que

$$\forall y, y' \in \mathcal{Y}, c(y, y') \geqslant 0 \text{ et } c(y, y) = 0.$$

But pricipal

- **Optimiser le modèle** : rendre la valeur $f(X_{n+1})$ la plus proche possible de la vraie valeur qu'aurait Y_{n+1} si elle avait été observée.

Définition (fonction de coût, perte)

Soit une fonction mesurable

$$c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

telle que $c(y, y')$ soit d'autant plus petit que y et y' sont similaires et telle que

$$\forall y, y' \in \mathcal{Y}, c(y, y') \geqslant 0 \text{ et } c(y, y) = 0.$$

Objectif : fournir un prédicteur f tel que $c(f(X_{n+1}), Y_{n+1})$ soit le plus petit "en moyenne".

Risque d'un prédicateur

Définition (risque d'un prédicteur)

On définit le risque d'un prédicteur $f \in \mathcal{F}$ par :

$$\mathcal{R}(f) = \mathcal{R}_{\mathbb{P}}(f) := \mathbb{E}\left[c(f(X), Y)\right]$$

Risque d'un prédicateur

Définition (risque d'un prédicteur)

On définit le risque d'un prédicteur $f \in \mathcal{F}$ par :

$$\mathcal{R}(f) = \mathcal{R}_{\mathbb{P}}(f) := \mathbb{E}\left[c(f(X), Y)\right]$$

Remarque

Le risque dépend de la fonction de coût c et de la mesure $\mathbb{P}$.

Risque d'un prédicateur

Définition (risque d'un prédicteur)

On définit le risque d'un prédicteur $f \in \mathcal{F}$ par :

$$\mathcal{R}(f) = \mathcal{R}_{\mathbb{P}}(f) := \mathbb{E}\left[c(f(X), Y)\right]$$

Remarque

Le risque dépend de la fonction de coût c et de la mesure $\mathbb{P}$.

Remarque

L'espérance n'est pas forcément finie.

Risque d'un prédicateur

Définition (risque d'un prédicteur)

On définit le risque d'un prédicteur $f \in \mathcal{F}$ par :

$$\mathcal{R}(f) = \mathcal{R}_{\mathbb{P}}(f) := \mathbb{E}\left[c(f(X), Y)\right]$$

Remarque

Ici, la fonction f n'est pas aléatoire, en fait la définition ci dessus doit se lire

$$\mathcal{R}(f) = \mathcal{R}_{\mathbb{P}}(f) := \mathbb{E}\left[c(f(X), Y) | \, D_n\right].$$

Tout ça pour quoi ?

Problème de prévision :

Il s'agit de trouver, à l'aide d'un échantillon D_n seul, un prédicteur $f \in \mathcal{F}$ tel que son risque $\mathcal{R}_{\mathbb{P}}(f)$ est minimal.

Rappel du plan

Comment trouver un tel prédicteur ?

Le risque de Bayes

- Situation idéale : $\mathbb{P}$ est connue.

Notre problème $\rightsquigarrow$ *Un problème d'optimisation*

Le risque de Bayes

- Situation idéale : $\mathbb{P}$ est connue.

Notre problème $\rightsquigarrow$ *Un problème d'optimisation*

Définition (Risque de Bayes)

Le *risque de Bayes* est définie par :

$$\mathcal{R}^* = \mathcal{R}^*_{\mathbb{P}} := \inf_{f \in \mathcal{F}} \mathcal{R}_{\mathbb{P}}(f).$$

C'est la plus petite valeur de risque envisageable pour un prédicteur.

La règle de Bayes

Définition

Un prédicteur optimal (au sens "pour $\mathbb{P}$") est un prédicteur $f^* \in \mathcal{F}$ tel que

$$f^* \in \operatorname{argmin}_{f \in \mathcal{F}} \{\mathcal{R}(f)\}.$$

Il est appelé *règle (ou prédicteur) de Bayes*.

La règle de Bayes

Définition

Un prédicteur optimal (au sens "pour $\mathbb{P}$") est un prédicteur $f^* \in \mathcal{F}$ tel que

$$f^* \in \operatorname{argmin}_{f \in \mathcal{F}} \{\mathcal{R}(f)\}.$$

Il est appelé *règle (ou prédicteur) de Bayes*.

Remarque (Existence et unicité)

Un prédicteur de Bayes n'existe pas toujours et s'il existe il n'est pas forcément unique.

Excès de risque

Définition

L'*excès de risque* ou *risque relatif* d'un prédicteur $f \in \mathcal{F}$ est défini par :

$$\ell(f^*, f) := \mathcal{R}(f) - \mathcal{R}^* \geqslant 0.$$

Il dépend de $\mathbb{P}$ et de c.

Excès de risque

Définition

L'*excès de risque* ou *risque relatif* d'un prédicteur $f \in \mathcal{F}$ est défini par :

$$\ell(f^*, f) := \mathcal{R}(f) - \mathcal{R}^* \geqslant 0.$$

Il dépend de $\mathbb{P}$ et de c.

Avantages :

- existe toujours
- niveau minimal toujours nul : prévision idéale

Règle d'apprentissage

- Situation générale : $\mathbb{P}$ est inconnue et on dispose d'un échantillon D_n de taille $n \in \mathbb{N}^*$ finie mais non fixée *a priori*.

Règle d'apprentissage

- Situation générale : $\mathbb{P}$ est inconnue et on dispose d'un échantillon D_n de taille $n \in \mathbb{N}^*$ finie mais non fixée *a priori*.

Définition

Une solution du problème proposé est appelée *règle d'apprentissage* et est définie comme une fonction mesurable

$$\widehat{f} : \bigcup_{n \geqslant 1} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}.$$

Règle d'apprentissage

Notation

Soit $\widehat{f}$ une règle d'apprentissage et D_n un échantillon. Alors

- la fonction mesurable $\widehat{f}(D_n) \in \mathcal{F}$ est un prédicteur qu'on note abusivement $\widehat{f}$.

Règle d'apprentissage

Notation

Soit $\widehat{f}$ une règle d'apprentissage et D_n un échantillon. Alors

- la fonction mesurable $\widehat{f}(D_n) \in \mathcal{F}$ est un prédicteur qu'on note abusivement $\widehat{f}$.
- pour tout $x \in \mathcal{X}$, on note abusivement $\widehat{f}(x) := \widehat{f}(D_n; x)$ la valeur en x de $\widehat{f}$.

Risque d'une règle d'apprentissage

Définition

Le *risque* de $\widehat{f}$ s'écrit

$$\mathcal{R}(\widehat{f}) = \mathbb{E}\left[c(\widehat{f}(X), Y)|D_n\right]$$

Risque d'une règle d'apprentissage

Définition

Le *risque* de $\widehat{f}$ s'écrit

$$\mathcal{R}(\widehat{f}) = \mathbb{E}\left[c(\widehat{f}(X), Y)|D_n\right]$$

Définition

Le *risque moyen* de $\widehat{f}$ avec n observation indépendantes et de même loi $\mathbb{P}$ s'écrit

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] = \mathbb{E}\left[c(\widehat{f}(X), Y)\right]$$

Consistance

Définition (Consistance)

Soit $\widehat{f}$ une règle d'apprentissage, $\mathbb{P}$ une loi de probabilité sur $\mathcal{X} \times \mathcal{Y}$ et D_n un échantillon de $n \geqslant 1$ variables *iid* de loi $\mathbb{P}$. Alors

1. $\widehat{f}$ est *faiblement consistante pour* $\mathbb{P}$ si

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \underset{n \to +\infty}{\longrightarrow} \mathcal{R}^*,$$

Consistance

Définition (Consistance)

Soit $\widehat{f}$ une règle d'apprentissage, $\mathbb{P}$ une loi de probabilité sur $\mathcal{X} \times \mathcal{Y}$ et D_n un échantillon de $n \geqslant 1$ variables *iid* de loi $\mathbb{P}$. Alors

1. $\widehat{f}$ est *faiblement consistante pour* $\mathbb{P}$ si

$$\mathbb{E}\left[\mathcal{R}(\widehat{f})\right] \underset{n \to +\infty}{\longrightarrow} \mathcal{R}^*,$$

2. $\widehat{f}$ est *fortement consistante pour* $\mathbb{P}$ si

$$\mathcal{R}(\widehat{f}) \underset{n \to +\infty}{\overset{p.s.}{\longrightarrow}} \mathcal{R}^*.$$

Consistance

Définition (Consistance)

Soit $\widehat{f}$ une règle d'apprentissage, $\mathbb{P}$ un loi de probabilité sur $\mathcal{X} \times \mathcal{Y}$ et D_n un échantillon de $n \geqslant 1$ variables *iid* de loi $\mathbb{P}$. Alors

1. $\widehat{f}$ est *faiblement consistante pour* $\mathbb{P}$ si

$$\mathcal{R}(\widehat{f}) - \mathcal{R}^* \xrightarrow{L^1} 0,$$

2. $\widehat{f}$ est *fortement consistante pour* $\mathbb{P}$ si

$$\mathcal{R}(\widehat{f}) - \mathcal{R}^* \xrightarrow{p.s.} 0.$$

Rappel du plan

Définition (Régression)

Lorsque Y est continue et univariée, on parle de *régression* pour parler de l'apprentissage par statistique supervisée.

Définition (Régression)

Lorsque Y est continue et univariée, on parle de *régression* pour parler de l'apprentissage par statistique supervisée.

Définition (Fonction de régression)

Si Y possède une espérance, on appelle *fonction de régression* l'application $\mathbb{P}_X$-presque sûrement unique :

$$\begin{array}{rccc} \eta & : & \mathcal{X} & \to & \mathbb{R} \\ & & X & \mapsto & \mathbb{E}[Y|X]. \end{array}$$

Définition (Régression)

Lorsque Y est continue et univariée, on parle de *régression* pour parler de l'apprentissage par statistique supervisée.

Définition (Fonction de régression)

Si Y possède une espérance, on appelle *fonction de régression* l'application $\mathbb{P}_X$-presque sûrement unique :

$$\begin{array}{rccc} \eta & : & \mathcal{X} & \to & \mathbb{R} \\ & & X & \mapsto & \mathbb{E}[Y|X]. \end{array}$$

On pose alors le *bruit* ε défini par :

$$\varepsilon(X) := Y - \eta(X)$$

Définition (Fonction de régression)

Si Y possède une espérance, on appelle *fonction de régression*, l'application $\mathbb{P}_X$-presque sûrement unique :

$$\begin{array}{rccc} \eta & : & \mathcal{X} & \to & \mathbb{R} \\ & & X & \mapsto & \mathbb{E}[Y|X]. \end{array}$$

On pose alors le *bruit* ε défini par :

$$\varepsilon := Y - \eta(X)$$

Remarque

Ces définitions nous permettent d'écrire

$$Y = \eta(X) + \varepsilon \text{ avec } \mathbb{E}[\varepsilon|X] = 0 \text{ p.s.}$$

Règle de régression par partition

Exemple (règle de régression par partition)

Soit $\mathcal{A}$ une partition au plus dénombrable et mesurable de $\mathcal{X}$. Pour tout $x \in \mathcal{X}$, on note $\mathcal{A}(x)$ l'unique élément de $\mathcal{A}$ qui contient x.
On pose, pour tout $n \in \mathbb{N}^*$, tout $x \in \mathcal{X}$ et tout $(x_i, y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$:

$$\widehat{f}_{\mathcal{A}}^{p-r}((x_i, y_i)_{1 \leqslant i \leqslant n}; x) = \frac{1}{\underbrace{|\{i \in [\![1, n]\!] / x_i \in \mathcal{A}(x)\}|}_{N_{\mathcal{A}(x)}(x_1, \ldots, x_n)}} \sum_{i=1}^{n} y_i \mathbb{1}_{x_i \in \mathcal{A}(x)}$$

avec la convention $\frac{0}{0} = 0$.

Ceci définit la *règle de régression par partition* associée à la partition $\mathcal{A}$, notée $\widehat{f}_{\mathcal{A}}^{p-r}$.

Règle de régression par partition cubique

Exemple (Règle de régression par partition cubique)

Lorsque $\mathcal{X}$ est un sous ensemble quelconque de $\mathbb{R}^p$, pour $p \in \mathbb{N}^*$ et $h > 0$ on définit la *partition cubique de pas h* par :

$$\mathcal{A}^{\mathrm{cub}}(h) := \left(\mathcal{X} \cap \prod_{i=1}^{p} [hk_i, h(k_i + 1)] \right)_{(k_1,\ldots,k_p)\in\mathbb{Z}^p}$$

La règle de partition associée à $\mathcal{A}^{\mathrm{cub}}(h)$, notée $\widehat{f}_h^{cub-r}$, est appelée *règle par partition cubique de pas h*.

Un coût classique en régression

Définition

On appelle *coût quadratique*, ou *coût des moindres carrés*, la fonction donnée par :

$$\begin{array}{rccc} c & : & \mathbb{R}^2 & \to & \mathbb{R}^+ \\ & & (y, y') & \mapsto & (y - y')^2. \end{array}$$

Un coût classique en régression

Définition

On appelle *coût quadratique*, ou *coût des moindres carrés*, la fonction donnée par :

$$\begin{array}{rccc} c & : & \mathbb{R}^2 & \to & \mathbb{R}^+ \\ & & (y, y') & \mapsto & (y - y')^2. \end{array}$$

Définition

En supposant que $\mathbb{E}[Y^2] < +\infty$, le *risque quadratique* associé au coût quadratique vaut donc

$$\forall f \in \mathcal{F}, \qquad \mathcal{R}(f) := \mathbb{E}\left[(f(X) - Y)^2\right].$$

Un coût classique en régression

Proposition

Si $\mathcal{Y} = \mathbb{R}$ et $\mathbb{E}[(Y - \eta(X))^2] < \infty$, alors pour le coût quadratique :

1. un prédicteur f est un prédicteur de Bayes si et seulement si :

$$f(X) = \eta(X) \quad \mathbb{P}_X - \text{p.s.},$$

Un coût classique en régression

Proposition

Si $\mathcal{Y} = \mathbb{R}$ et $\mathbb{E}[(Y - \eta(X))^2] < \infty$, alors pour le coût quadratique :

1. un prédicteur f est un prédicteur de Bayes si et seulement si :

$$f(X) = \eta(X) \quad \mathbb{P}_X - \text{p.s.},$$

2. le rique de Bayes vaut :

$$\mathcal{R}^* = \mathbb{E}\left[(Y - \eta(X))^2\right] = \mathbb{E}\left[var(Y|X)\right] = \mathbb{E}\left[\varepsilon^2\right],$$

Un coût classique en régression

Proposition

Si $\mathcal{Y} = \mathbb{R}$ et $\mathbb{E}[(Y - \eta(X))^2] < \infty$, alors pour le coût quadratique :

1. un prédicteur f est un prédicteur de Bayes si et seulement si :

$$f(X) = \eta(X) \quad \mathbb{P}_X - \text{p.s.},$$

2. le rique de Bayes vaut :

$$\mathcal{R}^* = \mathbb{E}\left[(Y - \eta(X))^2\right] = \mathbb{E}\left[var(Y|X)\right] = \mathbb{E}\left[\varepsilon^2\right],$$

3. l'excès de risque de tout prédicteur $f \in \mathcal{F}$ s'écrit :

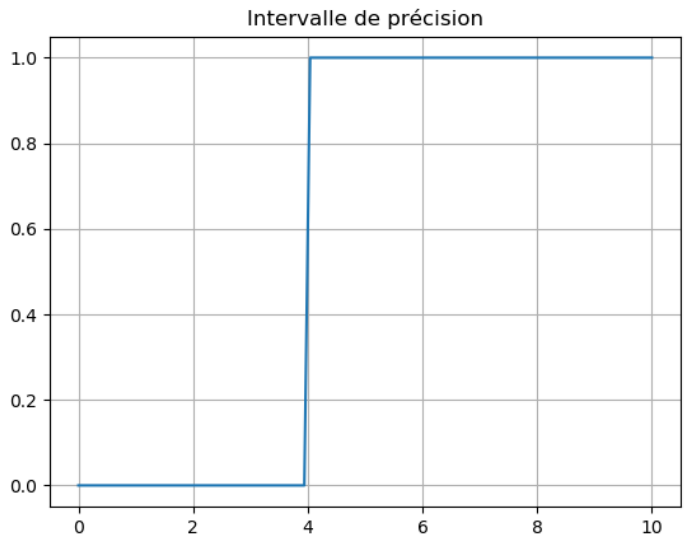
$$\ell(f^*, f) = \mathbb{E}\left[(f(X) - \eta(X))^2\right] = \|f - \eta\|^2_{L^2(P_X)}.$$

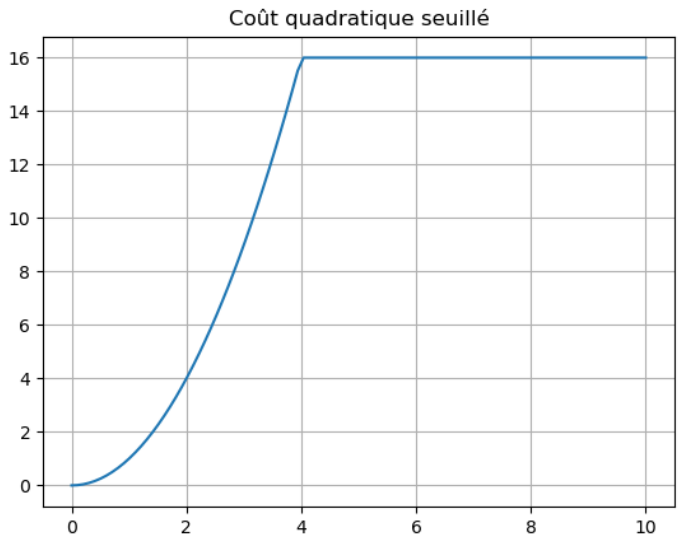
D'autres fonctions de coût

Voici des exemples de coûts sous la forme :

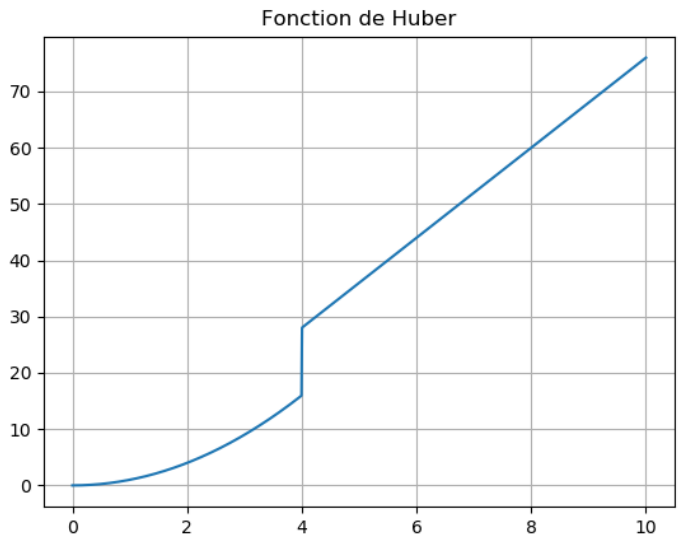
$$c(y, y') = \psi(y - y'),$$

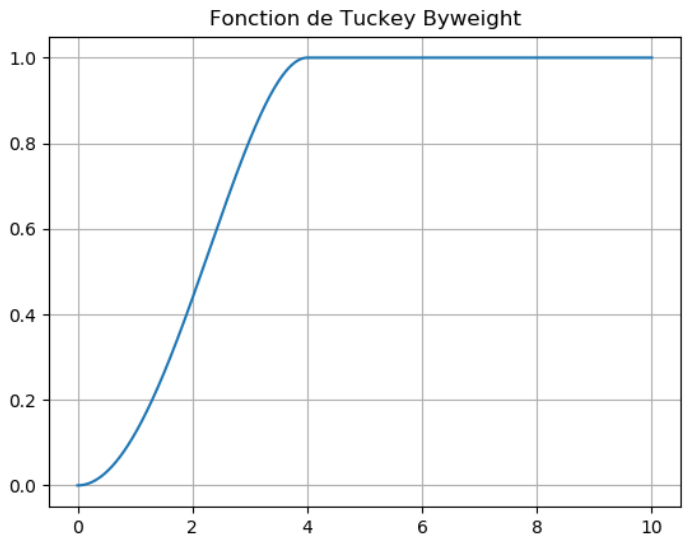
où $\psi : \mathbb{R} \to \mathbb{R}^+$ est paire.





Et encore d'autres...





Classification supervisée

Terminologie

Lorsque la variable d'intérêt Y ne prend qu'un nombre fini de valeurs, on parle de classification supervisée, de classifieur plutôt que de prédicteur, de classifieur de Bayes et de règle de classification.

Classification supervisée

Terminologie

Lorsque la variable d'intérêt Y ne prend qu'un nombre fini de valeurs, on parle de classification supervisée, de classifieur plutôt que de prédicteur, de classifieur de Bayes et de règle de classification.

Exemple

Le filtre à spam, reconnaissance de caractères,... Dans le cas où Y ne prend que deux valeurs on parle de classification binaire.

Classification binaire

On suppose désormais que $\mathcal{Y} = \{0, 1\}$.La fonction de régression est donnée par :

$$\eta(X) = \mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X)$$

Classification binaire

On suppose désormais que $\mathcal{Y} = \{0, 1\}$.La fonction de régression est donnée par :

$$\eta(X) = \mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X)$$

Exemple (Règle de classification par partition)

Soit $\mathcal{A}$ une partition au plus dénombrable mesurable de $\mathcal{X}$,

$$\widehat{f}_{\mathcal{A}}^{p-c}((x_i, y_i)_{1 \leqslant i \leqslant n}, x) := \begin{cases} 1 & \text{si } |\{i \in [\![1, n]\!], \quad y_i = 1, \ x_i \in \mathcal{A}(x)\}| > \\ & \quad |\{i \in [\![1, n]\!], \quad y_i = 0, \ x_i \in \mathcal{A}(x)\}| \\ 0 & \text{sinon.} \end{cases}$$

Ceci défini la règle d'apprentissage par partition associée à $\mathcal{A}$, notée $\widehat{f}_{\mathcal{A}}^{p-c}$.

Premier lien régression-classification

Remarque

Les règles d'apprentissage et de classification par partition sont liées. Étant donnée $\mathcal{A}$ une partition mesurable et au plus dénombrable de $\mathcal{X}$, pour tout échantillon D_n et tout $x \in \mathcal{X}$ on a l'égalité suivante :

$$\widehat{f}_{\mathcal{A}}^{p-c}(D_n; x) = \mathbb{1}_{\widehat{f}_{\mathcal{A}}^{p-r}(D_n;x) > \frac{1}{2}}$$

Premier lien régression-classification

Remarque

Les règles d'apprentissage et de classification par partition sont liées. Étant donnée $\mathcal{A}$ une partition mesurable et au plus dénombrable de $\mathcal{X}$, pour tout échantillon D_n et tout $x \in \mathcal{X}$ on a l'égalité suivante :

$$\widehat{f}_{\mathcal{A}}^{p-c}(D_n; x) = \mathbb{1}_{\widehat{f}_{\mathcal{A}}^{p-r}(D_n;x) > \frac{1}{2}}$$

Définition

On dit que $\widehat{f}_{\mathcal{A}}^{p-c}$ est la règle par *plug-in* associée à $\widehat{f}_{\mathcal{A}}^{p-r}$

Coût 0-1, définition

Définition

Une fonction naturelle en classification est le coût 0-1, défini par :

$$\forall y, y' \in \mathcal{Y}, \quad c(y, y') = \mathbb{1}_{y \neq y'}.$$

Remarque

Puisque $\mathcal{Y} = \{0, 1\}$, le coût 0-1 coïncide avec le coût quadratique. Le risque associé, appelé *risque 0-1* est défini par :

$$\forall f \in \mathcal{F}, \quad \mathcal{R}(f) = \mathbb{E}[\mathbb{1}_{f(X) \neq Y}] = \mathbb{P}(f(X) \neq Y).$$

Coût 0-1, proposition

Proposition

Toujours avec $\mathcal{Y} = \{0, 1\}$, et en considérant le coût 0-1,

1. Le classifieur défini par $f^*(x) := \mathbb{1}_{\eta(x)>1/2}$ pour tout $x \in \mathcal{X}$ est un classifieur de Bayes.
2. Un classificateur f^* est un classificateur de Bayes si et seulement si :

$$f(X) = \mathbb{1}_{\eta(X)>\frac{1}{2}} \quad \mathbb{P}_X \text{-presque sûrement},$$

sauf éventuellement en $\{\eta(X) = 1/2\}$.

3. Le rique de Bayes vaut :

$$\mathcal{R}^* := \mathbb{E}\left[min\{\eta(X), 1 - \eta(X)\}\right],$$

4. L'excès de risque de tout classificateur $f \in \mathcal{F}$ s'écrit :

$$\ell(f^*, f) = \mathbb{E}\left[|2\eta(X) - 1| \, \mathbb{1}_{f^*(X)\neq f(X)}\right]$$

Classification zéro-erreur

Remarque

Le risque de Bayes est nul si et seulement si $\eta(X) \in \{0, 1\}$ presque sûrement. On parle de cas *"zéro-erreur"*.

Dans le cas général, on a trois situations :

1. $\eta \approx 1/2$,
2. $|\eta(X)| \gg 1/2$,
3. $\eta(X)$ est à une distance intermédiaire de $1/2$.

Coût asymétique

On peut considérer d'autres mesures d'erreur en classification.

Coût asymétique

On peut considérer d'autres mesures d'erreur en classification.

Définition (Coût asymétrique)

Pour tout $\omega = (\omega_0, \omega_1) \in [0, +\infty[^2$ tel que $\omega_0 + \omega_1 > 0$, on appelle *coût asymétrique* la fonction définie par :

$$\begin{array}{rccc} c_\omega & : & \{0,,1\}^2 & \to & \mathbb{R}^+ \\ & & (y, y') & \mapsto & \omega_{y'} \mathbb{1}_{y \neq y'}. \end{array}$$

Le risque associé s'écrit, pour tout $f \in \mathcal{F}$:

$$\begin{aligned} \mathcal{R}^\omega(f) &= \mathbb{E}\left[\omega_Y \mathbb{1}_{f(X) \neq Y}\right] \\ &= \omega_1 \mathbb{P}(f(X) = 0 \text{ et } Y = 1) + \omega_0 \mathbb{P}(f(X) = 1 \text{ et } Y = 0). \end{aligned}$$

Coût asymétrique

Proposition

On suppose $\mathcal{Y} = 0, 1$ et on considère le coût asymétrique c_ω.

1. Le classifieur défini pour tout $x \in \mathcal{X}$ par $f^*_\omega(x) := \mathbb{1}_{\eta(x) > \omega_0/(\omega_0+\omega_1)}$ est un classifieur de Bayes,
2. Un classifieur f est de Bayes si et seulement si

$$f(X) = f^*_\omega(X) \qquad ps,$$

sauf éventuellement sur l'évènement $\{\eta(x) = \omega_0/(\omega_0 + \omega_1)\}$.

3. Le risque de Bayes vaut :

$$\mathcal{R}^{*,\omega} = \mathbb{E}\left[\min\left\{\omega_1 \eta(X), \omega_0(1 - \eta(X))\right\}\right].$$

4. L'excès de risque de tout classifieur $f \in \mathcal{F}$ s'écrit :

$$\ell^\omega(f^*_\omega, f) = (\omega_0 + \omega_1)\mathbb{E}\left[\left|\eta(X) - \frac{\omega_0}{\omega_0 + \omega_1}\right| \mathbb{1}_{f^*_\omega(X) \neq f(X)}\right].$$

Coût asymétrique VS les autres coûts

Remarque

Lorsque $\omega_0 = \omega_1 = 1$, le coût asymétrique c_ω est le coût 0-1 et les deux dernières propriétés énoncées coïncident.

Coût asymétrique VS les autres coûts

Remarque

Lorsque $\omega_0 = \omega_1 = 1$, le coût asymétrique c_ω est le coût 0-1 et les deux dernières propriétés énoncées coïncident.

Remarque

En classification binaire, avec $\mathcal{Y} = \{0, 1\}$, toute fonction de coût c non identiquement nulle sur $\mathcal{Y}$ s'écrit comme un coût asymétrique avec $\omega_0 = c(1, 0)$ et $\omega_1 = c(0, 1)$

Liens entre régression et classification

On se limite désormais au coût 0-1.
En deux temps :

Liens entre régression et classification

On se limite désormais au coût 0-1.
En deux temps :

- on estime la fonction de régression par $\widehat{\eta}$,

Liens entre régression et classification

On se limite désormais au coût 0-1.
En deux temps :

- on estime la fonction de régression par $\widehat{\eta}$,
- on utillise le classifieur $x \mapsto \mathbb{1}_{\widehat{\eta}(x)>1/2}$.

Liens entre régression et classification

On se limite désormais au coût 0-1.
En deux temps :

- on estime la fonction de régression par $\widehat{\eta}$,
- on utillise le classifieur $x \mapsto \mathbb{1}_{\widehat{\eta}(x)>1/2}$.

Définition

Soit $\widehat{\eta}$ une règle de régression. La règle de classification qui à tout échantillon $D_n \in (\mathcal{X} \times \{0,1\})^n$ et à tout $x \in \mathcal{X}$ associe

$$\widehat{f}_{\widehat{\eta}}(D_n; x) := \mathbb{1}_{\widehat{\eta}(D_n;x)>1/2}$$

est appelée *règle de classification par plug-in* associée à $\widehat{\eta}$.

Liens entre régression et classification

On se limite désormais au coût 0-1.
En deux temps :

- on estime la fonction de régression par $\widehat{\eta}$,
- on utillise le classifieur $x \mapsto \mathbb{1}_{\widehat{\eta}(x)>1/2}$.

Définition

Soit $\widehat{\eta}$ une règle de régression. La règle de classification qui à tout échantillon $D_n \in (\mathcal{X} \times \{0,1\})^n$ et à tout $x \in \mathcal{X}$ associe

$$\widehat{f}_{\widehat{\eta}}(D_n; x) := \mathbb{1}_{\widehat{\eta}(D_n;x)>1/2}$$

est appelée *règle de classification par plug-in* associée à $\widehat{\eta}$.

Exemple

Règles de classification par partition.

Liens entre régression et classification

On se limite désormais au coût 0-1.
En deux temps :

- on estime la fonction de régression par $\widehat{\eta}$,
- on utillise le classifieur $x \mapsto \mathbb{1}_{\widehat{\eta}(x)>1/2}$.

Définition

Soit $\widehat{\eta}$ une règle de régression. La règle de classification qui à tout échantillon $D_n \in (\mathcal{X} \times \{0,1\})^n$ et à tout $x \in \mathcal{X}$ associe

$$\widehat{f}_{\widehat{\eta}}(D_n; x) := \mathbb{1}_{\widehat{\eta}(D_n;x)>1/2}$$

est appelée *règle de classification par plug-in* associée à $\widehat{\eta}$.

ATTENTION : cette définition repose fortement sur la convention $\mathcal{Y} = \{0,1\}$!

Liens entre régression et classification

Proposition

Soit $\mathbb{P}$ une loi sur $\mathcal{X} \times \{0,1\}$, η la fonction de régression associée, $\widehat{\eta}$ une règle de régression et $\widehat{f}_{\widehat{\eta}}$ la règle de classification par plug-in associée. On a alors, pour le coût $0-1$ en classification :

$$\ell\left(f^*, \widehat{f}_{\widehat{\eta}}(D_n)\right) \leqslant 2\mathbb{E}\left[\left|\widehat{\eta}(D_n; X) - \eta(X)\right| \middle| D_n\right]$$

Liens entre régression et classification

Proposition

Soit $\mathbb{P}$ une loi sur $\mathcal{X} \times \{0,1\}$, η la fonction de régression associée, $\widehat{\eta}$ une règle de régression et $\widehat{f}_{\widehat{\eta}}$ la règle de classification par plug-in associée. On a alors, pour le coût $0-1$ en classification :

$$\begin{aligned}\ell\left(f^*, \widehat{f}_{\widehat{\eta}}(D_n)\right) &\leqslant 2\mathbb{E}\left[\left|\widehat{\eta}(D_n; X) - \eta(X)\right| \middle| D_n\right] \\ &\leqslant 2\sqrt{\mathbb{E}\left[\left|\left(\widehat{\eta}(D_n; X) - \eta(X)\right)^2\right| \middle| D_n\right]}.\end{aligned}$$

Liens entre régression et classification

Proposition

Soit $\mathbb{P}$ une loi sur $\mathcal{X} \times \{0,1\}$, η la fonction de régression associée, $\widehat{\eta}$ une règle de régression et $\widehat{f}_{\widehat{\eta}}$ la règle de classification par plug-in associée. On a alors, pour le coût $0-1$ en classification :

$$\ell\left(f^*, \widehat{f}_{\widehat{\eta}}(D_n)\right) \leqslant 2\mathbb{E}\left[\left|\widehat{\eta}(D_n; X) - \eta(X)\right| \middle| D_n\right]$$
$$\leqslant 2\sqrt{\mathbb{E}\left[\left|\left(\widehat{\eta}(D_n; X) - \eta(X)\right)^2\right| \middle| D_n\right]}.$$

En particulier, si $\widehat{\eta}$ est faiblement consistante pour $\mathbb{P}$ avec le coût quadratique, alors $\widehat{f}_{\widehat{\eta}}$ est faiblement consistante pour $\mathbb{P}$ en classification avec le coût $0-1$.