Data Mining

Formalisation du Data Mining supervisé

W. Toussile

[1]Département MSP
École Nationale Supérieure Polytechnique

15/11/2019

Section 1

Position du problème

Position du problème I

- Données : $\mathcal{D}_n = (x_i, y_i)_{i=1}^n$
- Objectifs :
  - Exploration : Décrire le lien $y_i \approx g(x_i)$ entre les entrées et sorties
  - Prédictif : Inférer sur le lien $y \approx g(x)$ entre entrée et sortie.
- Modélisation : En général, les données sont considérées comme des réalisaions de vecteurs aléatoires $(X_i, Y_i)_{i=1}^n$, iid de loi de probabilité inconnue $P_{(X,Y)}$ dans la pratique.
- Notations :
  - $X(\Omega)$ et $Y(\Omega)$ ensembles des valeurs des X_i et des Y_i resp.

Position du problème II

- On distingue :
  - Régression lorsque $Y(\Omega) \subseteq \mathbb{R}^d$
  - Classification supervisée lorsque $Y(\Omega)$ est fini non ordonné.
- Pb : Rechercher une application mesurable

$$g : X(\Omega) \to Y(\Omega)$$

  optimale au sens d'un critère qu'on se donne (risque)
- g est appelée prédicteur dans les deux cas, plus précisément
  - Régresseur dans les pb de régression
  - Classifieur dans les pb de classification supervisée

Section 2

Fonction de perte - Risque

Fonction de perte I

Definition

On appelle ainsi toute fonction $L : Y(\Omega) \times Y(\Omega) \to \mathbb{R}_+$ vérifiant:

$$\begin{cases} L(y', y) = 0 & \text{si } y' = y \\ L(y', y) > 0 & \text{sinon} \end{cases}$$

- $L(g(x), y)$ mesure la perte ponctuelle d'un prédicteur g pour la sortie y associée à une entrée x

Fonction de perte II

Exemples

- Régression : En général $Y(\Omega) \in \mathbb{R}$
  - Perte quadratique : $L(y', y) = (y' - y)^2$
  - Perte absolue : $L(y', y) = |y' - y|$
  - Perte $\mathbb{L}_p$ pour $p > 1$: $L(y', y) = |y' - y|^p$
- Clasification supervisée
  - Perte 0-1 : $L(y', y) = 1_{[y' \neq y]}$
  - Plus généralement $L(y', y) = c_{y',y} 1_{[y' \neq y]}$ avec $c_{y',y} > 0$

Rappels

- $\mathbb{E}[Z] = \sum_{z \in Z(\Omega)} zP(Z = z)$ si Z est discrète
- $\mathbb{E}[Z] = \int_{z \in Z(\Omega)} zp_Z(z)dz$ si Z est continue de densité de probabilité $p_Z(\cdot)$.

Exercise

1. *Soit* $Z \sim \mathcal{P}(\lambda)$ *de loi définie par* $P(Z = k) = \dfrac{e^{-\lambda}\lambda^k}{k!}$ *pour tout* $k \in \mathbb{N}$. *Montrer que* $\mathbb{E}[Z] = \mathbb{V}\text{ar}(Z) = \lambda$.
2. *Soit* Z *une v.a.r. de densité* $f_Z(z) = \alpha e^{-\lambda}$ *où* $\alpha \in \mathbb{R}$ *et* $\lambda \in \mathbb{R}_+^*$. *Déterminer* α *en fonction de* α, *puis calculer* $\mathbb{E}[Z]$.

Risque I

Definition

Pour une fonction de perte L, le risque d'un prédicteur est défini par :

$$\mathcal{R}_L(g) = \mathbb{E}_{(X,Y)}[L(g(X), Y)]$$

- Le risque mesure la perte moyenne d'un prédicteur
- Le prédicteur idéal est $g^* = \inf_{\mathcal{M}} \mathcal{R}_L(g)$, où $\mathcal{M}$ est un modèle de prédicteurs qu'on se donne
- Dans la pratique, g^* n'est pas calculable, mais sa forme peut guider la conception d'algorithme pour l'estimer à partir des données.

Risque II

Exemples

- Régression
  - Risque quadratique : $\mathcal{R}_L(g) = \mathbb{E}_{(X,Y)}[\|g(X) - Y\|^2]$
  - Risque absolu pour $Y(\Omega) \subset \mathbb{R}$: $\mathcal{R}_L(g) = \mathbb{E}_{(X,Y)}[|g(X) - Y|]$
- Classification supervisée
  - Risque 0-1 : $\mathcal{R}_L(g) = \mathbb{E}_{(X,Y)}[1_{[g(X)\neq Y]}] = P\left(g(X) \neq Y\right)$
  - Plus généralement :
    $\mathcal{R}_L(g) = \sum_{y',y\in Y(\Omega)} L(y,y')P\left(g(X) = y', Y = y\right)$

Rappels I

- Lorsque X ou Y sont continues, alors (X, Y) est continue de densité :
  - $f_{(X,Y)}(x, y) = f_{Y|X=x}(y) f_X(x)$ si les deux sont continues
  - $f_{(X,Y)}(x, y) = P(Y = y|X = x) f_X(x)$ si Y discrète.
- Lorsque X et Y sont discrètes, (X, Y) l'est aussi et

$$P(X = x, Y = y) = P(Y = y|X = x)P(X = x)$$

Rappels II

- X et Y continues

$$\mathbb{E}[\phi(X,Y)] = \int_{X_(\Omega)} \int_{Y(\Omega)} \phi(x,y) f_{(Y|X=x)}(y) dy f_X(x) dx$$

- X continue et Y discrète

$$\mathbb{E}[\phi(X,Y)] = \int_{X(\Omega)} \sum_{y \in Y(\Omega)} \phi(x,y) P(y=y|X=x) f_X(x) dx$$

- X et Y sont discrètes

$$\mathbb{E}[\phi(X,Y)] = \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} \phi(x,y) P(y=y|X=x) P(X=x)$$

Exercice d'application I

Section 3

Prédicteur idéal

Régresseur idéal I

Proposition

1. *Pour le risque quadratique, le régresseur idéal est défini par*

$$g^*(x) = \mathbb{E}[Y|X = x]$$

.

2. *Pour le risque absolu, il est défini par*

$$g^*(x) = Mediane(Y|X = x)$$

.

Régresseur idéal II

La preuve de la proposition précédente repose sur le fait que pour une v.a.r Z,

1. $\mu^* := \arg\min_\mu \mathbb{E}[\|Z-\mu\|^2] = \mathbb{E}[Z]$
2. $m^* := \arg\min_m \mathbb{E}[|Z-m|] = \mathit{Mediane}(Z)$

Exercice

1. Montrer ce qui précède
2. Appliquer ce qui précède pour montrer la proposition précédente

Classifieur idéal I

Proposition

Le classifieur idéal pour le risque 0-1 est donnée par

$$g^*(x) = \arg \max_{k \in Y(\Omega)} P(Y = k | X = x)$$

Preuve

Classifieur idéal II

Exercice d'application

Supposons que $f_{(X,Y)}(x,y) = \frac{1}{x} e^{-x-\frac{y}{x}} 1_{]0,+\infty[\times[0,+\infty[}(x,y)$.

1. Déterminer le régresseur idéal g^* pour le risque quadratique.
2. Calculer le risque de g^*.

Exercice d'application

On considère une classification binaire où $Y(\Omega) = \{0,1\}$, et le risque $\mathcal{R}_w(g) = \mathbb{E}[2w(Y)1_{[g(X)\neq Y]}]$, où $w(0), w(1) > 0$ avec $w(0) + w(1) = 1$.

1. Déterminer le classifieur de Bayes, puis son risque.
2. Quel est l'intérêt d'un tel risque?

Classification binaire I

- $|Y(\Omega)| = 2$, en général on pose $\mathcal{Y} = \{0, 1\}$ ou $\mathcal{Y} = \{-1, 1\}$
- Supposons $\mathcal{Y} = \{0, 1\}$ et posons $\eta(x) = P(Y = 1|X = x)$. Le classifieur de bayes est donné par :

$$g^*(x) = 1_{[\eta(x) \geq \frac{1}{2}]}, \tag{1}$$

- L'équation de la fronctière entre les deux classes étant $\eta(x) = \frac{1}{2}$

Exercise

Supposons que le vecteur aléatoire d'entrée X est continue. Montrer que l'équation de la fronctière entre les deux classes est

$$f_{(X|Y=1)}(x)P(Y = 1) = f_{(X|Y=0)}(x)(1 - P(Y = 1)).$$

Classification binaire II

Proposition

Nous supposons $\mathcal{Y} = \{0, 1\}$. Pour tout classifieur g, l'excès de risque est donné par

$$\mathcal{R}(g) - \mathcal{R}(g^*) = E_X\left[|2\eta(X) - 1| 1_{[g(X) \neq g^*(X)]}\right].$$

En conséquence, $\mathcal{R}^* := \mathcal{R}(g^*) \leq \mathcal{R}(g)$ pour tout classifieur g. De plus,

$$\mathcal{R}^* = \mathbb{E}\left[\min\left\{\eta(X), 1 - \eta(X)\right\}\right] \leq \frac{1}{2}.$$

Classification binaire III

Preuve

- Montrer que pour x,

$$\begin{aligned} P(Y \neq g(X)|X &= x) = 1 - \eta(x)(2\eta(x) - 1) \times 1_{[g(x)=0]} \\ &= \min\{\eta(x), 1 - \eta(x)\} \end{aligned}$$

- Montrer que

$$P(Y \neq g(X)|X = x) - P(Y \neq g^*(X)|X = x) = |\eta(x) - 1| \times 1_{[g(x) \neq g^*(x)}$$

- En déduire les résultats.

Section 4

Exemples d'algo pour estimer g^*

Algo des k-ppv en régression

- Données $D_n = (X_i, Y_i)_{i=1}^n$
- Choisir une dissimilarité et un nombre $k \in \mathbb{N}^*$ de voisins
- Pour x $X(\Omega)$

Algo

- Déterminer l'ensemble $V_k(x)$ des indices i des k plus proches voisins de x parmi les X_i
- Prédire $g^*(x)$ par

$$\widehat{g}_{D_n}(x) = \frac{1}{|V_k(x)|} \sum_{i \in V_k(x)} Y_i$$

Algo des k-ppv en classification

- Données $D_n = (X_i, Y_i)_{i=1}^n$
- Choisir une dissimilarité et un nombre $k \in \mathbb{N}^*$ de voisins
- Pour x $X(\Omega)$

Algo

- Déterminer l'ensemble $V_k(x)$ des indices i des k plus proches voisins de x parmi les X_i
- Prédire $g^*(x)$ par

$$\begin{aligned} \widehat{g}_{D_n}(x) &= \arg \max_{y \in Y(\Omega)} \frac{1}{|V_k(x)|} \sum_{i \in V_k(x)} 1_{[Y_i = y]} \\ &= \arg \max_{y \in Y(\Omega)} \sum_{i \in V_k(x)} 1_{[Y_i = y]} \end{aligned}$$

Section 5

Apprentissage dans la pratique

Apprentissage dans la pratique

- Rappelons que g^* n'est pas calculable sans hypothèses supplémentaires sur $P_{(X,Y)}$.
- Dans la pratique, on cherche un estimateur de g^* dans un ensemble de fonctions $\mathcal{M}$ donné que l'on peut appeler modèle, par $\mathcal{M}$ peut être l'ensemble des prédicteur des k-ppv, $k \in [K_{\text{max}}] := \{1, \ldots, K_{\text{max}}\}$.
- Posons $g_{\mathcal{M}} = \arg\min_{g \in \mathcal{M}} \mathcal{R}_L(g)$. On a

$$\mathcal{R}_L(g^*) \leq \mathcal{R}_L(g_{\mathcal{M}})$$

- Le modèle $\mathcal{M}$ doit être choisi de sorte que
  - $\mathcal{R}_L(g_{\mathcal{M}}) - \mathcal{R}_L(g^*)$ ne soit pas trop grand
  - on puisse approcher $g_{\mathcal{M}}$ à partir de données

Estimateur du minimum du risque empirique I

Definition

Il s'agit de $\widehat{\mathcal{R}}_{L,D_n}(g) = \frac{1}{n}\sum_{i=1}^{n} L(g(X_i), Y_i)$

- Perte quadratique en régression :

$$\widehat{\mathcal{R}}_{L,D_n}(g) = \frac{1}{n}\sum_{i=1}^{n} \|(g(X_i))_i - \mathbb{Y}\|^2$$

- Perte 0-1 en classification :

$$\widehat{\mathcal{R}}_{L,D_n}(g) = \frac{1}{n}\sum_{i} 1_{[g(X_i)\neq Y]}$$

Estimateur du minimum du risque empirique II

- Estimateur du minimum du risque empirique (EMRE) :

$$\widehat{g}_{L,D_n,\mathcal{M}} := \arg\min_{g\in\mathcal{M}} \widehat{\mathcal{R}}_{L,D_n}(g)$$

- Consistance : $\widehat{g}_{L,D_n,\mathcal{M}}$ est dit consistant lorsque

$$\mathcal{R}_L(\widehat{g}_{L,D_n,\mathcal{M}}) - \mathcal{R}_L(g_{\mathcal{M}}) \underset{n\to\infty}{\longrightarrow} 0$$

  On dit alors qu'on fait asymptotiquement au mieux.

- Enjeux
  - Choix du modèle $\mathcal{M}$
  - Choix d'une méthode d'estimation de g^* dans $\mathcal{M}$
  - Estimation "honnête" de $\mathcal{R}(\widehat{g})$

Estimateur du minimum du risque empirique III

Exemple : Régression polynômiale

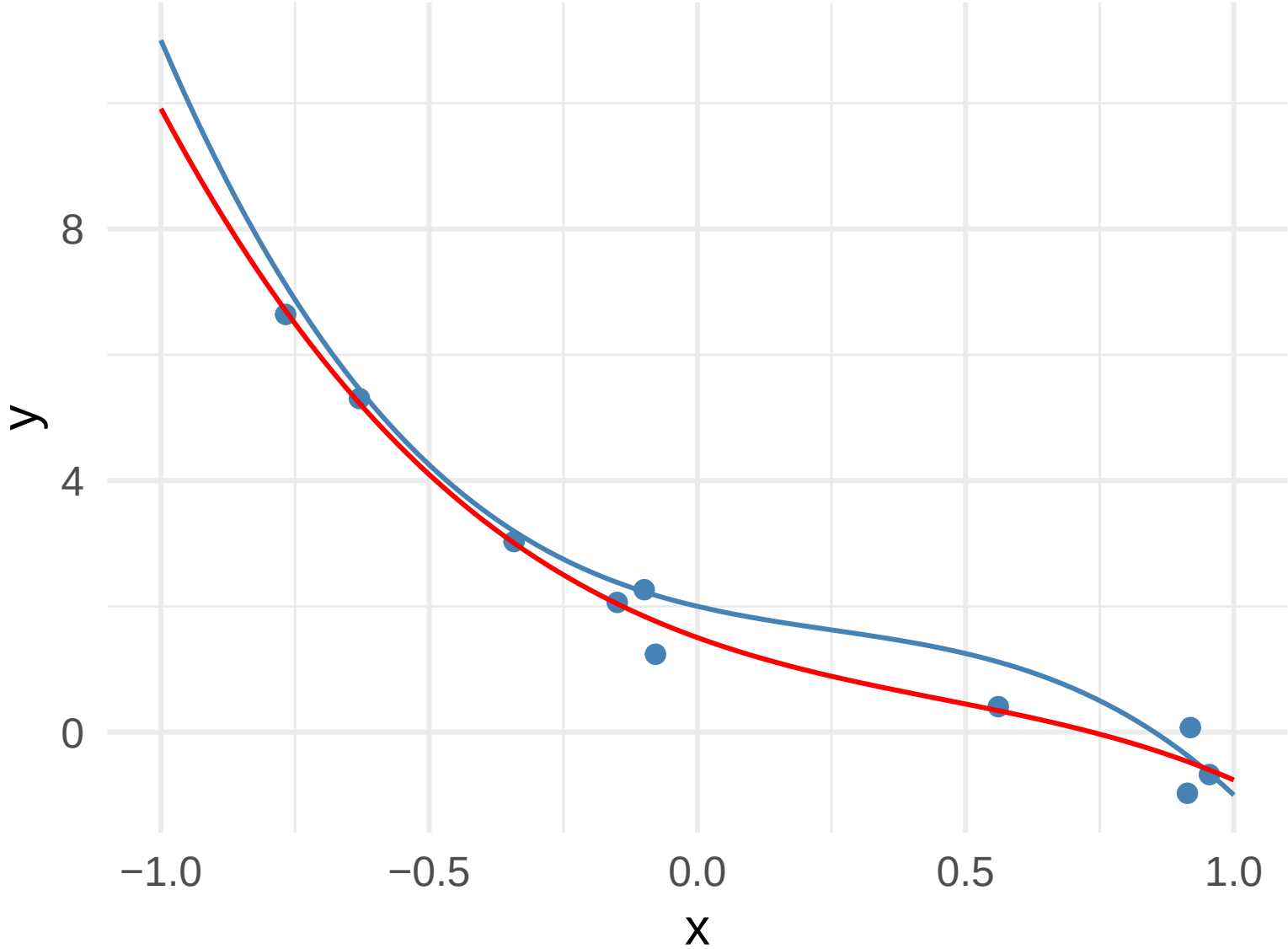
- Modèle

$$\begin{cases} Y_i & = & g(x_i) + \epsilon_i, \ i = 1, \cdots, n \\ \mathbb{E}[\epsilon_i] & = & 0 \\ \mathbb{C}\text{ov}(\epsilon_i, \epsilon_j) & = & \delta_{i,j}\sigma^2; \end{cases}$$

avec $x_i \in \mathbb{R}$.

- $\mathcal{M}_p = \left\{ g : X(\Omega) \to Y(\Omega) | \ \exists \ (\beta_j)_{j=0}^p, \ g(x) = \beta_0 + \sum_{j=1}^p \beta_j x^j \right\}$
-
  - Choix du modèle à travers de degré $p \in \mathbb{N}^*$
  - Méthode d'estimation pour le risque quadratique : les MC
  - Estimation du risque : par validation croisée

Estimateur du minimum du risque empirique IV



Section 6

Estimation du risque par validation croisée

Position du pb

- $\mathcal{R}\left(\widehat{g}_m\right)$ n'est pas calculable. Son estimateur naïf est le risque empirique

$$\widehat{\mathcal{R}}\left(\widehat{g}_m, \mathcal{A}_n\right) = \frac{1}{n} \sum_i L\left(\widehat{g}_m(X_i), Y_i\right). \tag{2}$$

- (2) est une estimation optimiste du risque de $\widehat{g}_n$. En effet, il est estimé à partir des données $\mathcal{A}_n$ qui servent de test.
- Exemple : Tout EMRE $\widehat{g}_m$ qui vérifie $\widehat{g}\left(X_i\right) = Y_i$ pour tout $i = 1, \ldots, n$ est telle $\widehat{\mathcal{R}}\left(\widehat{g}, \mathcal{A}_n\right) = 0$, mais est potentiellement mauvais sur de nouvelles données.

Estimation du risque par validation croisée

- On suppose n grand
- On partitionne $I = [n]$ en deux I_{train} et I_{test}
- On estime $\widehat{g}_{train} = \widehat{g}\left(\{(X_i, Y_i)\}_{i \in I_{train}}\right)$
- On estime $\mathcal{R}(\widehat{g})$ par $\widehat{\mathcal{R}}(\widehat{g}_{train})$

Leave one out

- Pour $j = 1, \cdots, n$, estimer $\widehat{g}^{(j)} = \widehat{g}(\{(X_i, Y_i)\}_{i \neq j})$
- Estimer $\mathcal{R}(\widehat{g})$ par $\widehat{\mathcal{R}}_n^{(loo)}(\widehat{g}) = \frac{1}{n} \sum_i L\left(\widehat{g}^{(i)}(X_i), Y_i\right)$

Par K folds

- Partitionner aléatoirement $I = [n]$ en K parties $I_k,\ k = 1, \cdots, K$
- Pour $k = 1, \cdots, K$, estimer $\widehat{g}^{(k)} = \widehat{g}(\{(X_i, Y_i)\}_{i \notin I_k})$
- Pour $k = 1, \cdots, K$, calculer
  $\widehat{\mathcal{R}}_k\left(\widehat{g}^{(k)}\right) = \frac{1}{|I_k|} \sum_{i \in I_k} L\left(\widehat{g}^{(k)}(X_i), Y_i\right)$
- Estimer $\mathcal{R}(\widehat{g})$ par $\frac{1}{K} \sum_k \widehat{\mathcal{R}}_k$

"Bootstrap"

- Pour $b = 1, \cdots, B$, tirer successivement avec remmise n entiers I_b de $[n]$
- Pour $b = 1, \cdots, B$, estimer $\widehat{g}^{(b)} = \widehat{g}(\{(X_i, Y_i)\}_{i \in I_b})$
- Pour $b = 1, \cdots, B$, calculer $\widehat{\mathcal{R}}_b\left(\widehat{g}^{(k)}\right) = \frac{1}{n - |I_b|} \sum_{i \notin I_b} L\left(\widehat{g}^{(b)}(X_i), Y_i\right)$
- Estimer $\mathcal{R}(\widehat{g})$ par $\frac{1}{B} \sum_k \widehat{\mathcal{R}}_b$