Data Science for Actuaries (ACT6100)

Arthur Charpentier

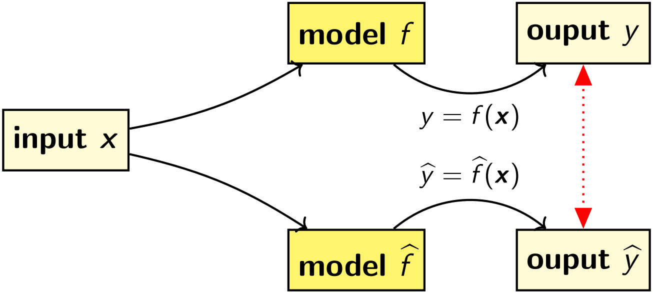
Supervisé # 1.1 (Concepts Fondamentaux)

automne 2Q20

https://github.com/freakonometrics/ACT6100/

Modèle

On suppose qu'il existe une fonction $f : \mathcal{X} \to \mathcal{Y}$, telle que $y = f(\boldsymbol{x})$.

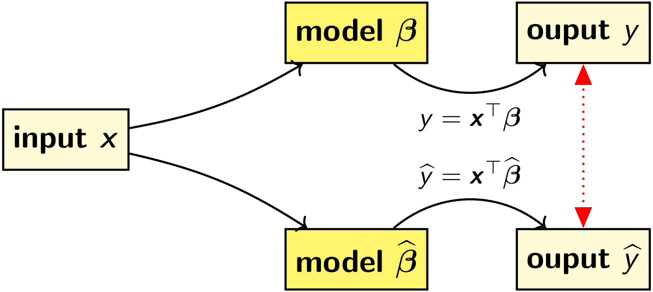


On dispose de données $\mathcal{D}_n = \{(y_i, \boldsymbol{x}_i)\}$ vues comme réalisation de n variables i.i.d. $(Y_i, \boldsymbol{X}_i)$.
$\mathcal{D}_n$ est une réalisation de $D_n = \{(Y_i, \boldsymbol{X}_i)\}$
On a associer un modèle $\widehat{f} = f(\cdot|\mathcal{D}_n) \in \mathcal{M}$ à partir des données $\mathcal{D}_n$

- $\widehat{y} = f(\boldsymbol{x}|\mathcal{D}_n)$ est la prévision associée à $\boldsymbol{x}$
- $\widehat{Y} = f(\boldsymbol{x}|D_n)$ est la prévision vue comme une variable aléatoire

Modèle

On suppose qu'il existe un paramètre $\boldsymbol{\beta}$, tel que $y = \boldsymbol{x}^\top\boldsymbol{\beta} + \epsilon$.

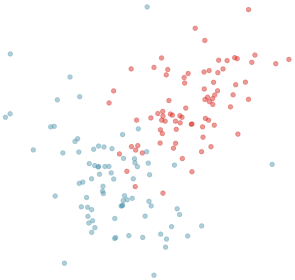


On dispose de données $\mathcal{D}_n = \{(y_i, \boldsymbol{x}_i)\}$ vues comme réalisation de n variables i.i.d. $(Y_i, \boldsymbol{X}_i)$.
$\mathcal{D}_n$ est une réalisation de $D_n = \{(Y_i, \boldsymbol{X}_i)\}$
On a associer un modèle (estimateur) $\widehat{\boldsymbol{\beta}}_n$ à partir des données $\mathcal{D}_n$

- $\widehat{y} = \boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}_n$ est la prévision associée à $\boldsymbol{x}$
- $\widehat{Y} = \boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_n$ est la prévision vue comme une variable aléatoire

Interpréter un modèle

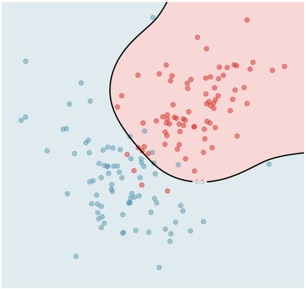
Considérons un problème de classification, $(x_{1,i}, x_{2,i}, y_i)$ avec $y_i \in \{0, 1\}$, et un classifieur $\widehat{f} : \mathcal{X} \to \{0, 1\}$ decision boundary $\mathcal{D}(\boldsymbol{x})$ — (région ambigüe)



- interprétable par nature

arbre de classification + modèle linéaire

- surgate local : au voisinage d'un point, on essaye de trouver un modele lineaire qui approche bien



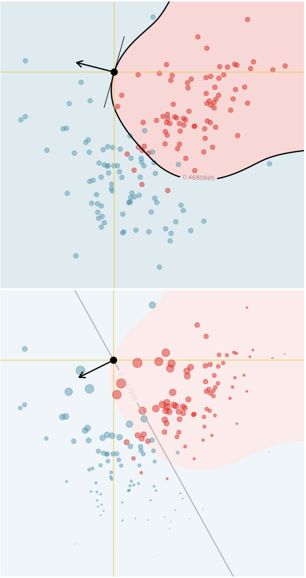
see also Local interpretable model-agnostic explanations (LIME)

Interpréter un modèle

- chercher le plan tangent à $\mathcal{D}(\boldsymbol{x})$, de vecteur normal $\vec{\boldsymbol{u}}_{(x_1,x_2)}$, interprété comme $\widehat{\boldsymbol{\beta}}_{(x_1,x_2)}$

- faire une régression linéaire locale et utiliser $\widehat{\boldsymbol{\beta}}_{(x_1,x_2)}$ pour interpréter localement, au voisinage de $\boldsymbol{x} = (x_1, x_2)$.

- ▶ surgate global : trouver un modèle linéaire proche / fidèle



Fonction de perte (coût) et risque

- On définit une fonction de perte (ou de coût) comme étant une fonction définie sur $\mathcal{Y} \times \mathcal{Y}$, à valeurs réelles, telle que $\ell(y, y') \geq 0$ et $\ell(y, y) = 0$.
- Ainsi, le risque d'un prédicteur $\widehat{f}$ pour f (inconnue) est

$$\mathcal{R}(\widehat{f}) = \mathbb{E}\Big[\ell\Big(Y, \widehat{f}(\mathbf{X})\Big)\Big],$$

- en régression, on travaille généralement avec une fonction de coût quadratique,

$$\ell(y, y') = (y - y')^2.$$

On obtient alors la fonction de risque

$$\mathcal{R}(\widehat{f}) = \mathbb{E}\Big[\Big(Y - \widehat{f}(\mathbf{X})\Big)^2\Big],$$

où $\widehat{f}(\mathbf{X})$ représente une prédiction.

- On obtient un risque quadratique

Risque empirique

- La distribution exacte de Y n'est pas connue, il n'est donc pas possible de calculer la fonction de risque. On construit la fonction de risque empirique

$$\widehat{\mathcal{R}}_n(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{f}(\mathbf{x}_i), y_i\right),$$

où $\{\mathbf{x}_i, y_i\}_{i=1,\dots,n}$ est un échantillon aléatoire

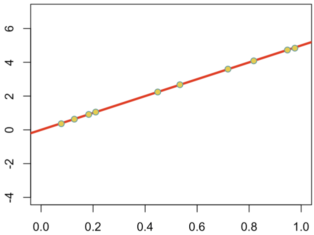
- Dans un contexte de régression, on travaille généralement avec l'erreur quadratique moyenne, ou *Mean Squared Error* (MSE), donnée par

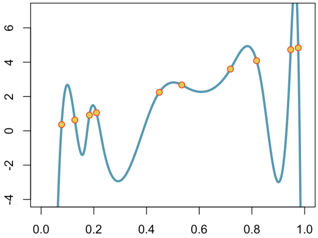
$$\text{EQM}_n = \text{MSE}_n = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \widehat{f}(\mathbf{x}_i)\right)^2.$$

Risque empirique

- sur les données suivantes $\{(x_1, y_1), \cdots, (x_n, y_n)\}$

$$\widehat{\mathcal{R}}_n(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\widehat{f}(\mathbf{x}_i), y_i\right) = 0 \quad \text{mais aussi } \widehat{f}$$





Notion de généralisation...

Base d'entrainement

- Dans la formule

$$\widehat{\mathcal{R}}IS_n(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}\ell\left(\widehat{f}_n(\mathbf{x}_i), y_i\right) = 0$$

  l'échantillon utilisé pour calculer l'erreur quadratique moyenne est le même que celui utilisé pour ajuster le modèle: erreur d'entrainement, in-sample risk

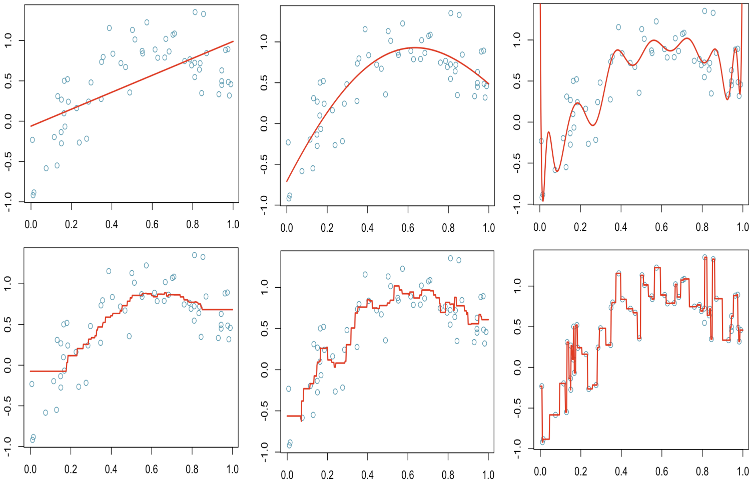
- si on construit

$$\widehat{f}_n = \underset{f\in\mathcal{M}}{\text{argmin}}\left\{\widehat{\mathcal{R}}_n^{IS}(f)\right\}$$

  on aura tendance à capturer beaucoup de bruit et à sur-ajuster les données : sur-apprentissage

Base d'entrainement

$\mathcal{M}_k$: régression polynomiale de degré k ou k-neareast neighbors



Base d'entrainement et base de validation

- Pour éviter ce problème de surapprentissage, on va diviser aléatoirement la base de données initiale en une base d'entrainement et une base de validation.
- La base d'entrainement, de taille $n_T < n$, sera utilisée pour estimer les paramètres du modèle,

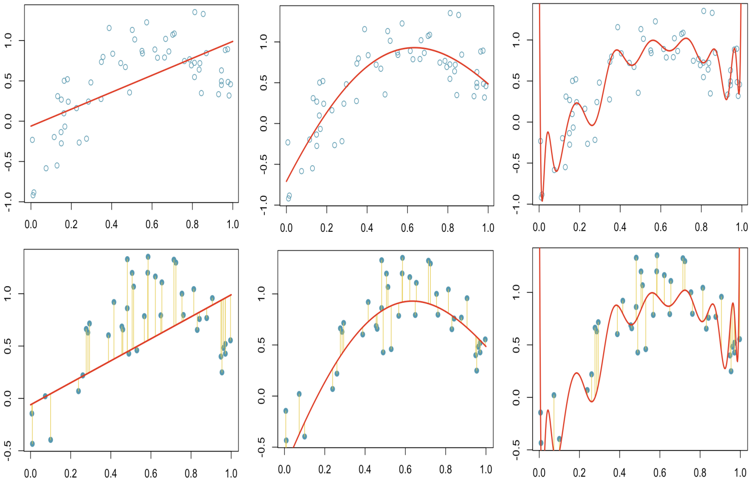
$$\widehat{f}_{n_T} = \underset{f \in \mathcal{M}}{\text{argmin}} \left\{ \widehat{\mathcal{R}}^{IS}_{n_T}(f) \right\}$$

- La base de validation, de taille $n_V = n - n_T$, sera utilisée pour sélectionner le modèle en minimisant une erreur quadratique moyenne de validation, out-of-sample risk

$$\widehat{\mathcal{R}}^{OS}_{n_V}(\widehat{f}_{n_T}) = \frac{1}{n} \sum_{i=1}^{n_V} \ell\left(\widehat{f}_{n_T}(\mathbf{x}_i), y_i\right) = 0$$

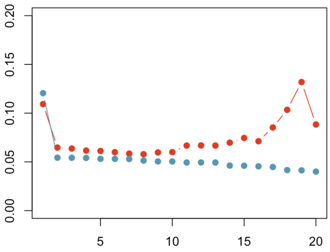
Base d'entrainement et de validation

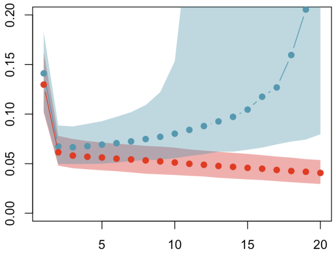
$\mathcal{M}_k$: régression polynomiale de degré k



Base d'entrainement et base de validation

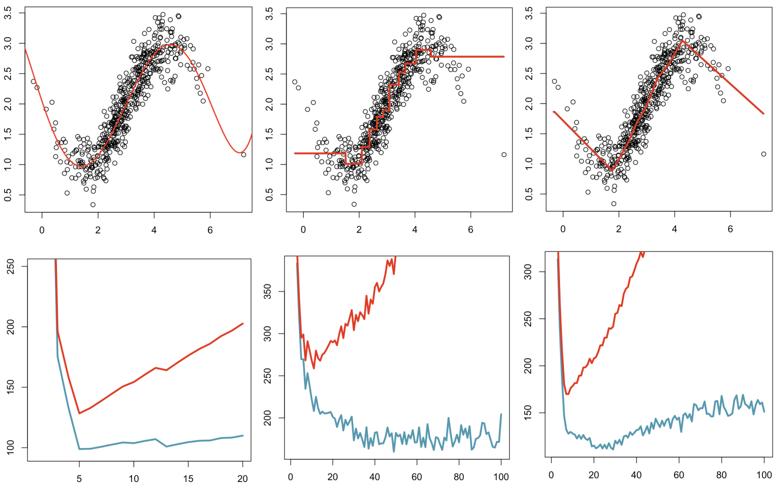
```
set.seed(1)
n=100
x=runif(n)
y=sin((x-1/6)*pi)+rnorm(n)/4
base=data.frame(x=x,y=y)
train=1:60
test=61:n
EQM = function(k){
reg=lm(y~poly(x,k),data=base[train,])
base$yp=predict(reg,newdata=base)
eqm_v=sum((base[test,"y"]-base[test,"yp"])^2)
eqm_t=sum((base[train,"y"]-base[train,"yp"])^2)
c(eqm_t/60,eqm_v/40)}
VE=Vectorize(EQM)(1:20)
```





Base d'entrainement et base de validation

Régression polynomiale, constante par morceaux, et linéaire par morceaux (splines)



Décomposition de l'erreur quadratique moyenne

On a

$$
\begin{aligned}
\mathbb{E}[\Big(Y-\widehat{f}(\mathbf{X})\Big)^2] &= \mathbb{E}[\Big(f(\mathbf{X})+\epsilon-\widehat{f}(\mathbf{X})\Big)^2] \\
&= \mathbb{E}[\Big(f(\mathbf{X})-\widehat{f}(\mathbf{X})\Big)^2]+\mathbb{E}[\epsilon^2] \\
&\quad +2\mathbb{E}[\epsilon\Big(f(\mathbf{X})-\widehat{f}(\mathbf{X})\Big)] \\
&= \mathbb{E}[\Big(f(\mathbf{X})-\widehat{f}(\mathbf{X})\Big)^2]+\mathbb{E}[\epsilon^2] \\
&= \mathbb{E}[\Big(f(\mathbf{X})-\widehat{f}(\mathbf{X})\Big)^2]+\mathrm{Var}[\epsilon] \\
&= \mathrm{Var}[\Big(f(\mathbf{X})-\widehat{f}(\mathbf{X})\Big)]+\mathbb{E}[\Big(f(\mathbf{X})-\widehat{f}(\mathbf{X})\Big)]^2+\mathrm{Var}[\epsilon] \\
&= \underbrace{\mathrm{Var}[\widehat{f}(\mathbf{X})]}_{(1)}+\underbrace{\mathbb{E}[\Big(f(\mathbf{X})-\widehat{f}(\mathbf{X})\Big)]^2}_{(2)}+\underbrace{\mathrm{Var}[\epsilon]}_{(3)}.
\end{aligned}
$$

Décomposition

- On obtient ainsi une erreur quadratique moyenne qui comprend trois termes: la variance de l'estimateur, le biais au carré et l'erreur stochastique (irréductible).
- Un *bon* modèle doit avoir simultanément un variance faible et un biais faible.
- Un modèle trop flexible n'est pas approprié car sa variance est généralement trop élevée: il est trop sensible au bruit présent dans les données.
- Un modèle peu flexible n'est pas approprié car son biais est généralement trop élevé: il n'arrive pas à capturer le comportement de la fonction f inconnue.
- compromis entre variance et biais

Un peu de formalisme

De manière générale, étant donnée une fonction de perte $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, le risque est

$$\mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{D_n}\left(\mathbb{E}_{Y|\boldsymbol{X}}\left[\ell(Y,\widehat{f}(\boldsymbol{X})|D_n)\right]\right)\right]$$

qu'on ne peut pas calculer sans connaître la loi de $(Y,\boldsymbol{X})$.
Si ℓ est la perte quadratique $\ell(y,\widehat{y}) = (y-\widehat{y})^2$,

$$\begin{aligned}
\mathcal{R}(\widehat{f}) &= \mathbb{E}_{D_n}\left(\mathbb{E}_{Y|\boldsymbol{X}}\left[\ell(Y,\widehat{f}(\boldsymbol{X})|D_n)\right]\right)\\
&= \left(\mathbb{E}_{Y|\boldsymbol{X}}(Y) - \mathbb{E}_{D_n}\left[\widehat{f}(\boldsymbol{X})|D_n)\right]\right)^2\\
&+ \mathbb{E}_{Y|\boldsymbol{X}}\left[\left(Y - \mathbb{E}_{Y|\boldsymbol{X}}(Y)\right)^2\right]\\
&+ \mathbb{E}_{D_n}\left[\left(\widehat{f}(\boldsymbol{X})|D_n) - \mathbb{E}_{D_n}\left[\widehat{f}(\boldsymbol{X})|D_n)\right]\right)^2\right]
\end{aligned}$$

On reconnaît le bias^2, l'erreur stochastique, et la variance de l'estimateur