Méthodes d'apprentissage

IFT603-712

Théorie de la décision

Par

Pierre-Marc Jodoin

Régression linéaire

RAPPEL

- Le modèle de **régression linéaire** est le suivant :

$$y_{\vec{w}}(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

$$\text{où } \vec{x} = (x_1, x_2, \ldots, x_d)^{\mathrm{T}}$$

- La prédiction correspond donc à
  - Une **droite** pour *d=1*
  - Un **plan** pour *d=2*
  - Un **hyperplan** pour *d>2*

2

Régression linéaire

RAPPEL

$$y_{\vec{w}}(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

$$y_{\vec{w}}(\vec{x}) = \vec{w}^{\mathrm{T}} \vec{x}$$

où $\vec{x}' = \left(1, x_1, x_2, \ldots, x_d\right)^{\mathrm{T}}$

3

Problème à résoudre

RAPPEL

$$\vec{w} = \arg\min_{\vec{w}} \sum_{n=1}^{N} \left(y_{\vec{w}}(\vec{x}_n) - t_n\right)^2$$

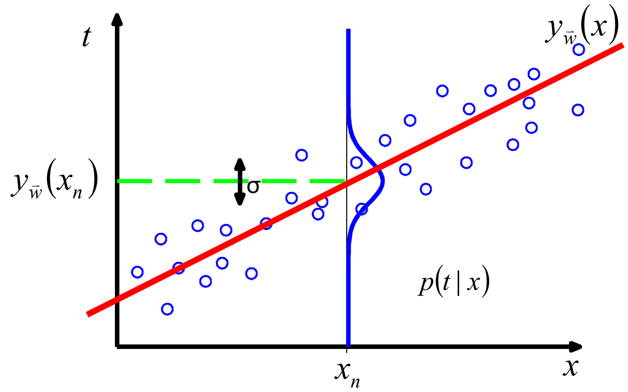
Il est très bien connu en technique d'apprentissage que cette solution est **optimale** lorsque le **bruit est gaussien.**

C'est même une question d'entrevue!

Formulation probabiliste

Loi conditionnelle (*maximum de vraisemblance*)

RAPPEL



Formulation probabiliste

RAPPEL

Pour entraîner le modèle $y_{\vec{w}}(\vec{x})$ nous passerons par une formulation probabiliste :

$$p(t \mid \vec{x}, \vec{w}, \Sigma) = N(t \mid y_{\vec{w}}(\vec{x}), \Sigma)$$

Hypothèse gaussienne

- Revient à supposer que les **cibles** sont des **versions bruitées** du vrai modèle

$$t_n = y_{\vec{w}}(\vec{x}_n) + \varepsilon$$

Bruit gaussien de moyenne 0
et de variance σ^2

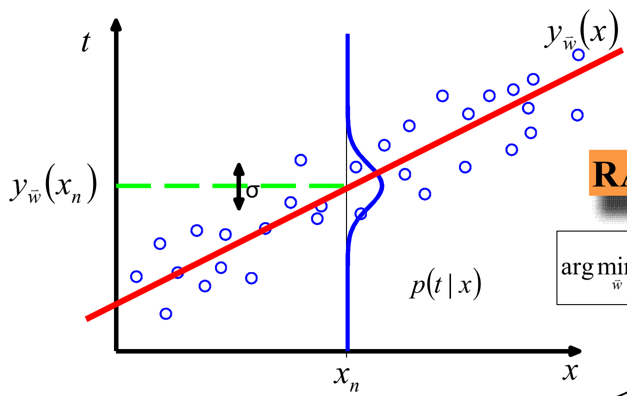
Maximum de vraisemblance

RAPPEL

$$\vec{w} = \arg\min_{\vec{w}} \sum_{n=1}^{N} \left(y_{\vec{w}}(\vec{x}_n) - t_n \right)^2$$

$$\vec{w}_{\mathrm{MV}} = \left(X^{\mathrm{T}} X \right)^{-1} X^{\mathrm{T}} T$$

Et bien sûr, on peut utiliser une **fonction de base** pour rendre le modèle non linéaire.



RAPPEL

$$\arg\min_{\vec{w}} \sum_{n=1}^{N} \left(y_{\vec{w}}(\vec{x}_n) - t_n \right)^2$$

Le meilleur modèle $y_{\vec{w}}(\vec{x}_n)$ selon correspond à la **moyenne** de la distribution **conditionnelle**

$$y_{\vec{w}}(\vec{x}_n) = E[t \mid \vec{x}] = \int t p(t \mid \vec{x}) dt$$

8

Comment prouver cette affirmation?

$$\arg\min_{y} \sum_{n=1}^{N} \left(y(x_n)-t_n\right)^2$$

$$\Big\Downarrow\ \text{Meilleure solution}$$

$$y(x)=E[t \mid x]=\int tp(t \mid x)dt$$

9

Preuve 1 (1.5.5 Bishop)

$$\arg\min_{y} \sum_{n=1}^{N} \underbrace{\left(y(\vec{x}_n)-t_n\right)^2}_{L(y(x_n),t_n)} \quad = \arg\min_{y} \underbrace{\frac{1}{N}\sum_{n=1}^{N} L\left(y(\vec{x}_n),t_n\right)}_{\substack{\text{Erreur moyenne:} \\ N\to\infty,\ E[L]}}$$

puisque $(\vec{x},t)$ est i.i.d de $p(\vec{x},t)$

$$\begin{aligned} E[L] &= \iint Lp(\vec{x},t)dxdt \\ &= \iint \left(y(\vec{x})-t\right)^2 p(\vec{x},t)dxdt \end{aligned}$$

10

Preuve 1

$$\underbrace{\arg\min_{y} \iint \left(y(\vec{x})-t\right)^2 p(\vec{x},t)\,dx\,dt}_{E[L]}$$

$$\frac{\delta E[L]}{\delta y}=0 \quad \Rightarrow 2\int \left(y(\vec{x})-t\right)p(\vec{x},t)\,dt=0$$

Calcul variationnel
Euler-Lagrange

11

Preuve 1

$$\cancel{2}\int \left(y(\vec{x})-t\right)p(\vec{x},t)\,dt=0$$

$$\int \left(y(\vec{x})p(\vec{x},t)-tp(\vec{x},t)\right)dt=0$$

$$\int y(\vec{x})p(\vec{x},t)\,dt-\int tp(\vec{x},t)\,dt=0$$

$$y(\vec{x})\underbrace{\int p(\vec{x},t)\,dt}_{\text{Marginalisation de t}}-\int tp(\vec{x},t)\,dt=0$$

$$y(\vec{x})p(\vec{x})-\int tp(\vec{x},t)\,dt=0$$

$$y(\vec{x})p(\vec{x})-\int tp(t\,|\,\vec{x})p(\vec{x})\,dt=0 \qquad \text{car } p(x,t)=p(t\,|\,x)p(x)$$

Preuve 1

$$y(\vec{x})p(\vec{x}) - \int tp(t \mid \vec{x})p(\vec{x})dt = 0$$

$$y(\vec{x})p\cancel{(\vec{x})} - \cancel{p(\vec{x})}\underbrace{\int tp(t \mid \vec{x})dt}_{\text{Expérance mathématique conditionnelle}} = 0$$

$$y(\vec{x}) - E[t \mid \vec{x}] = 0$$

$$\boxed{y(\vec{x}) = E[t \mid \vec{x}]}$$

Preuve 2 (1.5.5 Bishop)

$$L = (y(\vec{x}) - t)^2 = (y(\vec{x}) - E[t \mid x] + E[t \mid x] - t)^2$$

$$\begin{aligned} &= (y(\vec{x}) - E[t \mid \vec{x}])^2 \\ &\quad + 2(y(\vec{x}) - E[t \mid \vec{x}])(E[t \mid \vec{x}] - t) \\ &\quad + (E[t \mid \vec{x}] - t)^2 \end{aligned}$$

14

Preuve 2

$$E[L]=\iint (y(\vec{x})-t)^2\, p(x,t)\,dxdt$$

$$\begin{aligned}&=\iint (y(\vec{x})-E[t\mid\vec{x}])^2\, p(\vec{x},t)\,dxdt\\&+\iint 2(y(\vec{x})-E[t\mid\vec{x}])(E[t\mid\vec{x}]-t)\,p(\vec{x},t)\,dxdt\\&+\iint (E[t\mid\vec{x}]-t)^2\, p(\vec{x},t)\,dxdt\end{aligned}$$

$$\iint 2(y(\vec{x})-E[t\mid\vec{x}])(E[t\mid\vec{x}]-t)\,p(\vec{x},t)\,dxdt$$

15

Preuve 2

$$E[L]=\iint (y(\vec{x})-t)^2\, p(x,t)\,dxdt$$

$$\begin{aligned}&=\iint (y(\vec{x})-E[t\mid\vec{x}])^2\, p(\vec{x},t)\,dxdt\\&+\iint 2(y(\vec{x})-E[t\mid\vec{x}])(E[t\mid\vec{x}]-t)\,p(\vec{x},t)\,dxdt\\&+\iint (E[t\mid\vec{x}]-t)^2\, p(\vec{x},t)\,dxdt\end{aligned}$$

$$\iint 2(y(\vec{x})-E[t\mid\vec{x}])(E[t\mid\vec{x}]-t)\,p(t\mid\vec{x})\,p(\vec{x})\,dtdx$$

16

Preuve 2

$$E[L] = \iint \left(y(\vec{x}) - t\right)^2 p(x,t)\,dx\,dt$$

$$= \iint \left(y(\vec{x}) - E[t \mid \vec{x}]\right)^2 p(\vec{x},t)\,dx\,dt$$
$$+ \iint 2\left(y(\vec{x}) - E[t \mid \vec{x}]\right)\left(E[t \mid \vec{x}] - t\right) p(\vec{x},t)\,dx\,dt$$
$$+ \iint \left(E[t \mid \vec{x}] - t\right)^2 p(\vec{x},t)\,dx\,dt$$

$$\int 2\left(y(\vec{x}) - E[t \mid \vec{x}]\right) p(\vec{x}) \left\{ \int \left(E[t \mid \vec{x}] - t\right) p(t \mid \vec{x})\,dt \right\} dx$$

17

Preuve 2

$$E[L] = \iint \left(y(\vec{x}) - t\right)^2 p(\vec{x},t)\,dx\,dt$$

$$= \iint \left(y(\vec{x}) - E[t \mid \vec{x}]\right)^2 p(\vec{x},t)\,dx\,dt$$
$$+ \iint 2\left(y(\vec{x}) - E[t \mid \vec{x}]\right)\left(E[t \mid \vec{x}] - t\right) p(\vec{x},t)\,dx\,dt$$
$$+ \iint \left(E[t \mid \vec{x}] - t\right)^2 p(\vec{x},t)\,dx\,dt$$

$$\int 2\left(y(\vec{x}) - E[t \mid \vec{x}]\right) p(\vec{x}) \left\{ \int E[t \mid \vec{x}]\, p(t \mid \vec{x})\,dt - \int t\,p(t \mid \vec{x})\,dt \right\} dx$$

18

Preuve 2

$$E[L]=\iint \left(y(\vec{x})-t\right)^2 p(x,t)\,dx\,dt$$

$$\begin{aligned}
&=\iint \left(y(\vec{x})-E[t\mid\vec{x}]\right)^2 p(\vec{x},t)\,dx\,dt \\
&+\iint 2\left(y(\vec{x})-E[t\mid\vec{x}]\right)\left(E[t\mid\vec{x}]-t\right)p(\vec{x},t)\,dx\,dt \\
&+\iint \left(E[t\mid\vec{x}]-t\right)^2 p(\vec{x},t)\,dx\,dt
\end{aligned}$$

$$\int 2\left(y(\vec{x})-E[t\mid\vec{x}]\right)p(\vec{x})\left\{E[t\mid\vec{x}]\overbrace{\int p(t\mid\vec{x})\,dt}^{1}-E[t\mid\vec{x}]\right\}dx$$

Preuve 2

$$E[L]=\iint \left(y(\vec{x})-t\right)^2 p(x,t)\,dx\,dt$$

$$\begin{aligned}
&=\iint \left(y(\vec{x})-E[t\mid\vec{x}]\right)^2 p(\vec{x},t)\,dx\,dt \\
&+\iint 2\left(y(\vec{x})-E[t\mid\vec{x}]\right)\left(E[t\mid\vec{x}]-t\right)p(\vec{x},t)\,dx\,dt \\
&+\iint \left(E[t\mid\vec{x}]-t\right)^2 p(\vec{x},t)\,dx\,dt
\end{aligned}$$

$$\int 2\left(y(\vec{x})-E[t\mid\vec{x}]\right)p(\vec{x})\left\{\overbrace{E[t\mid\vec{x}]-E[t\mid\vec{x}]}^{0}\right\}dx$$

Preuve 2

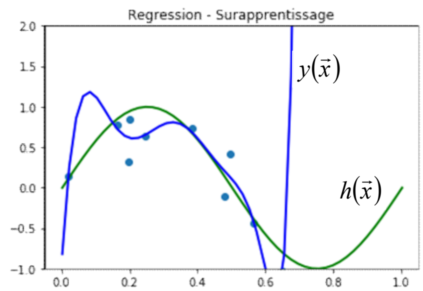
$$E[L] = \iint \left(y(\vec{x}) - t\right)^2 p(\vec{x}, t)\,dx\,dt$$

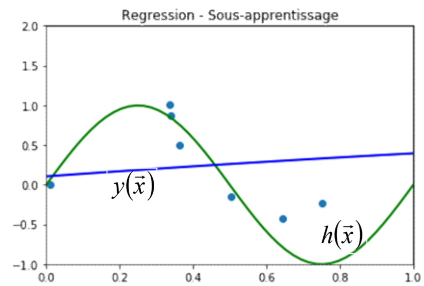
$$= \iint \left(y(\vec{x}) - E[t \mid \vec{x}]\right)^2 p(\vec{x}, t)\,dx\,dt + \iint \left(E[t \mid \vec{x}] - t\right)^2 p(\vec{x}, t)\,dx\,dt$$

$E[L]$ est minimum lorsque $\underline{y(\vec{x}) = E[t \mid \vec{x}]}$

Décomposition biais-variance
(Bishop 3.2)

On a vu les concepts de « sur-apprentissage » et de « sous-apprentissage ».

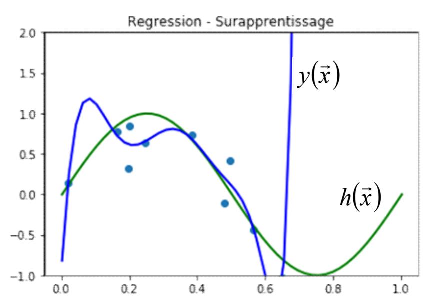


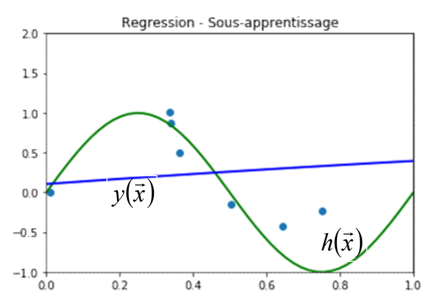


$y(\vec{x})$: modèle trouvé par apprentissage
$h(\vec{x})$: le meilleur modèle représentant les données

Décomposition biais-variance
(Bishop 3.2)

On a vu les concepts de « sur-apprentissage » et de « sous-apprentissage ».



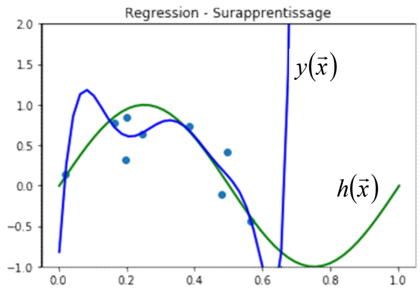


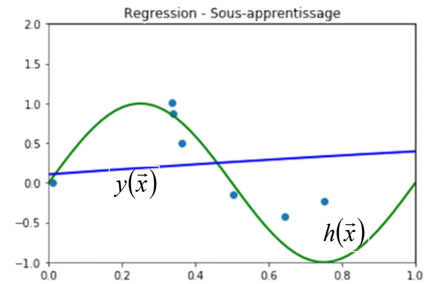
Le but d'un bon modèle est de trouver le **bon compromis biais-variance**

Décomposition biais-variance

(Bishop 3.2)

On a vu les concepts de « sur-apprentissage » et de « sous-apprentissage ».



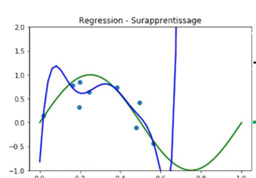


Variance élevée
Biais faible

Variance faible
Biais élevée

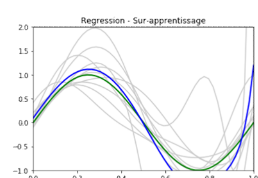
Variance élevée, Biais faible

D1



$y(\vec{x}; D_1)$

$h(\vec{x})$

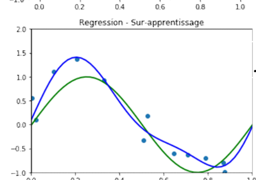


$y(\vec{x}; D_1), \ldots, y(\vec{x}; D_5)$

$1/5 \sum_{i=1}^{5} y(\vec{x}; D_i)$

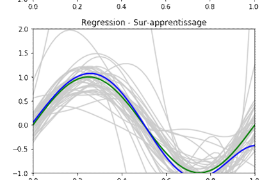
$h(\vec{x})$

D2



$y(\vec{x}; D_2)$

$h(\vec{x})$

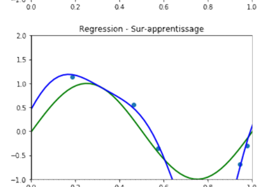


$y(\vec{x}; D_1), \ldots, y(\vec{x}; D_{40})$

$1/40 \sum_{i=1}^{40} y(\vec{x}; D_i)$

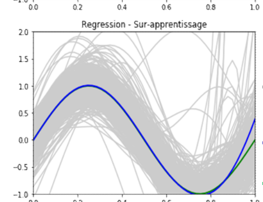
$h(\vec{x})$

D3



$y(\vec{x}; D_3)$

$h(\vec{x})$

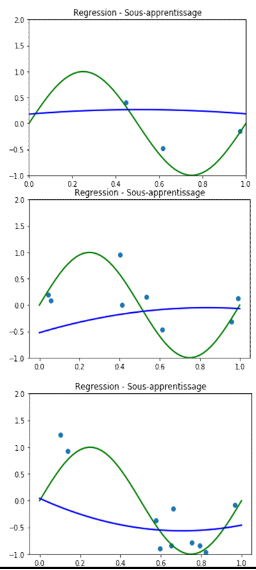


$y(\vec{x}; D_1), \ldots, y(\vec{x}; D_{500})$

$1/500 \sum_{i=1}^{500} y(\vec{x}; D_i)$

$h(\vec{x})$

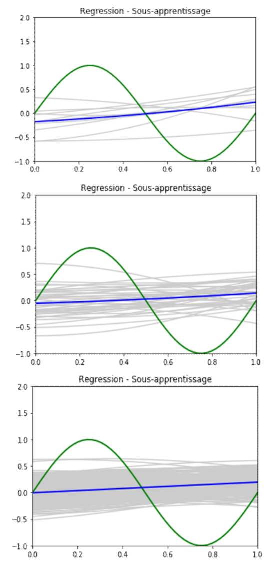
Variance faible, Biais élevé



D1: $y(\vec{x}; D_1)$ — $h(\vec{x})$

D2: $y(\vec{x}; D_2)$ — $h(\vec{x})$

D3: $y(\vec{x}; D_3)$ — $h(\vec{x})$



$y(\vec{x}; D_1), \ldots, y(\vec{x}; D_5)$

$1/5 \sum_{i=1}^{5} y(\vec{x}; D_i)$

$h(\vec{x})$

$y(\vec{x}; D_1), \ldots, y(\vec{x}; D_{40})$

$1/40 \sum_{i=1}^{40} y(\vec{x}; D_i)$

$h(\vec{x})$

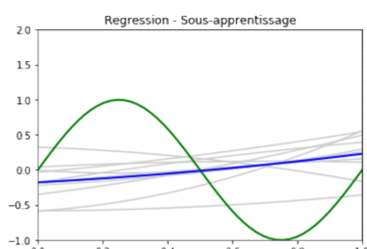
$y(\vec{x}; D_1), \ldots, y(\vec{x}; D_{500})$

$1/500 \sum_{i=1}^{500} y(\vec{x}; D_i)$

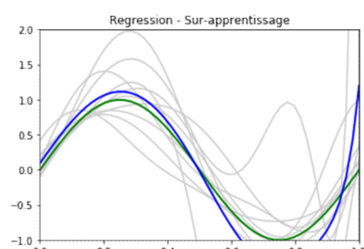
$h(\vec{x})$

Variance

Peut être vu intuitivement comme la **écart moyen entre les « courbes grises »** $y(\vec{x}; D_i)$



Variance faible



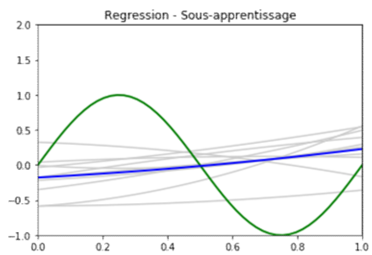
Variance élevée

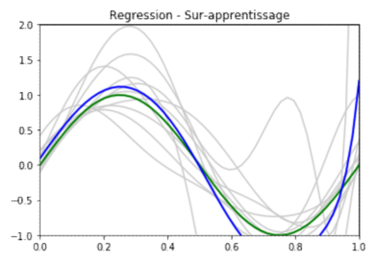
Dans le cas à 5 courbes:

$$\text{variance} = \frac{1}{5} \sum_{1=1}^{5} \Big(y(\vec{x}; D_i) - \underbrace{1/5 \sum_{i=1}^{5} y(\vec{x}; D_i)}_{\textbf{Courbe moyenne (bleue)}} \Big)^2$$

Biais

Peut être vu intuitivement comme la **différence entre les « courbes verte et bleue »**





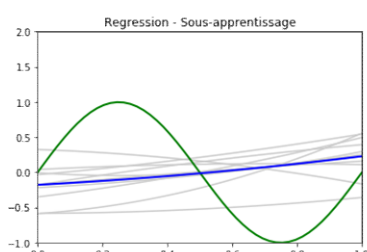
Biais élevé **Variance faible**

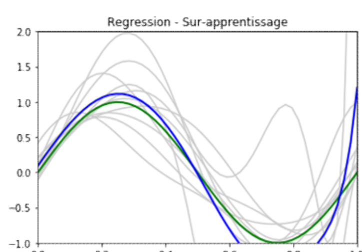
Dans le cas à 5 courbes:

$$\text{biais} = \frac{1}{5}\sum_{l=1}^{5}\Big(\underbrace{h(\vec{x})}_{\text{Courbe optimale (verte)}} - \underbrace{1/5\sum_{i=1}^{5} y(\vec{x};D_i)}_{\text{Courbe moyenne (bleue)}}\Big)^2$$

29

Compromis biais-variance





Biais élevé
Variance faible

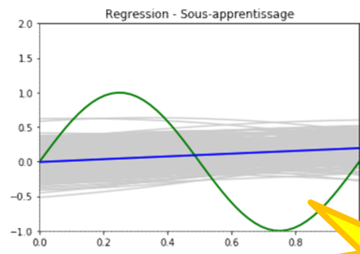
Biais faible
Variance élevée

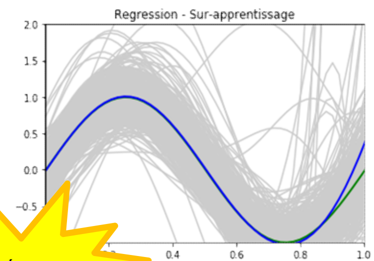
$$\text{variance} = \frac{1}{5}\sum_{l=1}^{5}\Big(y(\vec{x};D_i) - 1/5\sum_{i=1}^{5} y(\vec{x};D_i)\Big)^2$$

$$\text{biais} = \frac{1}{5}\sum_{l=1}^{5}\Big(h(\vec{x}) - 1/5\sum_{i=1}^{5} y(\vec{x};D_i)\Big)^2$$

30

Analyse quand $N \rightarrow \infty$





Espérance mathématique

Biais élevé
Variance faible

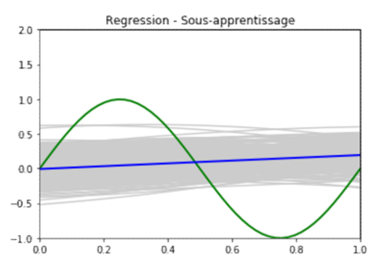
Biais faible
Variance élevée

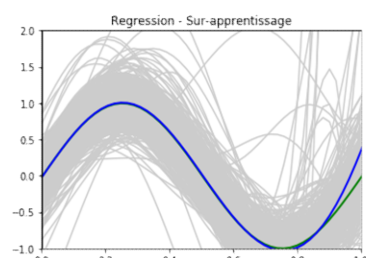
$$\text{variance} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \left(y(\vec{x}; D_i) - 1/N \sum_{i=1}^{N} y(\vec{x}; D_i) \right)^2$$

$$\text{biais} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \left(h(\vec{x}) - 1/N \sum_{i=1}^{N} y(\vec{x}; D_i) \right)^2$$

31

Analyse quand $N \rightarrow \infty$





Biais élevé
Variance faible

Biais faible
Variance élevée

$$\text{variance} = E_D\left[\left\{ y(\vec{x}; D) - E_D\left[y(\vec{x}; D) \right] \right\}^2 \right]$$

$$\text{biais} = E_D\left[\left\{ h(\vec{x}) - E_D\left[y(\vec{x}; D) \right] \right\}^2 \right]$$

32

Analyse formelle

On a démontré précédemment que

$$E[L] = \iint \{y(\vec{x}) - E[t \mid \vec{x}]\}^2 p(x,t)\,dxdt + \iint \{E[t \mid \vec{x}] - t\}^2 p(x,t)\,dxdt$$

où

$y(\vec{x})$: modèle entraîné à l'aide de D

$E[t \mid \vec{x}]$: modèle théorique optimal

(courbe verte)

Analyse formelle

Récriture

$$E[L] = \underbrace{\iint \{y(\vec{x};D) - h(\vec{x})\}^2 p(x,t)\,dxdt}_{\text{Mesure la performance du modèle y}} + \underbrace{\iint \{h(\vec{x}) - t\}^2 p(x,t)\,dxdt}_{\text{Mesure la magnitude du bruit dans les données}}$$

Analyse formelle

$$\{y(\vec{x};D)-h(\vec{x})\}^2 = \{y(\vec{x};D)-E_D[y(\vec{x};D)]+E_D[y(\vec{x};D)]-h(\vec{x})\}^2$$

$$\begin{aligned}&= \{y(\vec{x};D)-E_D[y(\vec{x};D)]\}^2\\ &\quad+2\{y(\vec{x};D)-E_D[y(\vec{x};D)]\}\{E_D[y(\vec{x};D)]-h(\vec{x})\}\\ &\quad+\{E_D[y(\vec{x};D)]-h(\vec{x})\}^2\end{aligned}$$

On peut démontrer que

$$\begin{aligned}E\left[\{y(\vec{x};D)-h(\vec{x})\}^2\right] &= E_D\left[\{y(\vec{x};D)-E_D[y(\vec{x};D)]\}^2\right]\\ &\quad+E_D\left[\{h(\vec{x})-E_D[y(\vec{x};D)]\}^2\right]\end{aligned}$$

Variance

Biais

36