Biais et Variance

August 21, 2019

Mots clefs. Après avoir lu le cours, ces mots doivent vous être familiés.

- Sur-apprentissage.
- Biais / Variance
- Flexibilité / interprétabilité

1 Biais-Variance

1.1 Un exemple de modèle linéaire avec transformation des input

Considérons un input $X \in [0, 1]$ quantitatif et un output $Y \in \mathbb{R}$ quantitatif. Considérons ν (=fréquence maximale) un paramètre. On définit le modèle suivant :

$$\begin{aligned} Y &= w_0 + \sum_{n=1}^{\nu} w_{2n} \cos(2\pi n X) + w_{2n-1} \sin(2\pi n X) \\ &= f_w(X) \end{aligned}$$

Plus ν est grand, et plus on a de chance de trouver un $\widehat{w}$ telle que

$$f_{\hat{w}} \simeq f^?$$

Mais, est-il toujours souhaitable de prendre un ν très grand ?

1.2 Flexibilité d'un modèle

Nous définissons la flexibilité d'un modèle par des exemples.

- un modèle linéaire pure et simple est très peu flexible. Il ne s'adapte qu'à des données proches d'un hyper-plan.
- Le modèle SinCos avec $\nu = 30$ est très flexible.
- quand $\widehat{f}$ est obtenu en minimisant

  $$\widehat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{Train} \operatorname{dist}(f(X_i) - Y_i) + \lambda \, Penalisation(f)$$

  Plus $\mathcal{F}$ est grand et plus le modèle est flexible.

  Plus λ est grand, plus on oblige le modèle a avoir des fonctions régulières, et moins le modèle est flexible.

- la technique des k-plus proches voisins

$$\widehat{f}(x) = \frac{1}{k} \sum_{X_i \in V_k(x)} Y_i$$

est très flexible quand $k = 1$, et peu flexible quand k est grand. Cas extrême : si k est égal au nombre de données dans $Train$, $\widehat{f}$ sera une fonction constante, partout égale à la moyenne des Y_i : aucune flexibilité.

- Quand $\widehat{f}$ est obtenue avec des noyaux gaussiens :

$$\widehat{f}(x) = \frac{\sum_{Train} N(x, X_i) Y_i}{\sum_{Train} N(x, X_i)}, \qquad N_\sigma(x, y) = e^{-\frac{1}{2}\left(\frac{x-y}{\sigma}\right)^2}$$

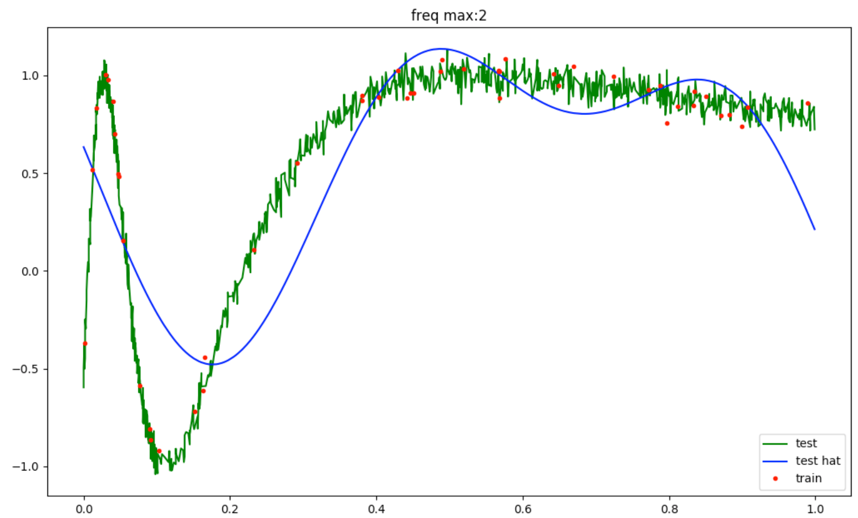
Plus σ est grand, plus la fonction $y \to N_\sigma(x, y)$ s'étale autour de x, et moins le modèle est flexible.

- De manière générale, plus un modèle s'adapte localement aux données, et plus il est flexible.

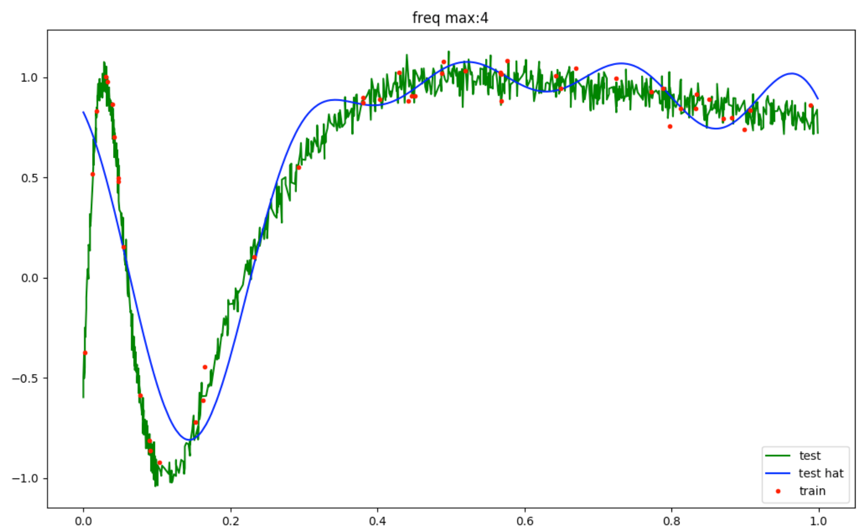
- Plus un modèle est flexible, plus il est complexe et moins il est interprétable.

1.3 Sur-apprentissage

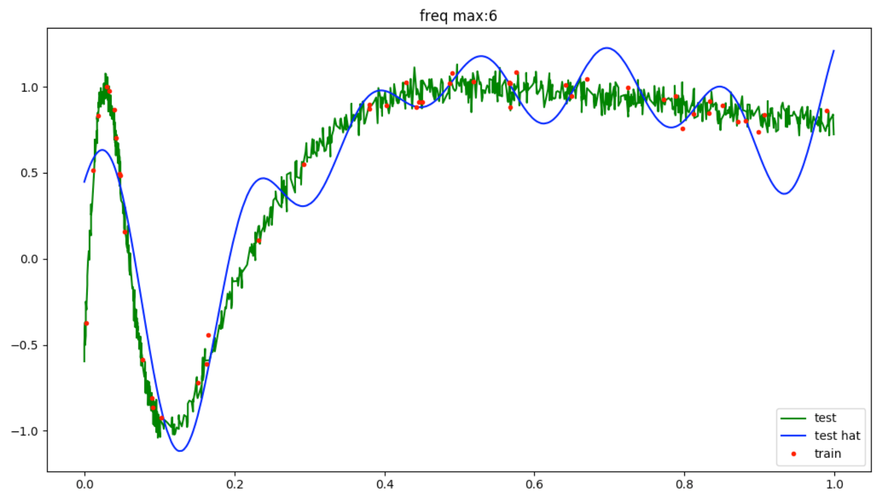
Observons notre modèle sin-cos pour différents ν. On dispose de données $Train$ avec lesquelles on entraine le modèle (= on calcule $\hat{w}$) et on dispose de données $Test$ très nombreuse qui nous permette de juger ce modèle.

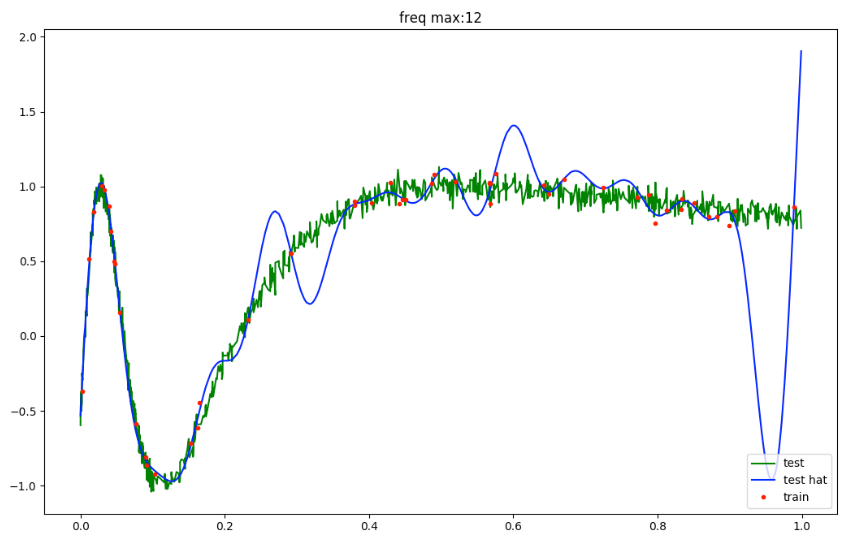


freq max:2
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
test
test hat
train



freq max:4
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0
test
test hat
train





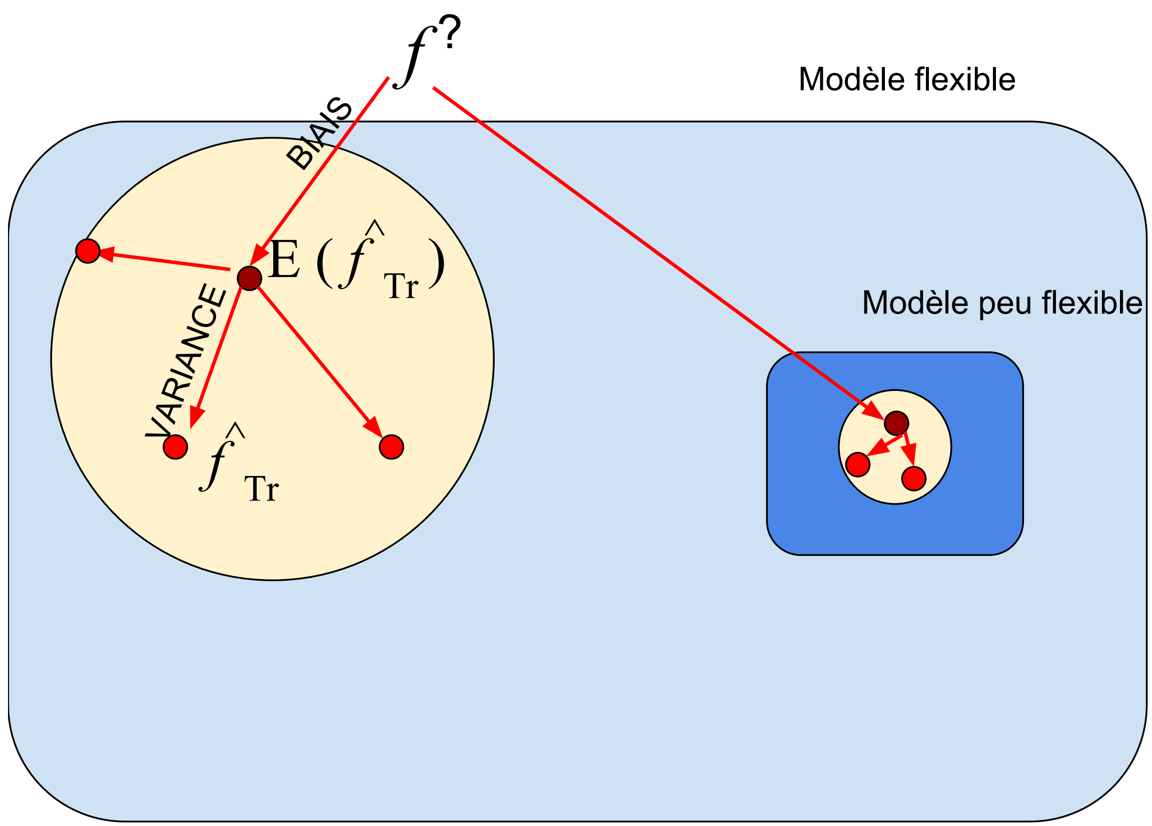
On remarque que plus ν est grand et plus l'estimation colle à $Train$ mais plus elle s'éloigne de $Test$ par endroit. Le modèle a appris par coeur les données

$Train$, et du coup il est mauvais pour estimer des données $Test$. Il y a eu sur-apprentissage.

1.4 Biais-Variance

Ainsi, un modèle flexible a une forte "probabilité" de s'éloigner de $Test$. Insistons sur le mot "probabilité". En effet, la fonction apprise $\widehat{f}$ dépend de l'échantillon d'apprentissage $Train$ qui est aléatoire. Pour montrer cette dépendance, par la suite, nous noterons $T_r = Train$ et $\widehat{f}_{T_r} = \widehat{f}$ (la fonction d'estimation construite à partir de T_r).

Avec $\nu = 30$, la fonction $\widehat{f}_{Tr}$ est très sensible aux variations de T_r. A l'inverse avec $\nu = 1$, $\widehat{f}_{Tr}$ sera quasi la identique d'un tirage à l'autre de T_r. On dit qu'un modèle flexible, a une grande "variance".



f?
Modèle flexible
BIAIS
$E(\hat{f}_{Tr})$
VARIANCE
$\hat{f}_{Tr}$
Modèle peu flexible

1.5 Quantifions

Considérons un x et sa prédiction $\widehat{f}_{T_r}(x)$. L'erreur quadratique pour l'observation (x,y) se décompose comme ceci :

$$\begin{aligned}\mathbf{E}\Big(\widehat{f}_{T_r}(x)-y\Big)^2 &= \mathbf{E}\Big(\widehat{f}_{T_r}(x)-\mathbf{E}[\widehat{f}_{T_r}(x)]+\mathbf{E}[\widehat{f}_{T_r}(x)]-y\Big)^2 \\ &= \underbrace{\mathbf{E}\Big(\widehat{f}_{T_r}(x)-\mathbf{E}[\widehat{f}_{T_r}(x)]\Big)^2}_{\text{variance}} + \underbrace{\Big(\mathbf{E}[\widehat{f}_{T_r}(x)]-y\Big)^2}_{\text{biais}}\end{aligned}$$

(la seule chose aléatoire ci-dessus c'est T_r)

Le premier terme s'appelle la variance en (x,y), le second terme s'appelle le biais de en (x,y).

Mais il est plus naturel d'estimer une erreur quadratique moyenne, une variance moyenne et un biais moyen, ce qui revient à intégrer sur tous les x,y :

$$\begin{aligned}&\int \mathbf{E}\Big(\widehat{f}_{T_r}(x)-y\Big)^2\mathbf{P}[X\in dx, Y\in dy] \\ =\ &\int \mathbf{E}\Big(\widehat{f}_{T_r}(x)-\mathbf{E}[\widehat{f}_{T_r}(x)]\Big)^2\mathbf{P}[X\in dx] \\ +\ &\int \Big(\mathbf{E}[\widehat{f}_{T_r}(x)]-y\Big)^2\mathbf{P}[X\in dx, Y\in dy]\end{aligned}$$

En plus court:

$$\mathbf{E}\Big(\widehat{f}_{T_r}(X) - Y\Big)^2 = \mathbf{E}\Big(\widehat{f}_{T_r}(X) - \mathbf{E}[\widehat{f}_{T_r}(X)]\Big)^2 \;\;+\;\; \mathbf{E}\Big(\mathbf{E}[\widehat{f}_{T_r}(X)] - Y\Big)^2$$

(Cette dernière formule est difficile à comprendre car l'espérance porte sur plusieurs niveaux d'aléatoire)

1.6 Estimons numériquement

Estimons maintenant ces quantités à l'aide de plusieurs réalisations d'échantillons d'apprentissage $(tr_i)_{i\in 1..I}$ et d'une réalisation d'un échantillon test $(x_j, y_j)_{j\in 1..J}$.

Ainsi la décomposition biais variance pour une valeur (x, y) :

$$\mathbf{E}\Big(\widehat{f}_{T_r}(x) - y\Big)^2 = \mathbf{E}\Big(\widehat{f}_{T_r}(x) - \mathbf{E}[\widehat{f}_{T_r}(x)]\Big)^2 \;\;+\;\; \Big(\mathbf{E}[\widehat{f}_{T_r}(x)] - y\Big)^2$$

S'estime par :

$$\begin{aligned}\mathrm{mean}_i\Big[(\widehat{f}_{tr_i}(x) - y)^2\Big] &\simeq \mathrm{mean}_i\Big[(\widehat{f}_{tr_i}(x) - \mathrm{mean}_k[\widehat{f}_{tr_k}(x)])^2\Big] \;\;+\;\; \Big(\mathrm{mean}_i[\widehat{f}_{tr_i}(x)] - y\Big)^2 \\ &\simeq \mathrm{std}_i^2\Big[\widehat{f}_{tr_i}(x)\Big] \;\;+\;\; \Big(\mathrm{mean}_i[\widehat{f}_{tr_i}(x)] - y\Big)^2\end{aligned}$$

La décomposition biais-variance moyenne :

$$\mathbf{E}\Big(\widehat{f}_{T_r}(X) - Y\Big)^2 = \mathbf{E}\Big(\widehat{f}_{T_r}(X) - \mathbf{E}[\widehat{f}_{T_r}(X)]\Big)^2 \quad + \quad \mathbf{E}\Big(\mathbf{E}[\widehat{f}_{T_r}(X)] - Y\Big)^2$$

s'estime par :

$$\text{mean}_j \ \text{mean}_i\Big[(\widehat{f}_{tr_i}(x_j) - y_j)^2\Big] \simeq \text{mean}_j \ \text{std}_i^2\Big[\widehat{f}_{tr_i}(x_j)\Big] \quad + \quad \text{mean}_j\Big(\text{mean}_i[\widehat{f}_{tr_i}(x_j)] - y_j\Big)^2$$

1.7 Remarque sur la "forme"

Ecrire du pseudo code aide beaucoup pour simplémenter des formules ou des algorithme. Au dessus, on a utilisé des notations proches de numpy :

- $\text{mean}_i[u]$ pour `np.mean(u,axis=i)`
- $\text{std}_i^2[u]$ pour `np.std(u,axis=i)**2`

Mathématiquement, la formule d'estimation aurait plutôt été :

$$\frac{1}{IJ}\sum_{i,j}\Big(\widehat{f}_{tr_i}(x_j) - y_j\Big)^2 \simeq \frac{1}{IJ}\sum_{i,j}\Big(\widehat{f}_{tr_i}(x_j) - \frac{1}{I}\sum_k \widehat{f}_{tr_k}(x_j)\Big)^2 \quad + \quad \frac{1}{J}\sum_j\Big(\frac{1}{I}\sum_k \widehat{f}_{tr_k}(x_j) - y_j\Big)$$

Ce qui est plus compliqué à transformer en code !

1.8 La décomposition biais-variance avec une courbe