Méthodes paramétriques

Introduction à l'apprentissage automatique – GIF-4101 / GIF-7005

Professeur : Christian Gagné

Semaine 2

UNIVERSITÉ LAVAL

2.3 Estimation paramétrique

Estimation paramétrique

- Ensemble de données $\mathcal{X} = \{x^t\}_{t=1}^N$ où $x^t \sim p(x)$
  - Variable indépendante et identiquement distribuée (iid)
- Estimation paramétrique
  - Famille de densités de probabilité $p(x|\theta)$
  - Estimation θ : les statistiques suffisantes de la densité
  - Avec une loi normale $\mathcal{N}(\mu,\sigma^2)$, $\theta = \{\mu,\sigma\}$
- Estimation de θ à partir de $\mathcal{X}$

Vraisemblance d'une estimation

- Vraisemblance d'une estimation paramétrée par θ

$$l(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta) = \prod_{t=1}^{N} p(x^t|\theta)$$

  - $p(x|\theta)$ est équivalent à la vraisemblance qu'un échantillon x^t soit obtenu étant donné θ
  - Comme les x^t sont iid, on fait un produit des vraisemblances

Maximum de vraisemblance

- Fonction log-vraisemblance

$$L(\theta|\mathcal{X}) \equiv \log l(\theta|\mathcal{X}) = \sum_{t=1}^{N} \log p(x^t|\theta)$$

  - $\log(ab) = \log(a) + \log(b)$
  - $\log(a^n) = n \log(a)$
- Simplification avec log des équations pour certaines densités (ex. loi normale)
- Estimation du maximum de vraisemblance : trouver θ rendant l'échantillonnage $\mathcal{X}$ le plus probable

$$\theta^* = \underset{\forall\theta}{\operatorname{argmax}}\, L(\theta|\mathcal{X})$$

Exemple : loi de Bernoulli

- Loi de Bernoulli : $P(x) = p^x(1-p)^{1-x}$, $x \in \{0, 1\}$
- Fonction de log-vraisemblance :

$$\begin{aligned} L(p|\mathcal{X}) &= \log \prod_{t=1}^{N} p^{(x^t)}(1-p)^{(1-x^t)} \\ &= \sum_{t=1}^{N} x^t \log p + \left(N - \sum_{t=1}^{N} x^t \right) \log(1-p) \end{aligned}$$

- Estimation du maximum de vraisemblance :

$$\frac{dL(p|\mathcal{X})}{dp} = 0 \quad \Rightarrow \quad \hat{p} = \frac{\sum_{t=1}^{N} x^t}{N}$$

Exemple : loi catégorielle

- Loi catégorielle : généralisation de Bernoulli à K états mutuellement exclusifs
  - État $\mathbf{x} = (x_1, x_2, \ldots, x_K)$, variables $x_i \in \{0, 1\}$ et $\sum_i x_i = 1$
  - Chaque variable x_i a une probabilité p_i, avec $\sum_i p_i = 1$
  - Probabilité d'état : $p(\mathbf{x}) = \prod_{i=1}^{K} p_i^{x_i}$
  - Expériences indépendantes : $\mathcal{X} = \{\mathbf{x}^t\}_{t=1}^{N}$
- Estimation du maximum de vraisemblance :

$$\frac{\partial L(p|\mathcal{X})}{\partial p_i} = 0 \quad \Rightarrow \quad \hat{p}_i = \frac{\sum_t x_i^t}{N}, \; i = 1, \ldots, K$$

Exemple : loi normale

- Loi normale : distribution paramétrée par une moyenne μ et un écart-type σ

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \ -\infty < x < \infty$$

- Vraisemblance selon un échantillonnage $\mathcal{X} = \{x^t\}_{t=1}^N$ avec $x^t \sim \mathcal{N}(\mu,\sigma^2)$

$$L(\mu,\sigma|\mathcal{X}) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x^t-\mu)^2}{2\sigma^2}$$

  - Maximum de vraisemblance avec $\frac{\partial L(\mu,\sigma|\mathcal{X})}{\partial\mu} = 0$ et $\frac{\partial L(\mu,\sigma|\mathcal{X})}{\partial\sigma} = 0$

$$\begin{aligned} m &= \frac{\sum_t x^t}{N} \\ s^2 &= \frac{\sum_t (x^t - m)^2}{N} \end{aligned}$$

Biais d'un estimateur

- $d(\mathcal{X})$, estimation de θ avec $\mathcal{X}$
- Qualité de l'estimation de $d(\mathcal{X})$: $(d(\mathcal{X}) - \theta)^2$
- Qualité de l'estimateur d :

$$r(d,\theta) = \mathbb{E}_{\mathcal{X}}\left[(d(\mathcal{X}) - \theta)^2\right]$$

  - Évaluation de d sur tous les échantillonnages $\mathcal{X}$ possibles
- Biais de l'estimateur

$$b_\theta(d) = \mathbb{E}_{\mathcal{X}}\left[d(\mathcal{X})\right] - \theta$$

  - Estimateur sans biais : $b_\theta(d) = 0$ pour toutes valeurs θ

Rappel : espérance mathématique

- Espérance d'une variable aléatoire continue X ayant une densité $f_X(x)$:

$$\mathbb{E}(X) = \int_{\mathbb{R}} x\, f_X(x)\, dx$$

  - Le théorème de transfert s'applique pour des fonctions mesurables de $g(X)$ de la variable aléatoire X :

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)\, f_X(x)\, dx$$

- Donc pour une constante a, l'espérance de $g(X) = a\,X$ est :

$$\mathbb{E}(a\,X) = \int_{\mathbb{R}} a\,x\, f_X(x)\, dx = a \int_{\mathbb{R}} x\, f_X(x)\, dx = a\,\mathbb{E}(X)$$

- Et pour la somme de deux fonctions de X, $g(X) = m(X) + n(X)$:

$$\mathbb{E}(m(X) + n(X)) = \int_{\mathbb{R}} (m(x) + n(X))\, f_X(x)\, dx = \mathbb{E}(m(X)) + \mathbb{E}(n(X))$$

Biais de l'estimateur m

- Supposons échantillons d'une densité de moyenne μ
    - m est un estimateur sans biais de μ

$$\mathbb{E}_{\mathcal{X}}[m] = \mathbb{E}_{\mathcal{X}}\left[\frac{\sum_t x^t}{N}\right] = \frac{1}{N}\sum_t \mathbb{E}_{\mathcal{X}}[x^t] = \frac{N\mu}{N} = \mu$$

- Variance de l'estimateur

$$\mathrm{Var}_{\mathcal{X}}(m) = \mathrm{Var}_{\mathcal{X}}\left(\frac{\sum_t x^t}{N}\right) = \frac{1}{N^2}\sum_t \mathrm{Var}_{\mathcal{X}}(x^t) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

    - Rappel : $\mathrm{Var}(x) = \mathbb{E}[(x - \mathbb{E}[x])^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$
    - Estimateur efficace : $\lim_{N\to\infty} \mathrm{Var}_{\mathcal{X}}(m) = 0$
- Estimateur convergent : $\lim_{N\to\infty} m = \mu$
    - Loi forte des grands nombres

Biais de l'estimateur s^2

- Écart-type σ d'une loi normale $\mathcal{N}(\mu,\sigma^2)$
  - s^2 est un estimateur avec maximum de vraisemblance de σ^2

$$s^2 = \frac{\sum_t (x^t - m)^2}{N} = \frac{\sum_t (x^t)^2 - Nm^2}{N}$$

  - Qualité de l'estimateur s^2

$$\begin{aligned}
\mathbb{E}_{\mathcal{X}}[(x^t)^2] &= \sigma^2 + \mu^2 \\
\mathbb{E}_{\mathcal{X}}[m^2] &= \sigma^2/N + \mu^2 \\
\mathbb{E}_{\mathcal{X}}[s^2] &= \frac{\sum_t \mathbb{E}_{\mathcal{X}}[(x^t)^2] - N\,\mathbb{E}_{\mathcal{X}}[m^2]}{N} \\
&= \frac{N(\sigma^2+\mu^2) - N(\sigma^2/N + \mu^2)}{N} = \frac{N-1}{N}\sigma^2 \neq \sigma^2
\end{aligned}$$

  - Estimateur s^2 est biaisé !

2.4 Classement bayésien

Classement bayésien

- Règle de Bayes pour le classement

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)} = \frac{p(x|C_i)P(C_i)}{\sum_{k=1}^{K} p(x|C_k)P(C_k)}$$

- Fonction discriminante correspondante ($p(x)$ le même $\forall C_i$)

$$\begin{aligned} \mathrm{h}_i(x) &= p(x|C_i)P(C_i) \\ &\equiv \log p(x|C_i) + \log P(C_i) \end{aligned}$$

- Avec $p(x|C_i)$ suivant une loi normale $\mathcal{N}(\mu_i,\sigma_i^2)$

$$\begin{aligned} p(x|C_i) &= \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right] \\ \mathrm{h}_i(x) &= -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i) \end{aligned}$$

Exemple de classement bayésien

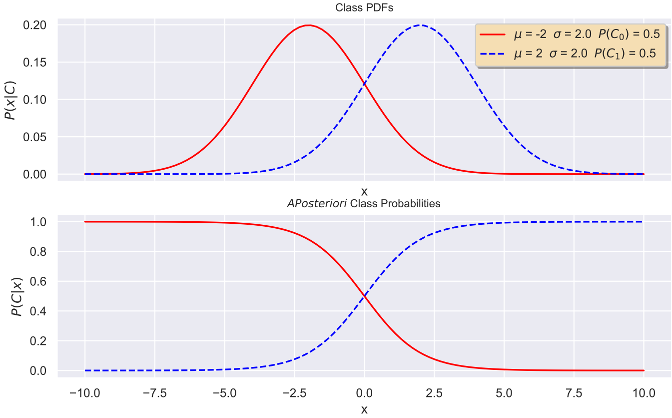
- Supposons jeu $\mathcal{X} = \{x^t, \mathbf{r}^t\}_{t=1}^N$ où $r_i^t = 1$ si $x^t \in C_i$ et $r_i^t = 0$ autrement
  - Estimation des probabilités a priori : $\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$
  - Estimation des moyennes : $m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t}$
  - Estimation des écarts-types : $s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$
- Fonction discriminante correspondante

$$\mathrm{h}_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

- Simplifications
  1. $-\frac{1}{2}\log 2\pi$ est une constante
  2. Supposons une variance égale, $\sigma_i = \sigma_j,\ \forall i,j$
  3. Supposons une probabilité a priori égale, $\hat{P}(C_i) = \hat{P}(C_j),\ \forall i,j$
- On fait alors un classement par la plus proche moyenne

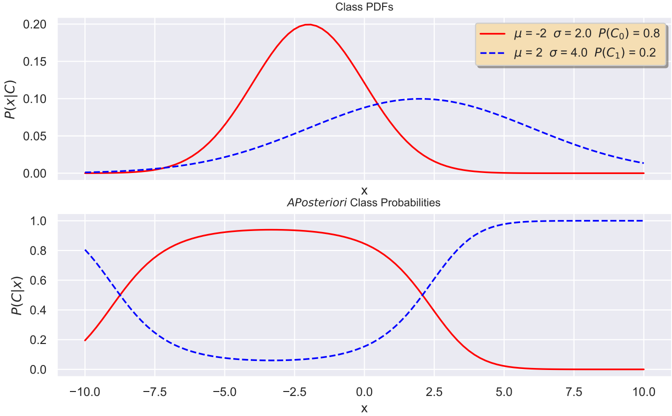
$$\mathrm{h}_i(x) = -(x - m_i)^2 \quad \Rightarrow \quad C_i = \operatorname*{argmin}_{C_k} |x - m_k|$$

Vraisemblances avec deux classes, même variance



Frontière : $\mathrm{h}_1(x) = \mathrm{h}_2(x) \quad \Rightarrow \quad (x - m_1)^2 = (x - m_2)^2 \quad \Rightarrow \quad x = \frac{m_1 + m_2}{2}$

Vraisemblances avec deux classes, variance différente



2.5 Régression

Régression

- Régression d'une fonction $\mathrm{f}(x)$
  - $r = \mathrm{f}(x) + \epsilon$
  - x : variable indépendante
  - $\mathrm{f}(x)$: variable dépendante
  - ϵ : bruit
- Approximation de $\mathrm{f}(x)$ à l'aide de l'estimateur (hypothèse) $\mathrm{h}(x|\theta)$
  - On peut supposer $\epsilon \sim \mathcal{N}(0,\sigma^2)$, un bruit blanc gaussien de moyenne nulle et variance constante σ^2

$$p(r|x) \sim \mathcal{N}(\mathrm{h}(x|\theta),\sigma^2)$$

Estimation selon le maximum de vraisemblance

- Log-vraisemblance avec ensemble d'échantillons $\mathcal{X} = \{x^t, r^t\}_{t=1}^N$ iid

$$
\begin{aligned}
p(x, r) &= p(x \cap r) = p(r|x)p(x) \\
L(\theta|\mathcal{X}) &= \log \prod_{t=1}^{N} p(x^t, r^t) = \log \prod_{t=1}^{N} p(r^t|x^t) + \log \prod_{t=1}^{N} p(x^t)
\end{aligned}
$$

- Comme $p(x^t)$ est indépendant de θ et $p(r|x) \sim \mathcal{N}(\mathrm{h}(x|\theta), \sigma^2)$

$$
\begin{aligned}
L(\theta|\mathcal{X}) &= \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(r^t - \mathrm{h}(x^t|\theta))^2}{2\sigma^2}\right] \\
&= \log\left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(r^t - \mathrm{h}(x^t|\theta))^2\right]\right] \\
&= -N\log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{N}(r^t - \mathrm{h}(x^t|\theta))^2
\end{aligned}
$$

Estimation selon les moindres carrés

- Estimation selon les moindres carrés

$$E(\theta|\mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}(r^t - \mathrm{h}(x^t|\theta))^2$$

- Maximiser la vraisemblance

$$L(\theta|\mathcal{X}) = -N\log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{N}(r^t - \mathrm{h}(x^t|\theta))^2$$

- $-N\log\left(\sqrt{2\pi}\sigma\right)$ et $1/\sigma^2$ sont indépendants de θ
  - $L(\theta|\mathcal{X}) = -\frac{1}{2}\sum_{t=1}^{N}(r^t - \mathrm{h}(x^t|\theta))^2$
  - $E(\theta|\mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}(r^t - \mathrm{h}(x^t|\theta))^2$ est l'erreur quadratique
  - Minimiser $E(\theta|\mathcal{X})$ donne une estimation selon les moindres carrés de θ
  - $\theta^*_{MV} = \underset{\forall\theta}{\operatorname{argmax}}\, L(\theta|\mathcal{X})$ est équivalent à $\theta^*_{MC} = \underset{\forall\theta}{\operatorname{argmin}}\, E(\theta|\mathcal{X})$

Régression linéaire

- Modèle linéaire de $\mathrm{h}(x|\theta)$

$$\mathrm{h}(x^t|w_1,w_0) = w_1 x^t + w_0$$

- Estimation de w_1 et w_0 selon $E(w_1,w_0|\mathcal{X})$

$$\begin{aligned}
\frac{\partial E(w_1,w_0|\mathcal{X})}{\partial w_0} &= \sum_{t=1}^{N}\left(-r^t + w_1 x^t + w_0\right) = 0 \\
\Rightarrow \sum_{t=1}^{N} r^t &= N w_0 + w_1 \sum_{t=1}^{N} x^t \\
\frac{\partial E(w_1,w_0|\mathcal{X})}{\partial w_1} &= \sum_{t=1}^{N}\left(-r^t x^t + w_1 (x^t)^2 + w_0 x^t\right) = 0 \\
\Rightarrow \sum_{t=1}^{N} r^t x^t &= w_0 \sum_{t=1}^{N} x^t + w_1 \sum_{t=1}^{N} (x^t)^2
\end{aligned}$$

Formulation matricielle (ordre 1)

- Formulation matricielle de l'estimation de w_1 et w_0 selon $E(w_1, w_0|\mathcal{X})$

$$
\begin{aligned}
\mathbf{Aw} &= \mathbf{y} \\
\text{où } \mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t (x^t)^2 \end{bmatrix}, \mathbf{w} &= \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}
\end{aligned}
$$

- Résolution avec $\mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$

Formulation matricielle (ordre k)

- Polynôme d'ordre k

$$\mathrm{h}(x^t|w_k,\ldots,w_2,w_1,w_0) = w_k(x^t)^k + \cdots + w_2(x^t)^2 + w_1x^t + w_0$$

- Résolution de l'équation $\mathbf{Aw} = \mathbf{y}$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t & \sum_t (x^t)^2 & \cdots & \sum_t (x^t)^k \\ \sum_t x^t & \sum_t (x^t)^2 & \sum_t (x^t)^3 & \cdots & \sum_t (x^t)^{k+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_t (x^t)^k & \sum_t (x^t)^{k+1} & \sum_t (x^t)^{k+2} & \cdots & \sum_t (x^t)^{2k} \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \\ \sum_t r^t (x^t)^2 \\ \vdots \\ \sum_t r^t (x^t)^k \end{bmatrix}$$

- En posant $\mathbf{A} = \mathbf{D}^\top\mathbf{D}$ et $\mathbf{y} = \mathbf{D}^\top\mathbf{r}$

$$\mathbf{D} = \begin{bmatrix} 1 & x^1 & (x^1)^2 & \cdots & (x^1)^k \\ 1 & x^2 & (x^2)^2 & \cdots & (x^2)^k \\ \vdots & & & & \end{bmatrix}, \mathbf{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

- Résolution selon $\mathbf{w} = (\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{r}$

Autres types d'erreurs

- Erreur quadratique

$$E(\theta|\mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left(r^t - \mathrm{h}(x^t|\theta)\right)^2$$

- Erreur quadratique relative

$$E(\theta|\mathcal{X}) = \frac{\sum_{t=1}^{N}\left(r^t - \mathrm{h}(x^t|\theta)\right)^2}{\sum_{t=1}^{N}(r^t - \bar{r})^2}$$

- Erreur absolue

$$E(\theta|\mathcal{X}) = \sum_{t=1}^{N}\left|r^t - \mathrm{h}(x^t|\theta)\right|$$

2.6 Compromis biais-variance

Compromis biais-variance

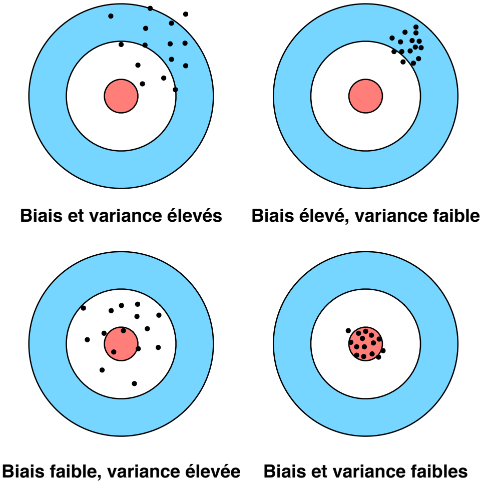
- Erreur quadratique espérée

$$\begin{aligned}\mathbb{E}\left[(\theta - c)^2\right] &= \mathbb{E}\left[(\theta - \mathbb{E}[\theta])^2\right] + (\mathbb{E}[\theta] - c)^2 \\ \mathbb{E}\left[(r - \mathrm{h}(x))^2 | x\right] &= \underbrace{\mathbb{E}\left[(r - \mathbb{E}[r|x])^2 | x\right]}_{\text{bruit}} + \underbrace{(\mathbb{E}[r|x] - \mathrm{h}(x))^2}_{\text{erreur quadratique}}\end{aligned}$$

  - Bruit : ne dépend pas de $\mathrm{h}(\cdot)$ ou $\mathcal{X}$ $\Rightarrow$ ne peut pas être retiré
  - Erreur quadratique : niveau de déviation de $\mathrm{h}(\cdot)$ par rapport à $\mathbb{E}[r|x]$
- Moyenne de $\mathrm{h}(\cdot)$ sur tous les $\mathcal{X} \sim p(r,x)$ possibles

$$\mathbb{E}_{\mathcal{X}}\left[(\mathbb{E}[r|x] - \mathrm{h}(x))^2 \, |x\right] = \underbrace{(\mathbb{E}[r|x] - \mathbb{E}_{\mathcal{X}}[\mathrm{h}(x)])^2}_{\text{biais}^2} + \underbrace{\mathbb{E}_{\mathcal{X}}\left[(\mathrm{h}(x) - \mathbb{E}_{\mathcal{X}}[\mathrm{h}(x)])^2\right]}_{\text{variance}}$$

Biais et variance



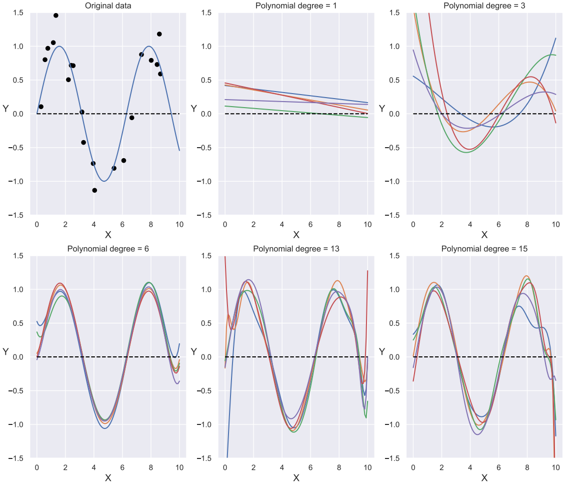
Exemple de compromis biais-variance

- Supposons différents jeux de données $\mathcal{X}_i = \{x_i^t, r_i^t\}$, $i = 1, \ldots, M$, à partir d'une fonction bruitée $\mathrm{f}(\cdot) + \epsilon$
  - En pratique, on ne connaît pas $\mathrm{f}(\cdot)$
  - $\mathrm{h}_i(x)$ généré par apprentissage sur $\mathcal{X}_i$
  - $\mathbb{E}[\mathrm{h}(x)] = \frac{1}{M}\sum_{i=1}^{M} \mathrm{h}_i(x)$
- Biais et variance associés

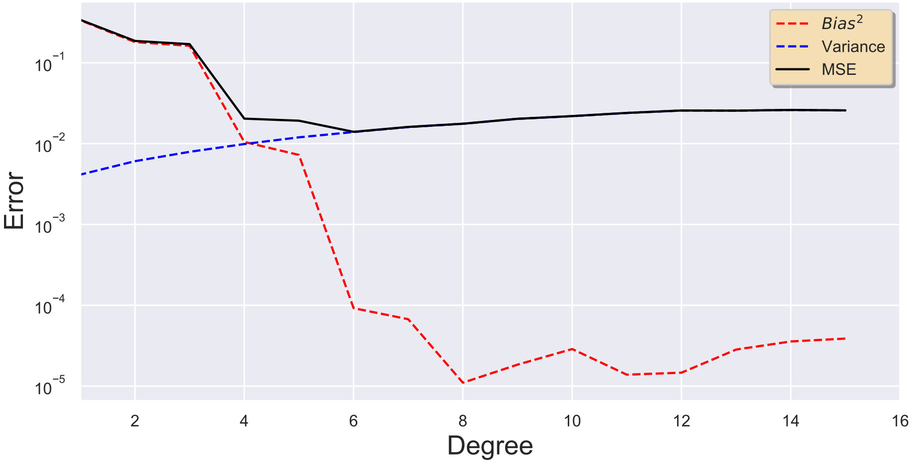
$$\begin{aligned}
\mathrm{biais}^2(\mathrm{h}) &= \frac{1}{N}\sum_{t=1}^{N}\left[\mathbb{E}[\mathrm{h}(x^t)] - \mathrm{f}(x^t)\right]^2 \\
\mathrm{variance}(\mathrm{h}) &= \frac{1}{NM}\sum_{t=1}^{N}\sum_{i=1}^{M}\left[\mathrm{h}_i(x^t) - \mathbb{E}[\mathrm{h}(x^t)]\right]^2
\end{aligned}$$

  - $\mathrm{h}_i(x^t) = c \;\Rightarrow$ biais constant, variance nulle (sous-apprentissage)
  - $\mathrm{h}_i(x^t) = \sum_i r_i^t / N \;\Rightarrow \downarrow$ biais, $\uparrow$ variance
  - Biais faible ou nul, variance élevée : sur-apprentissage

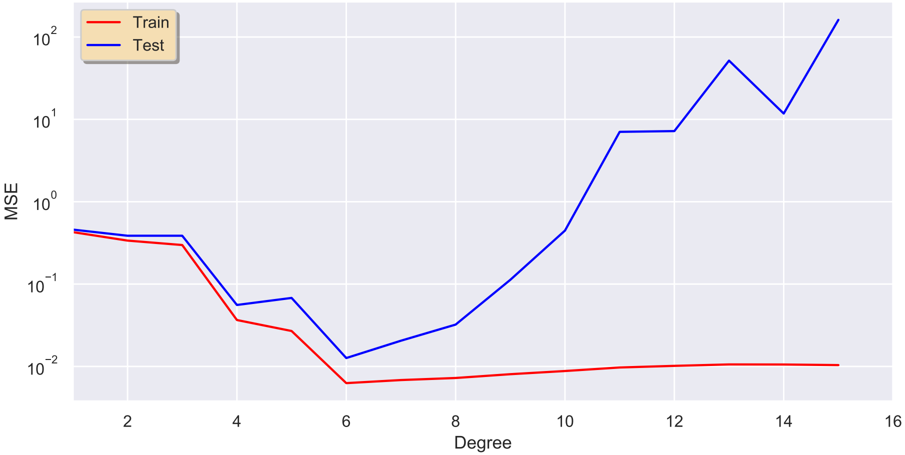
Complexité et compromis biais-variance



Erreur selon le compromis biais-variance



Erreur selon la complexité



Sélection de modèles

- En pratique, on ne peut pas calculer le biais et la variance d'un modèle
  - La validation croisée permet une mesure empirique de l'erreur totale
- Régularisation : intégrer une mesure de complexité dans l'optimisation

$$E' = (\text{erreur empirique}) + \lambda\ (\text{complexité du modèle})$$

  - λ contrôle la pénalité de complexité
  - λ généralement ajusté par validation croisée
- Mesures de complexité
  - Dimension Vapnik-Chervonenkis (VC-dim)
  - *Minimum description length* : description de taille minimale de la donnée