

①

Leçon n° 1: Introduction à l'apprentissage supervisé.

Objectifs

- Modélisation d'un problème d'apprentissage supervisé.
- Définition des métriques d'évaluation des performances
- Présentation de familles d'algorithmes classiques pour résoudre le problème
- Introduction de la nécessité de poser des hypothèses.
- Présentation des autres paradigmes classiques de l'apprentissage statistique.

I. Des données et des prédictions

Observations: $(x_i, y_i) \in X \times Y, i=1, \dots, n$
 (Données d'entraînement, Training data)

Inputs:

- Images, - Sons, - Videos, - Textes, ..., - \mathbb{R}^d , ...

Outputs:

- Classification binaire $\in \{0, 1\}$ ou $\in \{-1, 1\}$
- Classification multiclasse $\in \{1, \dots, k\}$
- \mathbb{R}^d
- ...

- ② Objectif: À partir d'un nouvel "input" x , "prédire" l'output y qui lui correspond le mieux. (Données de test, Testing Data)
- $y \approx f(x)$
 ↑
 apprise avec les données d'entraînement.



- y peut être une fonction aléatoire de x
 - f peut être complexe
 - On n'observe qu'un nombre fini de données d'entraînement, il faut donc trouver un compromis entre interpolation des données d'entraînement et extrapolation à de nouvelles données
 - La dimension des données d peut être élevée.
- Underfitting / Overfitting

II. Formalisation Mathématique

- Modélisation avec une densité de probabilité p sur $X \times Y$
- Hypothèse d'indépendance:
 $(x_1, y_1), \dots, (x_n, y_n) \stackrel{i.i.d}{\sim} p$
- Beaucoup d'abus de notations (ex $p = p^{\otimes n} = \dots$) mais n'introduit pas de problème.

③

1) Fonction de perte:

$$P: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$$

$P(y, \hat{y})$ = erreur induite par la prédiction de \hat{y} alors que la vérité était y .

Exemples:

- Classification binaire: $P(y, \hat{y}) = \mathbb{1}_{y \neq \hat{y}}$
- Classification multiclasse: $P(y, \hat{y}) = \mathbb{1}_{y \neq \hat{y}}$
- Régression: $P(y, \hat{y}) = (y - \hat{y})^2$
 $P(y, \hat{y}) = |y - \hat{y}|$
- ...

2) Risques

Maintenant que nous avons une fonction de perte, nous pouvons nous intéresser à l'erreur d'une fonction de prédiction f sur la population ou sur les données d'entraînement.

- Risque moyen / Erreur de test / Erreur de généralisation:

$$\mathcal{R}(f) := \mathbb{E}(P(y, f(x))) = \int_{\mathcal{X} \times \mathcal{Y}} P(y, f(x)) d\rho(x, y)$$

- Exemples

- Classification binaire:


$$\begin{aligned} \mathcal{R}(f) &= 0 \times \mathbb{P}(f(x) = y) + 1 \times \mathbb{P}(f(x) \neq y) \\ &= \mathbb{P}(f(x) \neq y) \end{aligned}$$

④ Classification multiclasse: Idem

• Régression:

$$P(y, s) = (y - s)^2 \Rightarrow R(f) = \mathbb{E}((y - f(x))^2)$$

$$P(y, s) = |y - s| \Rightarrow R(f) = \mathbb{E}(|y - f(x)|)$$

 On aimerait minimiser $R(f)$ en f , mais c'est impossible en pratique car il faudrait connaître la distribution des données. En revanche, l'est possible d'approximer le risque par une quantité empirique.

Risque empirique / Erreur d'entraînement

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^n P(y_i, f(x_i))$$

Pour justifier l'approximation $R(f) \approx \hat{R}(f)$, nous avons les deux résultats asymptotiques suivants (sous certaines hypothèses):

• $\hat{R}(f) \xrightarrow{P.S.} R(f)$ (Loi forte des grands nombres)

• $\sqrt{n} (\hat{R}(f) - R(f)) \xrightarrow{Loi} \mathcal{N}(0, \text{Var}(P(y, f(x))))$
(Loi forte des grands nombres).

⑤ 3) Risque de Bayes et Prédicteur de Bayes.

Proposition - Définition:

Le risque moyen $M(\cdot)$ est minimisé par le prédicteur de Bayes $f_* : X \rightarrow Y$ qui satisfait pour tout $x' \in X$

$$f_*(x') \in \operatorname{argmin}_{s \in Y} \underbrace{\mathbb{E}(P(y, s) | x = x')}_{:= r(s | x')}$$

Le risque de Bayes est défini comme le risque du prédicteur de Bayes

$$M^* := M(f_*) = \mathbb{E}_{x'} \left[\inf_{s \in Y} \mathbb{E}(P(y, s) | x = x') \right]$$

Preuve: $M(f) - M^* = M(f) - M(f_*)$

$$= \int_X \underbrace{r(f(x') | x')}_{\geq 0} - \underbrace{\min_{s \in Y} r(s | x')}_{\geq 0} dp(x')$$

≥ 0

□

Excès de risque : $M(f) - M^*$

⑥

• Exemples :

- Classification $P(y, s) = \mathbb{1}_{y \neq s}$

$$f_{\alpha}(x') \in \underset{S}{\operatorname{argmin}} P(y \neq s | x = x')$$

$$= \underset{S}{\operatorname{argmax}} P(y = s | x = x')$$

Il s'agit de l'output le plus probable

- Régression $P(y, s) = (y - s)^2$

$$f_{\alpha}(x') \in \underset{S}{\operatorname{argmin}} \mathbb{E}((y - s) | x = x')$$

$$= \underset{S}{\operatorname{argmin}} \left\{ \mathbb{E}((y - \mathbb{E}(y | x = x'))^2 | x = x') + (s - \mathbb{E}(y | x = x'))^2 \right\}$$

$$\text{donc } f_{\alpha}(x') = \mathbb{E}(y | x = x')$$

⑦ III. Paradigmes classiques d'apprentissage

1) Minimisation du risque empirique

Contexte: $f_0: X \rightarrow Y, \theta \in \Theta$.

Idee $R(f_0) \approx \hat{M}(f_0)$

Estimateur: $\hat{\theta} \in \arg\min_{\theta \in \Theta} \hat{M}(f_\theta) = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_\theta(x_i))$

Exemple: Régression Linéaire $f_0(x) = \theta^T \phi(x)$

$$\hat{M}(f_\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta^T \phi(x_i))^2$$

↑
feature map

Décomposition du risque Pour un estimateur $\hat{\theta}$,

$$M(f_{\hat{\theta}}) - M^* = \underbrace{\left(M(f_{\hat{\theta}}) - \inf_{\theta \in \Theta} M(f_{\theta'}) \right)}_{\text{Erreurs d'estimation}} + \underbrace{\left(\inf_{\theta \in \Theta} M(f_{\theta'}) - M^* \right)}_{\text{Erreurs d'approximation}}$$

Pour plus tard: Contrôle de l'erreur d'estimation

Soit $\theta' \in \arg\min_{\theta \in \Theta} M(f_{\theta'})$,

$$\textcircled{8} \quad M(p_{\hat{\sigma}}) - M(p_{\sigma'}) = \underbrace{(M(p_{\hat{\sigma}}) - \hat{M}(p_{\hat{\sigma}}))}_{\text{Erreur d'optimisation}} + \underbrace{(\hat{M}(p_{\hat{\sigma}}) - \hat{M}(p_{\sigma'}))}_{\text{Erreur d'approximation}} + \underbrace{(\hat{M}(p_{\sigma'}) - M(p_{\sigma'}))}_{\text{Erreur d'approximation}}$$

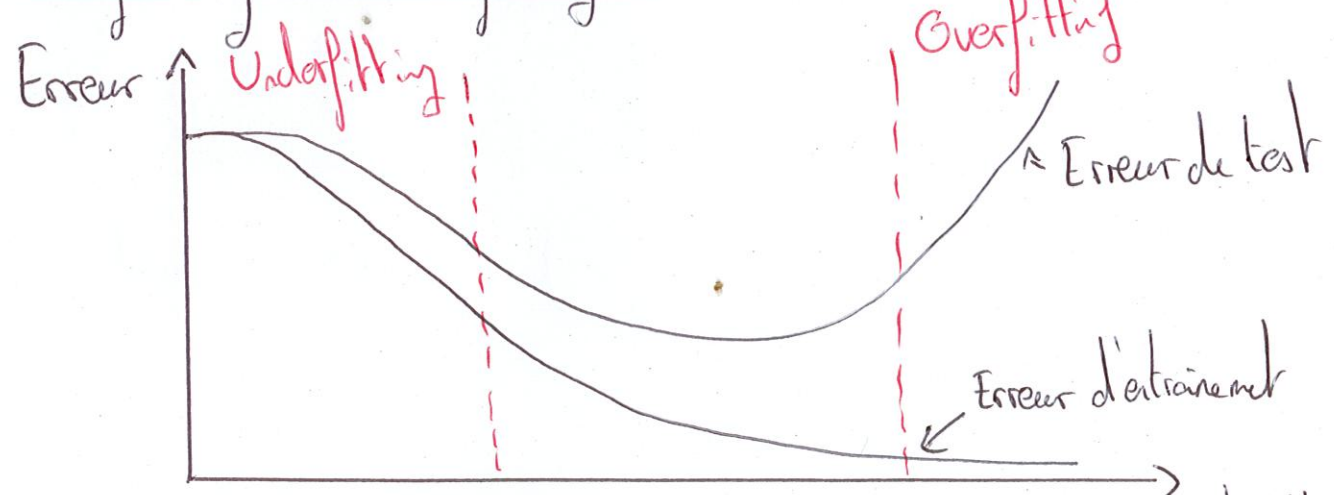
$$\leq 2 \sup_{\sigma \in \mathcal{H}} |\hat{M}(p_{\sigma}) - M(p_{\sigma})|$$

$$+ \sup_{\sigma \in \mathcal{H}} (\hat{M}(p_{\hat{\sigma}}) - \hat{M}(p_{\sigma}))$$

Maximale de RVs d'espérances nulles

Erreur d'optimisation

Underfitting VS Overfitting



Grosse erreur d'approximation:

- faible variance
- fort biais

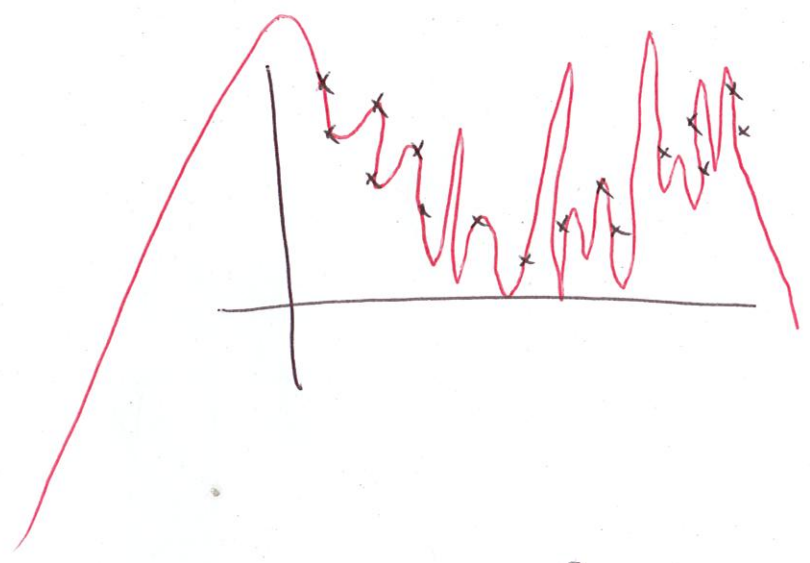
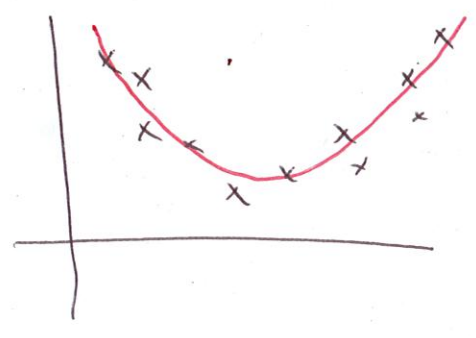
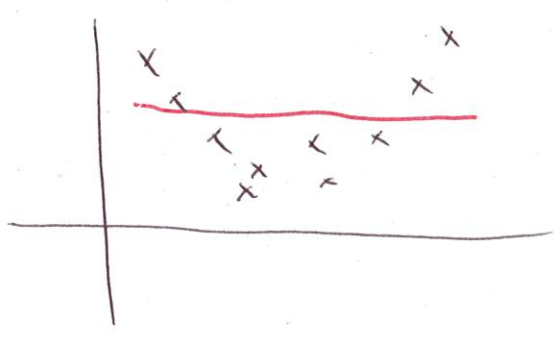
"Complexité" de \mathcal{H}

Grosse erreur d'estimation:

- faible biais
- forte variance

(9)

Exemple: Régression polynomiale et polynômes interpolateurs de Lagrange.



Régularisation: $\hat{\theta} = \arg \min_{\theta} \left\{ \sum_{i=1}^n (y_i - \hat{h}(x_i; \theta))^2 + \underbrace{\Omega(\theta)}_{\text{Pénalisation en fonction de la complexité}} \right\}$

cf. chapitres suivants

⑩ 2) Méthodes par moyennes locales



Approximer le prédicteur de Bayes en supposant une régularité en x .

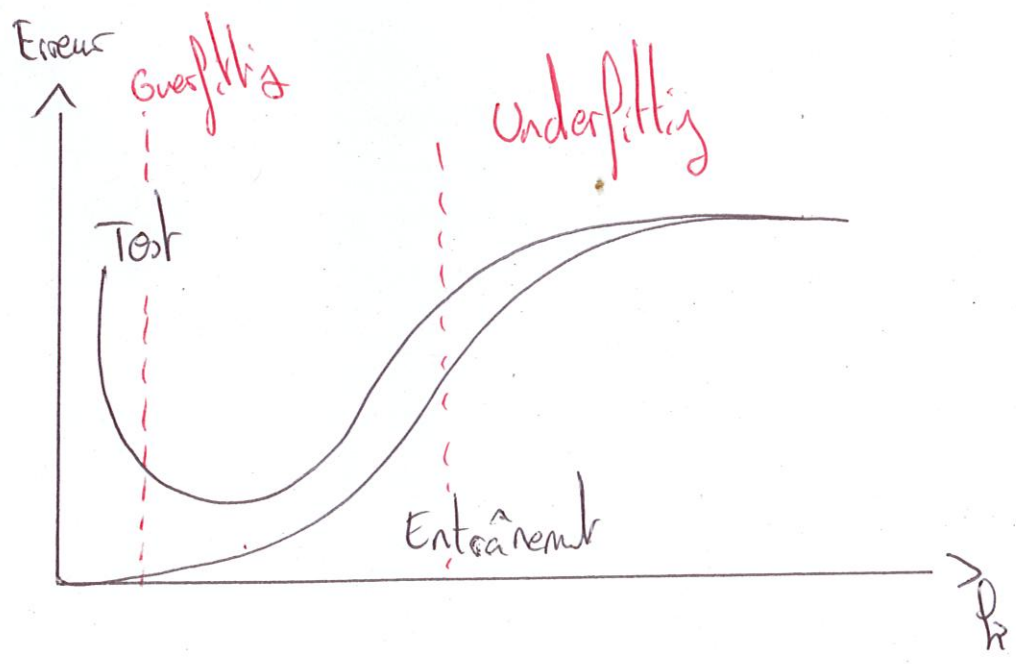
Exemple: Classificateur des k plus proches voisins

• Prédicteur de Bayes: $f_z(x') = \underset{J}{\operatorname{argmax}} P(y = J | x = x')$

• Prédicteur des k -plus proches voisins:

$f_k(x') =$ classe majoritaire parmi les k plus proches voisins de x' dans le jeu d'entraînement.

Underfitting VS Overfitting



⑪ IV. "No Free Lunch Theorem"

TLDR: Il n'existe pas d'algorithme magique fonctionnant bien sur toutes les distributions. Chaque algorithme a des avantages et des inconvénients, et il faut faire des hypothèses sur les données pour obtenir des résultats intéressants.

Théorème: Considérons un problème de classification binaire avec la fonction de perte du 0-1 tel que X est infini.

Soit \mathcal{P} l'ensemble des distributions de probabilités sur $X \times \{0, 1\}$.
Alors, pour tout n , et tout algorithme A ,

$$\sup_{p \in \mathcal{P}} \left(\mathbb{E} \left[\left| A \left(\text{Jeu de données de taille } n \text{ selon } p \right) - M^* \right| \right] \right) \geq \frac{1}{2}$$

⑫ V. Autres thématiques du Machine Learning

• Apprentissage non supervisé

On n'observe pas des couples (x, y) , mais simplement des x . Le but est d'apprendre la structure de X .

- ex:
- Analyse en composantes principales
 - Clustering
 - Modèles probabilistes

(12)

◦ Apprentissage en ligne :

Les données arrivent sous forme de flux.

◦ Apprentissage par renforcement

L'algorithme peut agir avec l'environnement. (ex: Jeux Vidéo)

◦ Génération de données :

Apprendre la structure des données de sorte à pouvoir en générer de nouvelles. (ex: Chat GPT, modèles de diffusion)