Machine Learning

Concepts et méthodologie

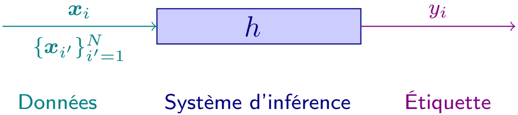
marie.szafranski@ensiie.fr

Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique. . .

Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique. . .



Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique...

$$\xrightarrow[\{\boldsymbol{x}_{i'}\}_{i'=1}^{N}]{\boldsymbol{x}_i} \boxed{\quad h \quad} \xrightarrow{y_i}$$

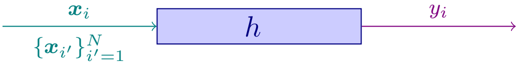
Données — Système d'inférence — Étiquette

Observations, exemples, individus

décrites par des variables explicatives

Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique...

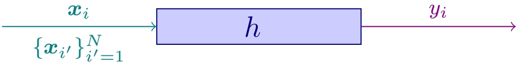


Données | Système d'inférence | Étiquette

Réponse, variable expliquée

Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique...

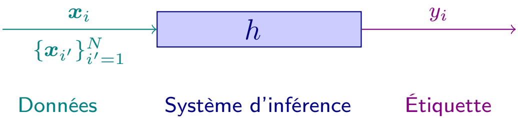


Données Système d'inférence Étiquette

Apprenant, prédicteur, classifieur, modèle

Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique...

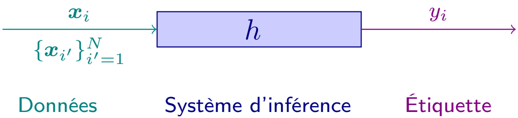


Apprentissage *automatique* vs apprentissage *statistique* ?

- Approche algorithmique pour l'inférence
- Approche théorique des garanties offertes par le modèle

Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique...



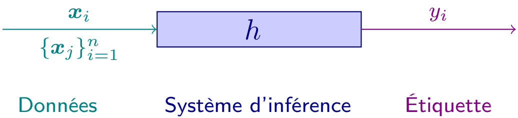
Apprentissage *automatique* vs apprentissage *statistique* ?

- Approche algorithmique pour l'inférence
- Approche théorique des garanties offertes par le modèle

Approches complémentaires et indissociables

Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique. . .



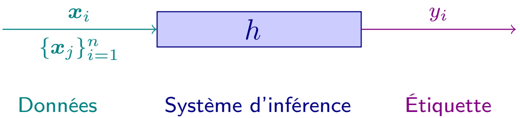
Exemples classiques

Catégorisation de texte

- x_i : suite finie de mots-clés caractérisant un document i $\qquad x_i \in \mathcal{X}$
- y_i : catégorie (auteur, sujet, etc.) $\qquad y_i \in \mathcal{C}$

Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique…



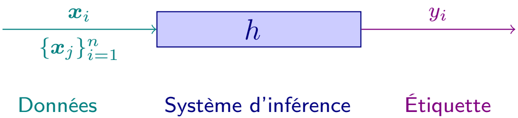
Exemples classiques

Reconnaissance de chiffres manuscrits

- $\boldsymbol{x}_i$: image i en niveau de gris sur M pixels $\qquad \boldsymbol{x}_i \in [0,1]^M$
- y_i : chiffre $\qquad y_i \in \{0, \cdots, 9\}$

Parlons « machine learning »

Apprentissage automatique, numérique, artificiel, statistique. . .



Exemples classiques

Aide au diagnostic médical

- $\boldsymbol{x}_i$: fréquence cardiaque, respiratoire, etc. du patient i $\qquad \boldsymbol{x}_i \in \mathbb{R}^M$
- y_i : diagnostic, prognostic $\qquad y_i \in \{\text{malade}, \text{sain}\},\ y_i \in \mathbb{N}$

Parlons « machine learning »

Différents apprentissages

- Non-supervisé $S_u = \{\boldsymbol{x}_u\}_{u=1}^N$
  $\rightarrow$ Inférer les étiquettes $\{y_u\}_{u=1}^N$ de S_u (classes inconnues)

- Supervisé et semi-supervisé $S_\ell = \{(\boldsymbol{x}_\ell, y_\ell)\}_{\ell=1}^N$ et S_u
  $\rightarrow$ Inférer les étiquettes y d'exemples $\boldsymbol{x}$ inédits

- Par renforcement $S_t = \{(e_t, a_t, r_t, e_{t+1})\}_{t=1}^N$
  $\rightarrow$ Inférer une politique $\Pi : a_t(e_t) \rightsquigarrow e_{t+1} : r_t$
  $\rightarrow$ Maximiser la somme des récompenses

Parlons « machine learning »

Différents apprentissages

- Non-supervisé $S_u = \{\boldsymbol{x}_u\}_{u=1}^N$
  $\rightarrow$ Inférer les étiquettes $\{y_u\}_{u=1}^N$ de S_u (classes inconnues)

- Supervisé et semi-supervisé $S_\ell = \{(\boldsymbol{x}_\ell, y_\ell)\}_{\ell=1}^N$ et S_u
  $\rightarrow$ Inférer les étiquettes y d'exemples $\boldsymbol{x}$ inédits

- Par renforcement $S_t = \{(e_t, a_t, r_t, e_{t+1})\}_{t=1}^N$
  $\rightarrow$ Inférer une politique $\Pi : a_t(e_t) \rightsquigarrow e_{t+1} : r_t$
  $\rightarrow$ Maximiser la somme des récompenses

Parlons « machine learning »

Différents apprentissages

- Non-supervisé $S_u = \{x_u\}_{u=1}^N$
  $\rightarrow$ Inférer les étiquettes $\{y_u\}_{u=1}^N$ de S_u (classes inconnues)

- Supervisé et semi-supervisé $S_\ell = \{(x_\ell, y_\ell)\}_{\ell=1}^N$ et S_u
  $\rightarrow$ Inférer les étiquettes y d'exemples x inédits

- Par renforcement $S_t = \{(e_t, a_t, r_t, e_{t+1})\}_{t=1}^N$
  $\rightarrow$ Inférer une politique $\Pi : a_t(e_t) \rightsquigarrow e_{t+1} : r_t$
  $\rightarrow$ Maximiser la somme des récompenses

Parlons « machine learning »

Apprentissage supervisé

$$S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$$
$$\forall i,\ (\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$$

Parlons « machine learning »

Apprentissage supervisé

$$S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$$
$$\forall i,\ (\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$$

Selon $\mathcal{Y}$

- Régression — $y \in \mathbb{R}$ ou $y \in \mathbb{R}^C$, avec $C \in \mathbb{N}$
- Classification ou discrimination — $y \in \{1, \cdots, C\}$, avec $C \in \mathbb{N}$

  Classification binaire — $C = 2$, $y \in \{\pm 1\}$

Parlons « machine learning »

Apprentissage supervisé

$$S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$$
$$\forall i,\ (\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$$

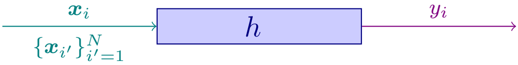
Selon $\mathcal{Y}$

- Régression — $y \in \mathbb{R}$ ou $y \in \mathbb{R}^C$, avec $C \in \mathbb{N}$
- Classification ou discrimination — $y \in \{1, \cdots, C\}$, avec $C \in \mathbb{N}$

  Classification binaire — $C = 2,\ y \in \{\pm 1\}$

Parlons « machine learning »

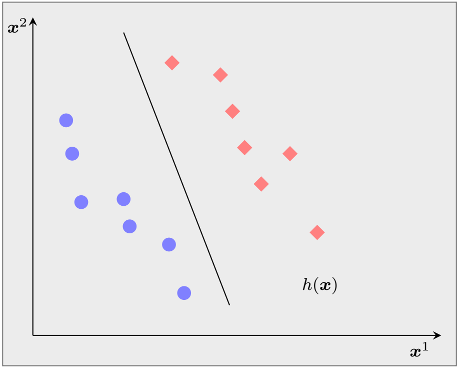
Apprentissage automatique, numérique, artificiel, statistique. . .



Données	Système d'inférence	Étiquette
$\in \mathcal{X}$	$\in \mathcal{H}$	$\in \mathcal{Y}$

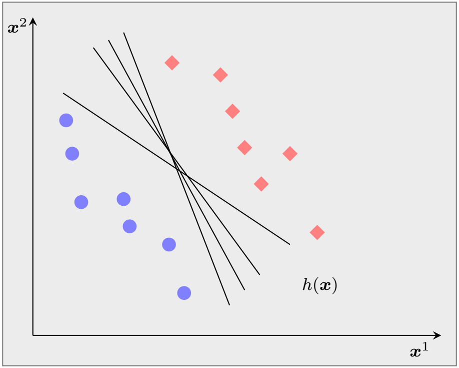
Quelques intuitions

- Espace de recherche $\mathcal{H}$ potentiellement infini $h \in \mathcal{H}$
- Capacité de généralisation de h
- Qualité de l'échantillon x



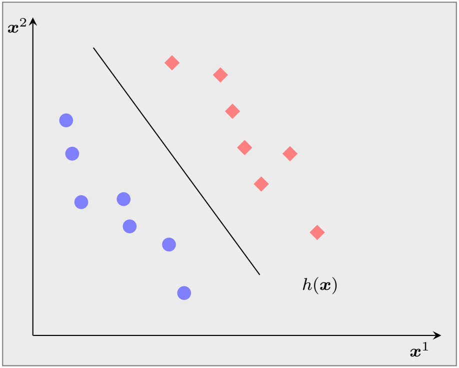
Quelques intuitions

- Espace de recherche $\mathcal{H}$ potentiellement infini $h \in \mathcal{H}$
- Capacité de généralisation de h
- Qualité de l'échantillon x



Quelques intuitions

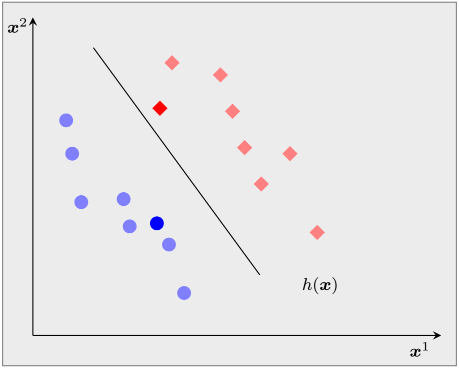
- Espace de recherche $\mathcal{H}$ potentiellement infini $h \in \mathcal{H}$
- Capacité de généralisation de h
- Qualité de l'échantillon x



Quelques intuitions

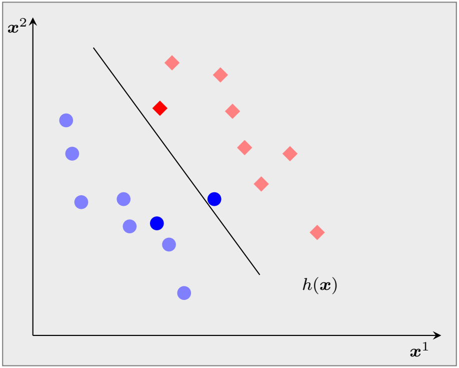
- Espace de recherche $\mathcal{H}$ potentiellement infini
- Capacité de généralisation de h
- Qualité de l'échantillon x

$h \in \mathcal{H}$



Quelques intuitions

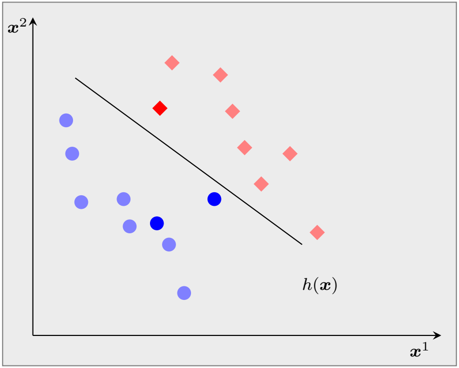
- Espace de recherche $\mathcal{H}$ potentiellement infini $h \in \mathcal{H}$
- Capacité de généralisation de h
- Qualité de l'échantillon x



Quelques intuitions

- Espace de recherche $\mathcal{H}$ potentiellement infini
- Capacité de généralisation de h
- Qualité de l'échantillon x

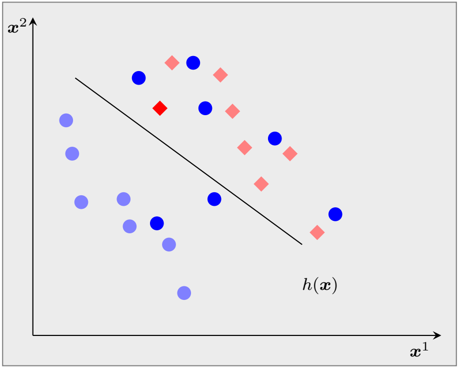
$h \in \mathcal{H}$



Quelques intuitions

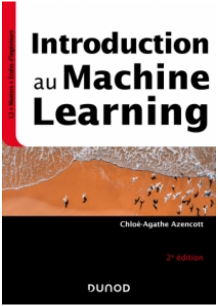
$h \in \mathcal{H}$

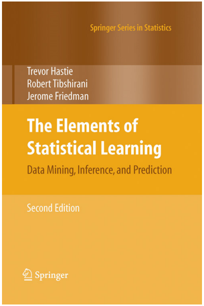
- Espace de recherche $\mathcal{H}$ potentiellement infini
- Capacité de généralisation de h
- Qualité de l'échantillon $\boldsymbol{x}$



Références en ligne (légalement)

WikiStat

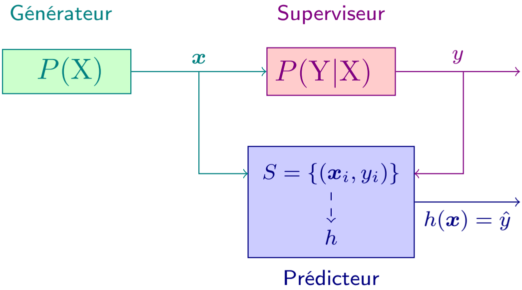




En français

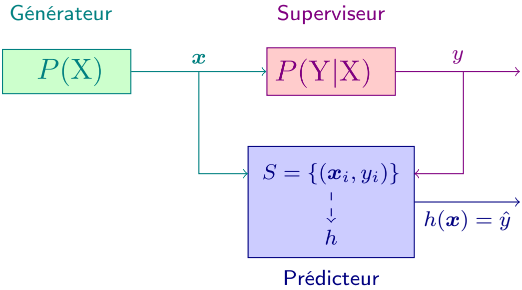
En français

Composante d'un modèle d'apprentissage



- Générateur de vecteurs aléatoires X $\quad \boldsymbol{x} \overset{\text{iid}}{\sim} P(\mathrm{X})$ (fixe mais inconnue)
- Superviseur de variables aléatoires Y $\quad y \overset{\text{iid}}{\sim} P(\mathrm{Y}|\mathrm{X})$ (fixe mais inconnue)
- Prédicteur $h \in \mathcal{H}$ $\quad \mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ (espace des hypothèses)

Composante d'un modèle d'apprentissage



$$S \overset{\text{iid}}{\sim} P(\mathrm{X}, \mathrm{Y}) = P(\mathrm{X}).P(\mathrm{Y}|\mathrm{X})$$

- Obtenir des valeurs de $\hat{y_i}$ proches de y_i, $\forall(\boldsymbol{x}_i, y_i) \in S$
- Prédire des valeurs $\hat{y}$ pour des exemples $\boldsymbol{x}$ inédits

Remarques

Une solution h déterministe, mais...

- Les données $\boldsymbol{x}$ sont peut-être très bruitées
- L'espace des description $\mathcal{X}$ est peut-être incomplet
- La solution $\boldsymbol{y}$ du problème cible n'est peut-être pas déterministe

Une solution h paramétrique ou non ?

- Pas d'*a priori* sur la forme de $P(\mathrm{X})$
- Pas de contrainte *a priori* sur l'espace de recherche $\mathcal{H}$

Remarques

Une solution h déterministe, mais...

- Les données $\boldsymbol{x}$ sont peut-être très bruitées
- L'espace des description $\mathcal{X}$ est peut-être incomplet
- La solution $\boldsymbol{y}$ du problème cible n'est peut-être pas déterministe

Une solution h paramétrique ou non ?

- Pas d'*a priori* sur la forme de $P(\mathrm{X})$
- Pas de contrainte *a priori* sur l'espace de recherche $\mathcal{H}$

Le but du jeu

Choisir la solution h qui « se trompe le moins »

Fonction de perte

Définition

Coût, critère

Quantifier l'erreur comise par h pour évaluer y

$$\begin{array}{rccc} L: & \mathcal{Y} \times \mathcal{Y} & \to & \mathbb{R}^+ \\ & (y, h(\boldsymbol{x})) & \mapsto & L(y, h(\boldsymbol{x})) \end{array}$$

0 si $y = h(\boldsymbol{x})$ et > 0 sinon

Fonction de perte

Définition

Coût, critère

Quantifier l'erreur comise par h pour évaluer y

$$\begin{array}{rccc} L: & \mathcal{Y} \times \mathcal{Y} & \to & \mathbb{R}^+ \\ & (y, h(\boldsymbol{x})) & \mapsto & L(y, h(\boldsymbol{x})) \end{array}$$

0 si $y = h(\boldsymbol{x})$ et > 0 sinon

$|y - h(\boldsymbol{x})|$

$(y - h(\boldsymbol{x}))^2$

$\mathbb{I}\left[y \neq h(\boldsymbol{x})\right]$

? ? ?

Fonction de perte

Définition

Coût, critère

Quantifier l'erreur comise par h pour évaluer y

$$L: \begin{array}{ccc} \mathcal{Y} \times \mathcal{Y} & \to & \mathbb{R}^+ \\ (y, h(\boldsymbol{x})) & \mapsto & L(y, h(\boldsymbol{x})) \end{array}$$

0 si $y = h(\boldsymbol{x})$ et > 0 sinon

$|y - h(\boldsymbol{x})|$

$(y - h(\boldsymbol{x}))^2$

$\mathbb{I}\left[y \neq h(\boldsymbol{x})\right]$

? ? ?

Fonction de perte

Définition

Coût, critère

Quantifier l'erreur comise par h pour évaluer y

$$L : \begin{array}{ccc} \mathcal{Y} \times \mathcal{Y} & \rightarrow & \mathbb{R}^+ \\ (y, h(\boldsymbol{x})) & \mapsto & L(y, h(\boldsymbol{x})) \end{array}$$

0 si $y = h(\boldsymbol{x})$ et > 0 sinon

$|y - h(\boldsymbol{x})|$

$(y - h(\boldsymbol{x}))^2$

$\mathbb{I}\left[y \neq h(\boldsymbol{x})\right]$

$???$

Fonction de perte

Définition

Coût, critère

Quantifier l'erreur comise par h pour évaluer y

$$L : \begin{array}{ccc} \mathcal{Y} \times \mathcal{Y} & \to & \mathbb{R}^+ \\ (y, h(\boldsymbol{x})) & \mapsto & L(y, h(\boldsymbol{x})) \end{array}$$

0 si $y = h(\boldsymbol{x})$ et > 0 sinon

$|y - h(\boldsymbol{x})|$

$(y - h(\boldsymbol{x}))^2$

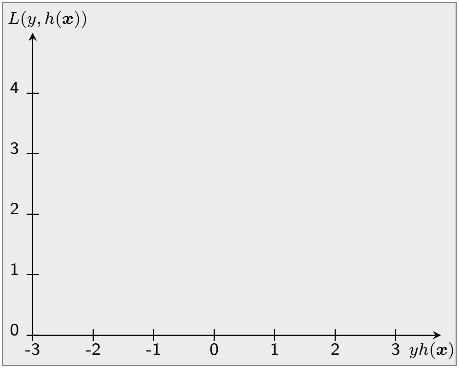
$\mathbb{I}\left[y \neq h(\boldsymbol{x})\right]$

? ? ?

Fonction de perte

Définir une perte $L(y, h(\boldsymbol{x}))$ cohérente avec la tâche

Classification binaire : $y = \{\pm 1\}$



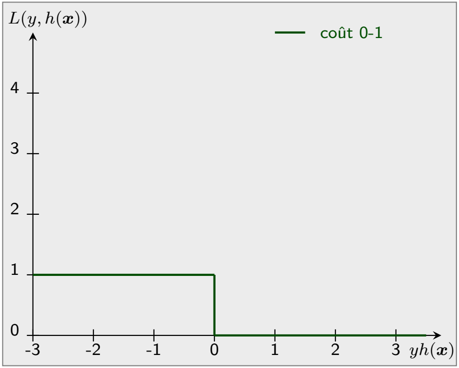
Soit $d(\boldsymbol{x}) = \text{sign}\,(h(\boldsymbol{x}))$

- $L(y, h(\boldsymbol{x})) = I_{\text{sign}}\,[y \neq h(\boldsymbol{x})]$
- $L(y, h(\boldsymbol{x})) = [1 - yh(\boldsymbol{x})]_+$

Fonction de perte

Définir une perte $L(y, h(\boldsymbol{x}))$ cohérente avec la tâche

Classification binaire : $y = \{\pm 1\}$



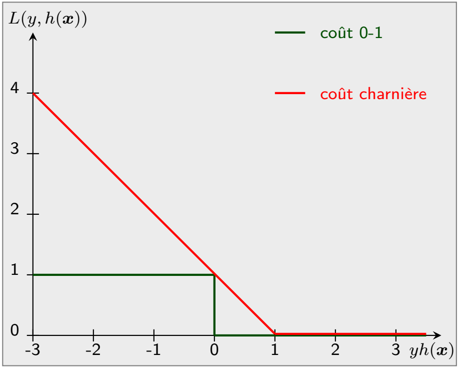
Soit $d(\boldsymbol{x}) = \text{sign}\,(h(\boldsymbol{x}))$

- $L(y, h(\boldsymbol{x})) = I_{\text{sign}}\,[y \neq h(\boldsymbol{x})]$
- $L(y, h(\boldsymbol{x})) = [1 - yh(\boldsymbol{x})]_+$

Fonction de perte

Définir une perte $L(y, h(\boldsymbol{x}))$ cohérente avec la tâche

Classification binaire : $y = \{\pm 1\}$



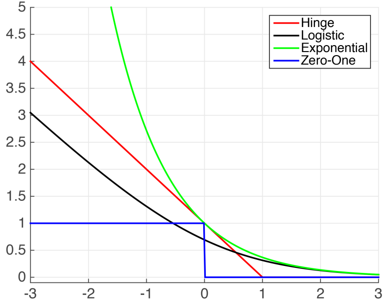
Soit $d(\boldsymbol{x}) = \operatorname{sign}(h(\boldsymbol{x}))$

- $L(y, h(\boldsymbol{x})) = I_{\text{sign}}[y \neq h(\boldsymbol{x})]$
- $L(y, h(\boldsymbol{x})) = [1 - yh(\boldsymbol{x})]_+$

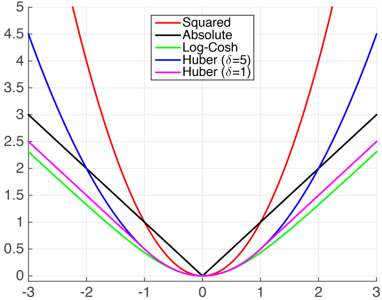
Fonction de perte

Définir une perte $L(y, h(\boldsymbol{x}))$ cohérente avec la tâche

Classification binaire



Régression



Crédits : `http://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote10.html`

Fonction de perte et risque réel

Fonction de perte

Coût, critère

Quantifier l'erreur comise par h pour évaluer y

$$\begin{array}{rccc} L: & \mathcal{Y}\times\mathcal{Y} & \rightarrow & \mathbb{R}^+ \\ & (y, h(\boldsymbol{x})) & \mapsto & L(y, h(\boldsymbol{x})) \end{array}$$

0 si $y = h(\boldsymbol{x})$ et > 0 sinon

Risque réel

$$R(h) = E[L(\mathrm{Y}, h(\mathrm{X}))] = P(h(\mathrm{X}) \neq \mathrm{Y})$$

- $R(h) = \int\int_{\mathcal{X}\times\mathcal{Y}} L(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)d\boldsymbol{x}dy$ Régression
- $R(h) = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} L(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)d\boldsymbol{x}$ Classification

Fonction de perte et risque réel

Fonction de perte

Coût, critère

Quantifier l'erreur comise par h pour évaluer y

$$L: \begin{array}{ccc} \mathcal{Y} \times \mathcal{Y} & \to & \mathbb{R}^+ \\ (y, h(\boldsymbol{x})) & \mapsto & L(y, h(\boldsymbol{x})) \end{array}$$

0 si $y = h(\boldsymbol{x})$ et > 0 sinon

Risque réel

$$R(h) = E[L(\mathrm{Y}, h(\mathrm{X}))] = P(h(\mathrm{X}) \neq \mathrm{Y})$$

- $R(h) = \int\int_{\mathcal{X}\times\mathcal{Y}} L(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)d\boldsymbol{x}dy$ Régression
- $R(h) = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} L(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)d\boldsymbol{x}$ Classification

Fonction de perte et risque réel

Fonction de perte

Coût, critère

Quantifier l'erreur comise par h pour évaluer y

$$\begin{array}{rccc} L: & \mathcal{Y}\times\mathcal{Y} & \rightarrow & \mathbb{R}^+ \\ & (y, h(\boldsymbol{x})) & \mapsto & L(y, h(\boldsymbol{x})) \end{array}$$

0 si $y = h(\boldsymbol{x})$ et > 0 sinon

Risque réel

$$R(h) = E[L(\mathrm{Y}, h(\mathrm{X}))] = P(h(\mathrm{X}) \neq \mathrm{Y})$$

- $R(h) = \int\int_{\mathcal{X}\times\mathcal{Y}} L(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)d\boldsymbol{x}dy$ Régression
- $R(h) = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} L(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)d\boldsymbol{x}$ Classification

Fonction de perte et risque réel

Fonction de perte

Coût, critère

Quantifier l'erreur comise par h pour évaluer y

$$L : \begin{array}{ccc} \mathcal{Y} \times \mathcal{Y} & \to & \mathbb{R}^+ \\ (y, h(\boldsymbol{x})) & \mapsto & L(y, h(\boldsymbol{x})) \end{array}$$

0 si $y = h(\boldsymbol{x})$ et > 0 sinon

Risque réel

$h^\star = \operatorname{argmin}_h R(h)$

$$R(h) = E[L(\mathrm{Y}, h(\mathrm{X}))] = P(h(\mathrm{X}) \neq \mathrm{Y})$$

- $R(h) = \int\int_{\mathcal{X}\times\mathcal{Y}} L(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)d\boldsymbol{x}dy$ Régression
- $R(h) = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} L(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)d\boldsymbol{x}$ Classification

Risque empirique

Risque réel

$$R(h) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} L(y, h(\boldsymbol{x})) P(\boldsymbol{x}, y) d\boldsymbol{x}$$

Risque empirique

Risque réel

$P(\mathrm{X}, \mathrm{Y})$ et en particulier $P(\mathrm{Y}|\mathrm{X})$ inconnus

$$R(h) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} L(y, h(\boldsymbol{x})) P(\boldsymbol{x}, y) d\boldsymbol{x}$$

Risque empirique

Risque réel

$P(\mathrm{X}, \mathrm{Y})$ et en particulier $P(\mathrm{Y}|\mathrm{X})$ inconnus

$$R(h) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} L(y, h(\boldsymbol{x})) P(\boldsymbol{x}, y) d\boldsymbol{x}$$

Hypothèse de loi uniforme

Hypothèse : $P(\boldsymbol{x}_i, y_i) = 1/N$

Risque empirique

Risque réel

$P(\mathrm{X}, \mathrm{Y})$ et en particulier $P(\mathrm{Y}|\mathrm{X})$ inconnus

$$R(h) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} L(y, h(\boldsymbol{x})) P(\boldsymbol{x}, y) d\boldsymbol{x}$$

Hypothèse de loi uniforme

Hypothèse : $P(\boldsymbol{x}_i, y_i) = 1/N$

Risque : $R_{emp}(h) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, h(\boldsymbol{x}_i))$

Solution : $h^\star_S = \operatorname{argmin}_{h \in \mathcal{H}} R_{emp}(h)$

Décomposition du risque empirique

Soit $h^\star_S = \operatorname{argmin}_{h \in \mathcal{H}} R_{emp}(h)$

$$R(h^\star_S) = R(h^-) + \underbrace{[R(h^\star) - R(h^-)]}_{\text{Biais}} + \underbrace{[R(h^\star_S) - R(h^\star)]}_{\text{Variance}}$$

- Difficulté intrinsèque du problème $\rightsquigarrow$ incompressible (quantité de bruit)
- Adéquation de l'espace de recherche $\mathcal{H}$
- Erreur liée au principe de Minimisation du Risque Empirique (MRE)

Décomposition du risque empirique

Soit $h^\star_S = \operatorname{argmin}_{h \in \mathcal{H}} R_{emp}(h)$

$$R(h^\star_S) = R(h^-) + \underbrace{[R(h^\star) - R(h^-)]}_{\text{Biais}} + \underbrace{[R(h^\star_S) - R(h^\star)]}_{\text{Variance}}$$

- Difficulté intrinsèque du problème $\rightsquigarrow$ incompressible (quantité de bruit)
- Adéquation de l'espace de recherche $\mathcal{H}$
- Erreur liée au principe de Minimisation du Risque Empirique (MRE)

Décomposition du risque empirique

Soit $h^\star_S = \operatorname{argmin}_{h \in \mathcal{H}} R_{emp}(h)$

$$R(h^\star_S) = R(h^-) + \underbrace{[R(h^\star) - R(h^-)]}_{\text{Biais}} + \underbrace{[R(h^\star_S) - R(h^\star)]}_{\text{Variance}}$$

- Difficulté intrinsèque du problème $\rightsquigarrow$ incompressible (quantité de bruit)
- Adéquation de l'espace de recherche $\mathcal{H}$
- Erreur liée au principe de Minimisation du Risque Empirique (MRE)

Décomposition du risque empirique

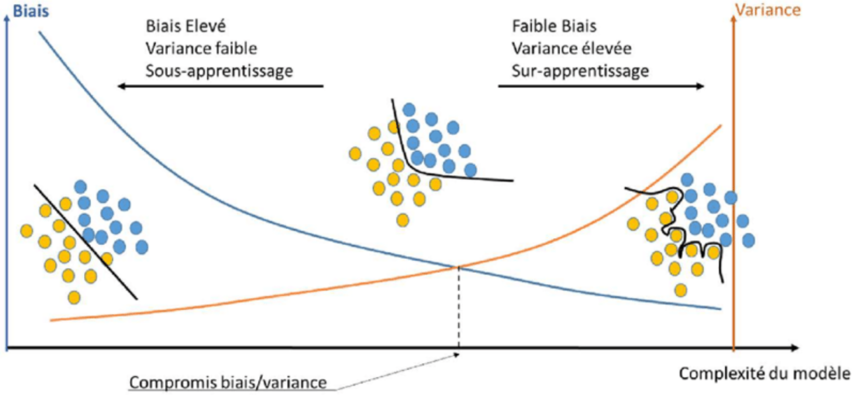
Soit $h^\star_S = \text{argmin}_{h \in \mathcal{H}} R_{emp}(h)$

$$R(h^\star_S) = R(h^-) + \underbrace{[R(h^\star) - R(h^-)]}_{\text{Biais}} + \underbrace{[R(h^\star_S) - R(h^\star)]}_{\text{Variance}}$$

- Difficulté intrinsèque du problème $\rightsquigarrow$ incompressible (quantité de bruit)
- Adéquation de l'espace de recherche $\mathcal{H}$
- Erreur liée au principe de Minimisation du Risque Empirique (MRE)

Décomposition du risque empirique

Dilemme Biais vs Variance



Crédits : Pascal Scalar – introduction au DM

Quelques questions

L'objectif est-il réalisable ?

$\rightsquigarrow$ consistance du processus d'apprentissage

Quelques questions

L'objectif est-il réalisable ?

$\rightsquigarrow$ consistance du processus d'apprentissage

MRE : processus consistant dans $\mathcal{H}$ $\qquad h^\star_S \xrightarrow[N\to\infty]{} h^\star$

Quelques questions

L'objectif est-il réalisable ?

$\rightsquigarrow$ consistance du processus d'apprentissage

MRE : processus consistant dans $\mathcal{H}$

$$h_S^\star \xrightarrow[N \to \infty]{} h^\star$$

Si oui, comment ?

- Construction d'algorithmes
- Contrôle de la vitesse de convergence

Quelques questions

L'objectif est-il réalisable ?

$\rightsquigarrow$ consistance du processus d'apprentissage

MRE : processus consistant dans $\mathcal{H}$

$$h_S^\star \xrightarrow[N\to\infty]{} h^\star$$

Si oui, comment ?

- Construction d'algorithmes
- Contrôle de la vitesse de convergence

Vers la minimisation du risque structurel

Régression polynomiale $D = 1$

Données $S = \{(t_i, x_i)\}_{i=1}^{N}$

Processus réel $t = h(x)$

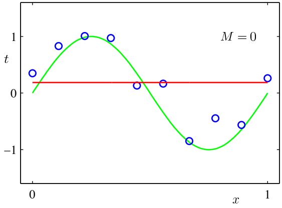
Modèle $\hat{t}_{i'} = h_m(x_{i'}) = \sum_{m=0}^{M} \alpha_m x_{i'}^{m}$

Régression polynomiale $D = 1$

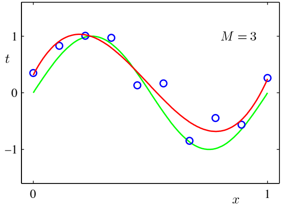
Données $S = \{(t_i, x_i)\}_{i=1}^N$

Processus réel $t = h(x)$

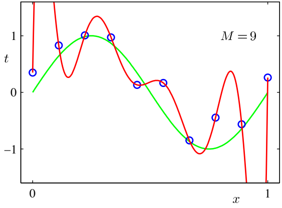
Modèle $\hat{t}_{i'} = h_m(x_{i'}) = \sum_{m=0}^{M} \alpha_m x_{i'}^m$



Sous-apprentissage



Apprentissage « correct »



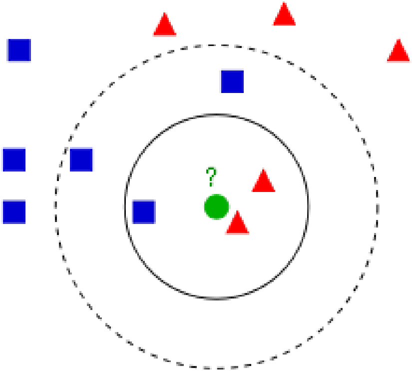
Sur-apprentissage

Exemple tiré du cours de Vincent Charvillat

Les K plus proches voisins (K-nn)

Données $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, $y_i = 1$ ou $y_i = -1$

Modèle h_K avec $K = \{3 —, 5 \text{ - - -}\}$



Exemple tiré du cours de Wikipedia

En résumé et au delà

Apprentissage supervisé

- Construire un modèle $\hat{y} = h_{\boldsymbol{p}}(\boldsymbol{x})$
- À partir d'un ensemble d'apprentissage $S_\ell = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_\ell}$
- Erreur empirique $\rightarrow$ optimiste
  Estimation imparfaite de l'erreur de généralisation

En résumé et au delà

Apprentissage supervisé

- Construire un modèle $\hat{y} = h_{\boldsymbol{p}}(\boldsymbol{x})$
- À partir d'un ensemble d'apprentissage $S_\ell = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_\ell}$
- Erreur empirique → optimiste
  Estimation imparfaite de l'erreur de généralisation

Méthodologie

- Comment évaluer les performances d'un modèle ?
- Comment choisir un bon modèle ?

En résumé et au delà

Apprentissage supervisé

- Construire un modèle $\hat{y} = h_{p}(\boldsymbol{x})$
- À partir d'un ensemble d'apprentissage $S_\ell = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_\ell}$
- Erreur empirique → optimiste
  Estimation imparfaite de l'erreur de généralisation

Méthodologie

- Comment évaluer les performances d'un modèle ?
- Comment choisir un bon modèle ?

⇝ vignette WikiStat