Statistiques descriptives Analyse de données Algorithmes

(7) APPRENTISSAGE

Jérôme Lacaille
Expert Émérite Safran

APPRENTISSAGE

1. **ESTIMATION**
2. **BIAIS ET VARIANCE**
3. **OPTIMISATION EMPIRIQUE**
4. **ERREUR DE GÉNÉRALISATION**
5. **CAPACITÉ DE GÉNÉRALISATION**

→ **ROBUSTESSE**

LA MALÉDICTION DES GRANDES DIMENSIONS – EN DEUX REMARQUES

1. **Plus le nombre de dimensions d'un espace est grand, plus les points sont éloignés les uns des autres.**
   - Par exemple en dimension d deux points tirés aléatoirement suivant une loi gaussienne ont une distance au carré dont la loi est un $\chi^2(d)$ de moyenne égale à d.
   - Ainsi la distance moyenne ente deux points tirés au hasard en dimension d est d'ordre $\sqrt{d}$ qui augmente avec la dimension.
   - Il faut donc de plus en plus de points pour décrire ce qui se passe dans un tel espace.

2. **Le volume d'une boule est de plus en plus concentré sur sa surface quand la dimension augmente.**
   - En effet, soit deux boules de rayons respectifs $r_1 < r_2$, le volume d'une sphère en dimension d est $V_d(r) = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma\left(\frac{n}{2}+1\right)}$
     - donc $\frac{(V_d(r_2)-V_d(r_1))}{V_d(r_2)} = 1 - \left(\frac{r_1}{r_2}\right)^d \xrightarrow{d\to\infty} 1.$
   - Ainsi tous les points sont à équidistance du centre.
   - La notion de distance euclidienne perd son sens habituel.

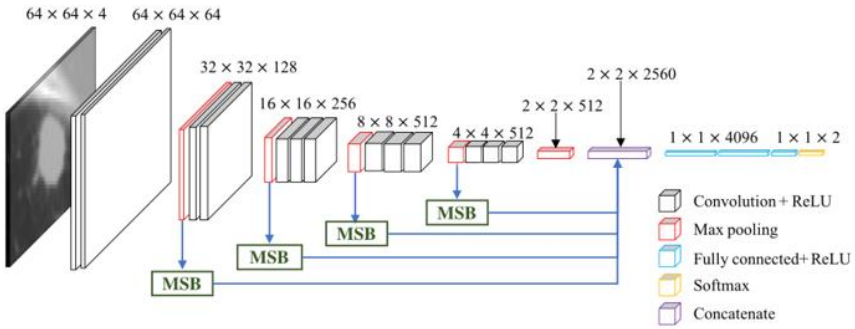
POURQUOI RÉDUIRE LA DIMENSION ?

Il existe deux éléments qui sont très liées :

- le nombre de facteurs utilisés pour réaliser une prédiction
- le nombre de paramètres de la fonction de prédiction.

Si le nombre de facteurs augmente, il est logique de s'attendre aussi à une augmentation du nombre de paramètres

- Si par exemple on observe l'architecture d'un réseau de neurone on s'aperçoit que plus le nombre de cellules intermédiaires (dites cachées) augmente, plus le nombre de paramètres va aussi augmenter.
- Pourtant certains réseaux de neurones très gros généralisent bien.



Derrière ce phénomène se cache une notion de « complexité » du modèle utilisé.

- La complexité d'un modèle est mesurée par sa capacité de généralisation ou **dimension de Vapnik-Chervonenkis (VC-dimension)**
- C'est la capacité d'un modèle à apprendre des règles plutôt que mémoriser « par cœur » un jeu de données spécifique.

ESTIMATION PAR UN APPRENTISSAGE SUPERVISÉ

L'apprentissage est un algorithme qui permet de sélectionner parmi un ensemble de fonctions $\mathcal{F}$ un élément $F \in \mathcal{F}$ qui estime au mieux une relation $x \longmapsto y$.

La relation n'est a priori pas connue. Il existe deux variables aléatoires X et Y liées entre elles et c'est cette liaison que l'on cherche à modéliser.

- $X : \Omega \longrightarrow \mathbb{R}^d$
  - Les difficultés principales de l'estimation proviennent du fait que la dimension p des entrées est en général très grande.
- $Y : \Omega \longrightarrow \mathbb{R}^r$ ou $Y : \Omega \longrightarrow \{a, b, ..\}$ un ensemble de catégories ordonnées ou non.
  - La classification est un cas particulier de l'estimation avec l'espace d'arrivée discret.
  - En général on prendra $r = 1$, l'estimation multivariée se généralise facilement.

La relation entre X et Y n'est cependant connue qu'à travers un ensemble d'exemples :

$$\left\{\left(\boldsymbol{x^{(1)}}, \boldsymbol{y^{(1)}}\right), \ldots, \left(\boldsymbol{x^{(N)}}, \boldsymbol{y^{(N)}}\right)\right\}$$

L'ensemble de fonctions dans lesquels choisir la relation est paramétré

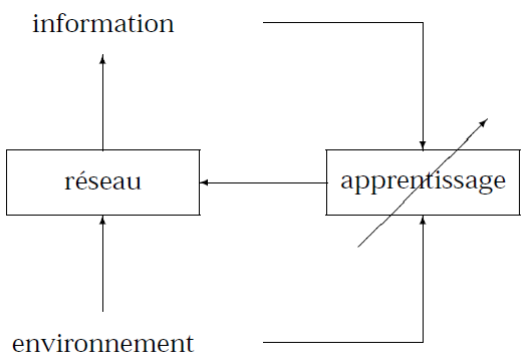
$$\boldsymbol{\mathcal{F} = \{F_w \mid w \in \mathbb{R}^p\}}$$

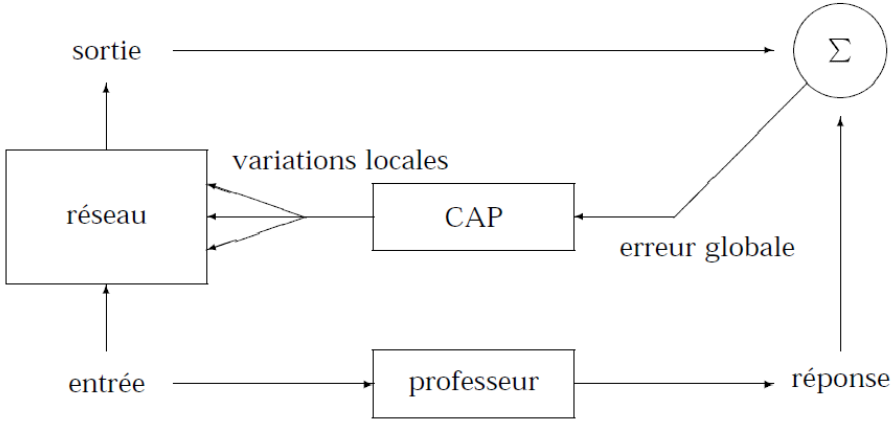
- La question principale est de trouver un « bon » $\boldsymbol{w_0}$: s'assurer que la fonction obtenue « généralise ».

CRÉDIT ASSIGNMENT PROBLEM

L'apprentissage du modèle consiste à mettre à jour des poids $(w_{i,j})$.

- Un algorithme de supervision compare la sortie obtenue à la sortie souhaitée, en déduit une erreur. La difficulté est ensuite de répartir cette erreur sur les différents poids du réseau (c'est le *Credit Assignment Problem* de Minsky).





BIAIS ET VARIANCE (1)

On dispose d'une base d'apprentissage D formée de n couples d'entrées et de réponses associées :

$$D = \left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$$

Soit $\hat{y} = F_w(x)$ une fonction d'estimation paramétrée par le vecteur w.

- Par exemple w peut être les paramètres d'une régression ou les poids d'un réseaux de neurones.

Notons $e(X, Y)$, l'erreur d'estimation :

$$e(X, Y) = Y - \hat{Y} = Y - F_w(X)$$

Le risque moyen $R_0(w)$ est l'espérance du carré de cette erreur sur la distribution réelle des entrées (X, Y).

$$R_0(w) = E\left[e(X, Y)^2\right] = E[Y - F_w(X)]^2$$

C'est bien entendu la fonction que l'on cherche à minimiser en w.

BIAIS ET VARIANCE (2)

On peut réécrire R_0 en faisant intervenir la projection $g(X) = E[Y|X]$

$$\begin{aligned} \boldsymbol{R_0(w)} &= \boldsymbol{E[Y - g(X) + g(X) - F_w(X)]^2} \\ &= \boldsymbol{E[Y - g(X)]^2 + 2E\big[\big(Y - g(X)\big)\big(g(X) - F_x(X)\big)\big] + E[g(X) - F_w(X)]^2} \end{aligned}$$

- Or $\boldsymbol{\epsilon = Y - g(X) \perp X}$, donc $\boldsymbol{E[\epsilon g(X)] = 0}$ et $\boldsymbol{E[\epsilon F_w(X)] = 0}$

Finalement

$$\boldsymbol{R_0(w) = E\big[\epsilon^2\big] + E[g(X) - F_w(X)]^2}$$

- Le premier terme est indépendant de w, donc la minimisation du risque moyen R_0 revient à minimiser l'écart à la projection sur l'espace engendré par le vecteur X.

- Si $w_0 = \arg\min_w R_0(w)$, alors la fonction F_{w_0} approxime au mieux $x \mapsto E[y|x]$.
- $R_0 \geq \boldsymbol{E\big[\epsilon^2\big]}$ donc on ne peut pas faire mieux que $E[y|x]$.

RISQUE EMPIRIQUE (1)

Notons R_D la moyenne des carrés des erreurs sur l'ensemble d'exemples D, c'est le risque empirique :

$$R_D(w) = E_D[e^2] = \frac{1}{n}\sum_{i=1}^{n} e^2\left(x^{(i)}, y^{(i)}\right)$$

- L'apprentissage sur le jeu d'exemples D revient à minimiser ce risque empirique et conduit à trouver un paramètre w_D qui estime au mieux $E_D[y|x]$.
- Soit la fonction $D \mapsto w_D$ qui à un ensemble d'exemples D associe le paramètre w_D minimisant le risque empirique, on peut voir la fonction d'estimation comme une fonction de l'ensemble d'exemples plutôt que du paramètre w.

$$F_D(x) = F_{w_D}(x)$$

On va étudier le comportement de cette fonction sur l'ensemble des expériences possibles $\mathcal{D} = \{D|D \in \mathcal{P}(X(\Omega))\}$ et on notera E_* la moyenne suivant toutes ces expériences.

- Pour x fixé :

$$E_*[g(x) - F_D(x)]^2 = E_*[g(x) - E_*F_D(x) + E_*F_D(x) - F_D(x)]^2$$

$$= \underbrace{E_*[g(x) - E_*F_D(x)]^2}_{\text{Biais}} + \underbrace{E_*[E_*F_D(x) - F_D(x)]^2}_{\text{Variance}} + \cancel{E_*\left(g(x) - E_*F_D(x)\right)\left(E_*F_D(x) - F_D(x)\right)}$$

Ne dépend pas de D — *De moyenne nulle*

SAFRAN Snecma

RISQUE EMPIRIQUE (2)

- Pour toute observation $\boldsymbol{x}$:

$$\boldsymbol{E_*[g(x) - F_D(x)]^2 = [g(x) - E_* F_D(x)]^2 + E_*[E_* F_D(x) - F_D(x)]^2}$$

Biais — Variance

La fonction $\boldsymbol{D \longmapsto F_D(x)}$ s'interprète comme la réponse à l'entrée $\boldsymbol{x}$ d'un estimateur dont l'apprentissage a été basé sur le lot $\boldsymbol{D}$.

1. L'estimateur est sans biais si $\boldsymbol{E_* F_D(x) = g(x)}$, c'est-à-dire si l'architecture des fonctions $\boldsymbol{F_w}$ permet de réaliser en moyenne l'espérance conditionnelle.
2. Plus le modèle de fonctions sera complexe, plus on aura de chance que la fonction $\boldsymbol{g}$ puisse être modélisée correctement, le biais va donc diminuer.
3. Cependant, plus le modèle est complexe, plus les résultats des apprentissages seront divers et donc la variance va augmenter.
4. Il est donc important de réduire la complexité du modèle, par exemple en utilisant la régularité des données pour construire des estimateurs plus simples.

GÉNÉRALISATION (1)

Hypothèses :

- x et y sont liés par une fonction inconnue $y = g(x)$.
- $P(x)$ est la probabilité que l'environnement génère l'état x.
- $P(x, y) = P(y|x)P(x)$ est la probabilité conjointe d'observer un couple exemple (x, y) que l'on utilisera pendant l'apprentissage.
- $P(y|x)$ est la relation que l'on cherche à estimer par une fonction $y = F_w(x)$.

Notations :

- $R_0(w) = \int e\left(y, F_w(x)\right)^2 dP(x, y)$ est le risque réel.
- $R_D(w) = \frac{1}{n}\sum_{i=1}^{n} e\left(y^{(i)}, F_w\left(x^{(i)}\right)\right)^2$ est le risque empirique sur le jeu d'exemples D.

- $w_0 = \arg\min_w R_0(w)$ est inconnu.
- $w_D = \arg\min_w R_D(w)$ est le résultat de l'apprentissage.

Problème :

Sous quelles condition F_{w_D} est proche de F_{w_0} ?

- En fait, nous ne sommes pas intéressés par la proximité des paramètres.

CONVERGENCE PRESQUE SURE

$$Z_w = e\left(Y, F_w(X)\right)^2$$

Comme (X, Y) est un vecteur aléatoire de loi $P(x, y)$ et que e^2 est mesurable, Z_w est aussi une variable aléatoire et :

- $R_D(w)$ est la moyenne empirique de Z_w pour le jeu de données exemples D.
- $R_0(w)$ est l'espérance de Z_w sous la loi P.

La loi forte des grands nombres nous assure que pour un w donné :

$$R_D(w) \xrightarrow[n\to\infty]{ps} R_0(w)$$

- Cela justifie l'utilisation du risque empirique, mais rien n'assure que le vecteur des poids w_D qui minimise R_D minimise aussi R_0

CONVERGENCE UNIFORME

Proposition :

- Si on peut assurer la convergence uniforme en probabilité par rapport à w de R_D vers R_0, alors $R_0(w_D)$ converge vers la plus petite valeur possible pour le risque : $R_0(w_0)$.

- Preuve :

  Exprimons la convergence uniforme de R_D vers R_0 :

  $$\forall \epsilon > 0, \qquad P\{sup_w |R_0(w) - R_D(w)| \geq \epsilon\} \xrightarrow[n \to +\infty]{} 0$$

  Alors

  $$\forall \alpha > 0, \forall \epsilon > 0, \exists D \in \mathcal{D}, P\{sup_w |R_0(w) - R_D(w)| \geq \epsilon\} \leq \alpha$$

  Dans ce cas, avec la probabilité $1 - \alpha$

  $$\begin{cases} R_0(w_D) - R_D(w_D) & < & \epsilon \\ R_D(w_0) - R_0(w_0) & < & \epsilon \end{cases}$$

  Or par définition $R_D(w_D) \leq R_D(w_0)$ donc en sommant membre à membre les deux équations précédentes on obtient :

  $$R_0(w_D) - R_0(w_0) < 2\epsilon$$

  Soit

  $$\forall \alpha > 0, \forall \epsilon > 0, \exists D \in \mathcal{D}, P\{|R_0(w_D) - R_0(w_0)| \geq 2\epsilon\} \leq \alpha$$

GRANDES DÉVIATIONS

Inégalité de Hoeffding :

- Soient $(Z_i)_{i=1\dots n}$ des variables réelles i.i.d. (indépendantes et identiquement distribuées)
- Bornées : $\forall i,\ Z_i \in [a,b]$
- De moyenne $\mu = E(Z_i)$

- Alors, si $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$ est la moyenne empirique,

$$\forall \epsilon > 0, \qquad P(|\bar{\mu}_n - \mu| \geq \epsilon) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

BORNE DE L'ERREUR DE GÉNÉRALISATION

Hypothèses :

- On suppose que le risque est borné, $R_0(w) \in [0,1]$
  - Par exemple si $|e(y,x)| \leq 1$.
- On considère des ensembles de fonctions finis : $\text{card}\,\mathcal{F} = |\mathcal{F}| < +\infty$
  - Si $\mathcal{F} = \{F_w | \, w \in \mathbb{R}^p\}$, il suffit de s'assurer que le nombre de fonctions produites ainsi est fini ou par exemple de quantifier l'ensemble des w possibles.

Théorème :

$$P\left\{\sup_w |R_0(w) - R_D(w)| \leq \epsilon\right\} \geq 1 - \delta \quad si \quad \epsilon \geq \sqrt{\frac{\log|\mathcal{F}| + \log(2/\delta)}{2n}}$$

Remarques :

- Si on quantifie chaque paramètre sur q valeurs, alors $|\mathcal{F}| \sim q^p$ (c'est l'entropie de Shannon).
  - Pour avoir une erreur de l'ordre de ϵ, il faut que $q \sim d/\epsilon$ (intervalles de taille ϵ/m) (Pythagore).
- Il faut que $\frac{\log|\mathcal{F}|}{n} \ll \epsilon^2$ soit $\frac{p}{n} \ll \epsilon^2$ et donc que le nombre d'exemples (n) soit très très grand devant le nombre de paramètres (p).

DÉMONSTRATION DE LA BORNE

On fait une majoration très brutale

$$
\begin{aligned}
P\left\{\sup_w |R_0(w) - R_D(w)| \geq \epsilon\right\} &\leq \sum_w P\{|R_0(w) - R_D(w)| \geq \epsilon\} \\
&\leq \sum_w 2\exp(-2n\epsilon^2) \\
&\leq 2|\mathcal{F}|e^{-2n\epsilon^2}
\end{aligned}
$$

- La seconde inégalité est une application de Hoeffding avec $a = 0$ et $b = 1$.

- Le résultat est obtenu si on pose $\delta \geq 2|\mathcal{F}|e^{-2n\epsilon^2}$ soit :

$$\epsilon^2 \geq \frac{\log(2|\mathcal{F}|/\delta)}{2n}$$

MALÉDICTION DES GRANDES DIMENSIONS

Corolaire

S'il existe une constante C, et un nombre β tels que

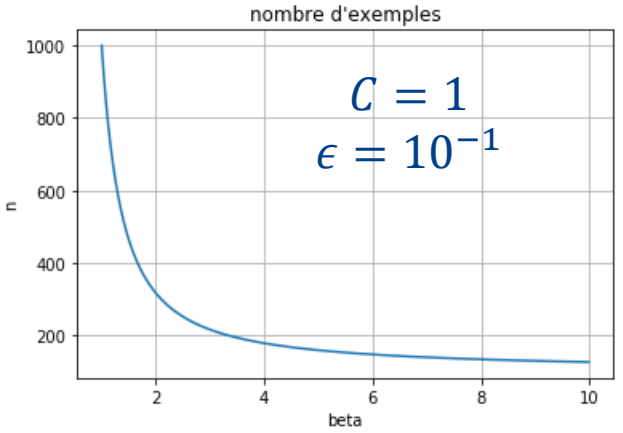
$$R_0(w_0) \leq C\ (\log|\mathcal{F}|)^{-\beta}$$

Alors $\forall \epsilon > 0, \forall n,$ on pose $\log|\mathcal{F}| = \log(2/\delta) = n\epsilon^2$

$$P(R_0(w_D) \leq 3\epsilon) \geq 1 - 2e^{-C^{\frac{1}{\beta}}\epsilon^{-\frac{1}{\beta}}}$$

Si

$$n \geq C^{\frac{1}{\beta}}\epsilon^{-2-\frac{1}{\beta}}$$



(Application du théorème précédent en choisissant $log|\mathcal{F}| = \log(2/\delta) = n\epsilon^2$)

- Plus l'erreur de généralisation décroit vite avec la complexité de l'ensemble de fonctions (plus β est grand) moins on a besoin d'exemples pour assurer une erreur d'approximation petite.

Cas des fonction Lipchitziennes

- Si la fonctions F à estimer par les F_w est uniformément $\alpha -$ lipchitzienne de $\mathbb{R}^d \longrightarrow \mathbb{R}$
  - $\exists C > 0$ et un polynôme $Q_\alpha(x)$ de degré inférieur à α tel que
    $$\forall (x, x'), |F(x) - Q_\alpha(x')| \leq C\|x - x'\|^\alpha$$

- Alors si $R_0(w_0) \leq C\ (\log|\mathcal{F}|)^{-\beta}$
  $$\beta \leq \alpha/d$$

DÉMONSTRATION PRÉCÉDENTE

Majoration par l'inégalité de Hoeffding avec probabilité $1-\delta$

$$R_0(w_D) - R_D(w_D) \leq \epsilon \text{ et } R_D(w_0) - R_0(w_0) \leq \epsilon$$

En sommant les deux équations :

$$R_0(w_D) \leq 2\epsilon + R_0(w_0) + \left(R_D(w_D) - R_D(w_0)\right) \leq 2\epsilon + R_0(w_0)$$

Car le terme de droite est négatif : $R_D(w_D) - R_D(w_0) \leq 0$

Il suffit donc d'avoir

$$R_0(w_0) \leq C\ (\log|\mathcal{F}|)^{-\beta} \leq \epsilon$$

Soit

$$(\log|\mathcal{F}|)^{-\beta} \leq \frac{\epsilon}{C} \Rightarrow \log|\mathcal{F}| \geq \left(\frac{C}{\epsilon}\right)^{\frac{1}{\beta}}$$

On choisit de prendre n, ϵ et δ tels que $\log|\mathcal{F}| = \log(2/\delta) = n\epsilon^2$, ce qui vérifie la condition du théorème. Ainsi

$$\log\left(\frac{2}{\delta}\right) = \log|\mathcal{F}| \geq \left(\frac{C}{\epsilon}\right)^{\frac{1}{\beta}}$$

$$\delta \leq 2e^{-\left(\frac{C}{\epsilon}\right)^{\frac{1}{\beta}}} \quad \text{d'où} \quad 1-\delta \geq 1-2e^{-\left(\frac{C}{\epsilon}\right)^{\frac{1}{\beta}}}$$

SIMPLIFICATION : UNE CLASSIFICATION BINAIRE

Pour simplifier l'étude on considère un cadre binaire : on suppose que $y \in \{0,1\}$.

Soit

$$\mathcal{F} = \left\{ F_w, \quad F_w : \mathbb{R}^p \to \{0,1\} \right\}$$

une famille de classifications binaires réalisées par un réseau F.

Posons

$$\mathcal{S} = \left\{ x^{(1)}, \ldots x^{(N)} \mid x^{(i)} \in \mathbb{X} \right\}$$

Alors, pour F fixé (w fixé), le réseau définitune partition binaire de $\mathcal{S}$ (une dichotomie).

$$\begin{aligned} \mathcal{S}_0(F) &= \left\{ x \in \mathcal{S} \mid F(x) = 0 \right\} \\ \mathcal{S}_1(F) &= \left\{ x \in \mathcal{S} \mid F(x) = 1 \right\}. \end{aligned}$$

Notons $\Delta_{\mathcal{F}}(\mathcal{S})$ le nombre de dichotomies de $\mathcal{S}$ réalisées par la famille de fonctions $\mathcal{F}$.

$$\Delta_{\mathcal{F}}(\mathcal{S}) = \#\left\{ \{\mathcal{S}_0(F), \mathcal{S}_1(F)\} \mid F \in \mathcal{F} \right\}.$$

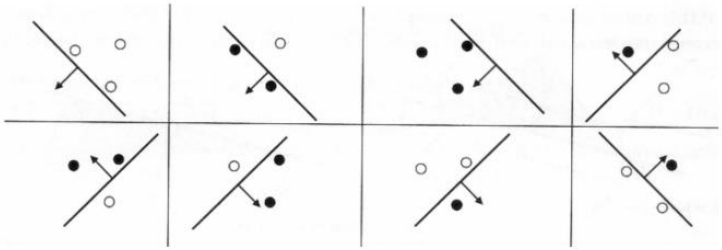
Notons aussi $\Delta_{\mathcal{F}}(N)$ le maximum du nombre précédent pour tous les jeux d'exemples de taille N.

$$\Delta_{\mathcal{F}}(N) = \max_{\mathcal{S} \mid \#\mathcal{S} = N} \Delta_{\mathcal{F}}(\mathcal{S}).$$

DIMENSION DE VAPNIK

DEFINITION 2 *On dit que la famille* $\mathcal{S}$ *est* éclatée *par* $\mathcal{F}$ *si*

$$\Delta_{\mathcal{F}}(\mathcal{S}) = 2^{\#\mathcal{S}}$$



En dimension 2 trois points sont « éclatables » par un modèle linéaire

DEFINITION *La dimension de Vapnik* h *de* $\mathcal{F}$ *est le cardinal maximum d'un ensemble* $\mathcal{S}$ *qui peut être éclaté par* $\mathcal{F}$ *:*

$$\begin{aligned} h &= \max\left\{\#\mathcal{S} \mid \mathcal{S} \subset \mathbb{X} \text{ et } \Delta_{\mathcal{F}}(\mathcal{S}) = 2^{\#\mathcal{S}}\right\} \\ &= \max\left\{N \mid \Delta_{\mathcal{F}}(N) = 2^{N}\right\} \end{aligned}$$

THÉORÈME **(Vapnik)** $\forall \epsilon > 0$, *on a*

$$\mathrm{P}\left\{\sup_{w} |\lambda_0(w) - \lambda_{\mathcal{D}}(w)| > \epsilon\right\} < \left(\frac{2eN}{h}\right)^{h} \exp(-\epsilon^2 N)$$

$$\mathrm{P}\left\{\sup_{w} \frac{|\lambda_0(w) - \lambda_{\mathcal{D}}(w)|}{\sqrt{\lambda_0(w)}} > \epsilon\right\} < \left(\frac{2eN}{h}\right)^{h} \exp\left(-\frac{\epsilon^2 N}{4}\right).$$

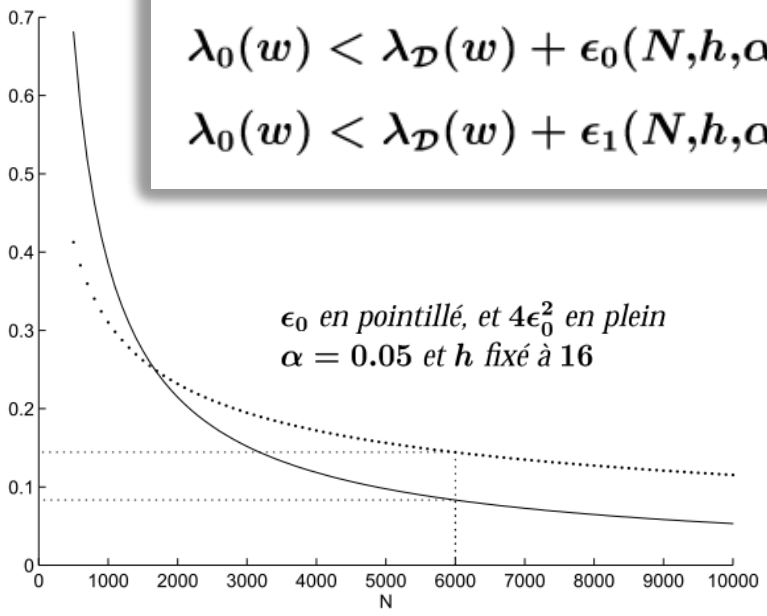
ERREUR DE GÉNÉRALISATION

$$\alpha = \left(\frac{2eN}{h}\right)^h \exp(-\epsilon_0^2 N)$$

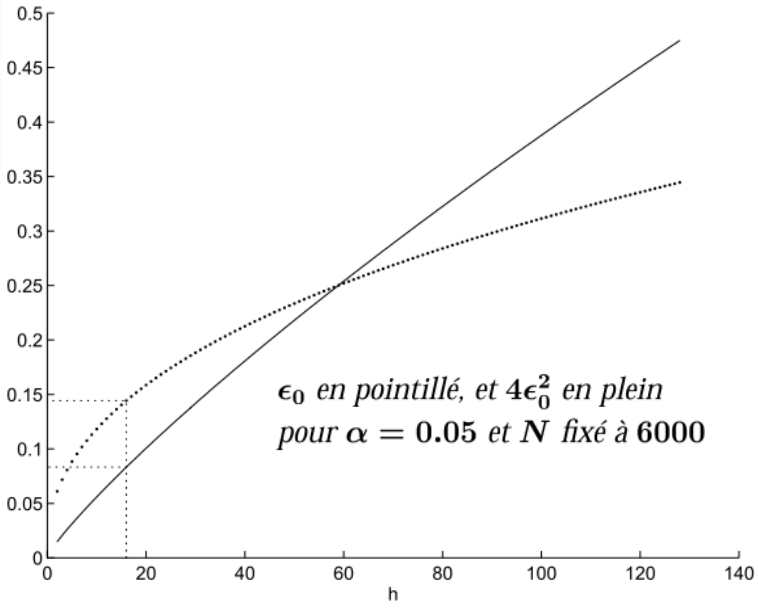
$$\begin{cases} \epsilon_0(N,h,\alpha) = \sqrt{\dfrac{h}{N}\left[\log\dfrac{2N}{h}+1\right] - \dfrac{1}{N}\log\alpha} \\ \epsilon_1(N,h,\alpha,\lambda_{\mathcal{D}}(w)) = 2\epsilon_0^2(N,h,\alpha)\left[1+\sqrt{1+\dfrac{\lambda_{\mathcal{D}}(w)}{\epsilon_0^2(N,h,\alpha)}}\right] \end{cases}$$

Avec une probabilité $1-\alpha$, il existe $\mathcal{D}$ assez grand tel que :

$$\lambda_0(w) < \lambda_{\mathcal{D}}(w) + \epsilon_0(N,h,\alpha)$$
$$\lambda_0(w) < \lambda_{\mathcal{D}}(w) + \epsilon_1(N,h,\alpha,\lambda_{\mathcal{D}}(w))$$



ϵ_0 *en pointillé, et* $4\epsilon_0^2$ *en plein*
$\alpha = 0.05$ *et* h *fixé à* 16



ϵ_0 *en pointillé, et* $4\epsilon_0^2$ *en plein*
pour $\alpha = 0.05$ *et* N *fixé à* 6000

ROBUSTESSE

1. **BAGGING**
2. **VALIDATION CROISÉE**

BAGGING

Bagging = Bootstrap Averaging (Leo Breiman, 1994 - 1996).

C'est une méthode ensembliste. Elle s'applique aux arbres de décision, mais aussi à toute autre méthode de régression ou de classification.

- On tire avec remise des échantillons $D_1, \dots, D_B$ dans D.
  - Si la taille des D_j est la même que celle de D on parle de **bootstrap**.
- On crée un modèle $Y = f_j(X), j = 1 \dots B$, par apprentissage sur chaque échantillon.
- Pour l'estimation :
  - On choisit la moyenne des résultats pour une régression

$$\hat{y} = \frac{1}{B} \sum_j f_j(x)$$

  - On vote la catégorie majoritaire pour une classification

$$\hat{y} = \arg\max_k \sum_{j=1}^{B} \mathbb{I}_{f_j(x)=k}$$

VALIDATION CROISÉE

Deux méthodes sont usuellement utilisées pour calculer l'erreur de généralisation d'une régression ou d'une classification.

- K-fold : on partitionne de manière déterministe D en K parties disjointes (folds).
  - Pour $k = 1 \dots K$, on crée le sous échantillon D_k formé de la réunion de toutes les parties sauf la $k^{ième}$.
    - On apprend l'estimateur sur S_k.
    - On calcule l'erreur e_k sur $S \backslash S_k$.
  - On estime la moyenne des erreurs dont on peut obtenir un intervalle de confiance par approximation gaussienne.
    - $\bar{e} = \frac{1}{K} \sum_k e_k$.
- Leave-one-out : c'est la même méthode que K-fold, mais avec $K = N$. Donc on implémente un estimateur sur toutes les observations sauf une que l'on va tester sur celle que l'on a mis de coté.

La plupart des heuristiques d'apprentissage d'arbres de décision limitent la croissance des arbres par validation croisée.

A SUIVRE

MACHINES DE BOLTZMANN