La théorie statistique de l'apprentissage
ou
l'approche de Vapnik

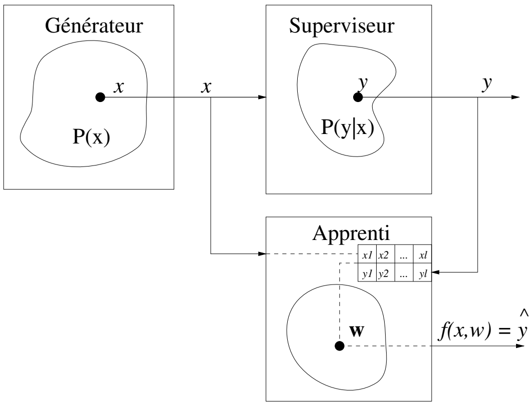
- V.N. Vapnik. *The Nature of Statistical Learning Theory, Springer Series in Statistics.* Springer-Verlag, 1995.
- K. Kotz and N.L Johnson. *Encyclopedia of Statistical Sciences*, John Wiley & sons, 1985–1989.
- V.N. Vapnik. *Estimation of Dependences Based on Empirical Data, Springer Series in Statistics.* Springer-Verlag, 1982.

Plan

1. Le problème de l'apprentissage fonctionnel
2. Consistance des processus d'apprentissage
3. Bornes sur les vitesses de convergence des processus d'apprentissage
4. Contrôle de la capacité de généralisation d'un processus d'apprentissage

Les 3 composantes du modèle d'apprentissage

1. Un **générateur** de vect.a $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, iid $\sim P(\boldsymbol{x})$ (inconnu)
2. Un **superviseur** de v.a $y \in \mathcal{Y} \subseteq \mathbb{R}$, iid $\sim P(y|\boldsymbol{x})$ (inconnu)
3. Un **apprenti** capable d'implémenter $f(\boldsymbol{x}, \boldsymbol{w}) \in \mathcal{H}$



Exemples d'apprentis

- Décompositions en séries
  - Polynômes de degré K : $\mathcal{H} = \{f | f(x) = \sum_{k=0}^{K} a_k x^k\}$, $\boldsymbol{w} = \boldsymbol{a}$
  - Séries de Fourier tronquées :

$$\mathcal{H} = \{f | f(x) = a_0 + \sum_{k=1}^{K} a_k \sin(kx) + b_k \cos(kx)\}, \ \boldsymbol{w} = (\boldsymbol{a}^T \boldsymbol{b}^T)^T$$

- Splines : $\mathcal{H} = \{f \in L^2 | f' \in L^2, \|f'\|^2 \leq c\}$, $\boldsymbol{w} = c$
- Nadaraya-Watson : $f(\boldsymbol{x}) = \dfrac{\sum_{i=1}^{\ell} y_i K_\sigma(\boldsymbol{x} - \boldsymbol{x}_i)}{\sum_{i=1}^{\ell} K_\sigma(\boldsymbol{x} - \boldsymbol{x}_i)}$, $\boldsymbol{w} = \sigma$
- MLP, RBF, ... : $f(\boldsymbol{x}) = \sum_{k=1}^{K} c_k g(\boldsymbol{x}, \boldsymbol{v}_k)$, $\boldsymbol{w} = (\boldsymbol{c}^T \boldsymbol{v}_1^T \ldots \boldsymbol{v}_\ell^T)$

Objectif de l'apprenti

Sélectionner $f(\boldsymbol{x}, \boldsymbol{w}) \in \mathcal{H}$ qui approche **au mieux** la réponse du superviseur.

A partir d'un échantillon $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$, $(\boldsymbol{x}_i, y_i)$ iid $\sim P(\boldsymbol{x}, y)$

Au mieux : défini par une fonction de perte $L(y, f(\boldsymbol{x}, \boldsymbol{w}))$ entre les réponses du superviseur y, et les réponses de l'apprenti $f(\boldsymbol{x}, \boldsymbol{w})$.

Espérance de perte = risque = $R(\boldsymbol{w}) = \displaystyle\int L(y, f(\boldsymbol{x}, \boldsymbol{w}))P(\boldsymbol{x}, y)d\boldsymbol{x}dy$

Au mieux = minimisant le risque $R(\boldsymbol{w})$ sur $\mathcal{H}$

$\Rightarrow$ rôle central de L

Exemples de fonctions de perte

coût quadratique	$L = (y - f(\boldsymbol{x}))^2$	$\rightarrow f^*(\boldsymbol{x}) = \mathbb{E}(y \vert \boldsymbol{x})$
coût absolu	$L = \vert y - f(\boldsymbol{x}) \vert$	$\rightarrow \mathbb{P}(y \leq f^*(\boldsymbol{x}) \vert \boldsymbol{x}) = 1/2$
coût 0/1	$\begin{cases} L = 0 \text{ si } y = f(\boldsymbol{x}) \\ L = 1 \text{ si } y \neq f(\boldsymbol{x}) \end{cases}$	$\rightarrow f^*(\boldsymbol{x}) = \mathbb{I}_{\mathbb{P}(y=1 \vert \boldsymbol{x}) > \frac{1}{2}}$
Entropie croisée	$L = -y \ln(f(\boldsymbol{x})) - (1-y)\ln(1 - f(\boldsymbol{x}))$	$\rightarrow f^*(\boldsymbol{x}) = \mathbb{P}(y = 1 \vert \boldsymbol{x})$

HEUDIASYC UMR 6599 - Université de Technologie de Compiègne CNRS Centre National de la Recherche Scientifique

Le principe de la minimisation du risque empirique (MRE)

On veut minimiser $R(\boldsymbol{w}) = \int L(y, f(\boldsymbol{x}, \boldsymbol{w}))P(\boldsymbol{x}, y)d\boldsymbol{x}dy$

Mais $P(\boldsymbol{x}, y) = ??$

Heuristique : minimisation de $R_{\text{emp}}(\boldsymbol{w}) = \frac{1}{\ell}\sum_{i=1}^{\ell} L(y_i, f(\boldsymbol{x}_i, \boldsymbol{w}))$

Pour le théoricien

principe d'induction par MRE pas justifiable *per se*

même si MRE = maximisation de la vraisemblance.

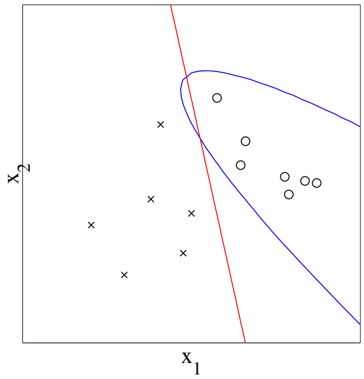
L'application de la MRE pour choisir $\boldsymbol{w}$ défini un processus d'apprentissage, qui est l'objet d'étude

MRE et maximum de vraisemblance

2 classes gaussiennes, matrices de covariance différentes.

Taux d'erreur : MV : 25%

Séparation linéaire : 17%



14 pts en dimension 2 $\simeq$ 350 pts en dim 10 $\simeq$ 1400 pts en dim 20

Questions

- L'objectif est-il réalisable ?
  - Conditions de consistance du processus d'apprentissage
  - Vitesse de convergence
- Et si oui, comment ?
  - Contrôle de la vitesse de convergence
  - Construction d'algorithmes d'apprentissage

Consistance des processus d'apprentissage

Définition : Un estimateur est dit consistant si il converge en probabilité vers la valeur estimée quand $\ell \to \infty$.

Quel intéret pour l'apprentissage à échantillon fini ?

- Logique : Est-ce que l'apprenti peut réaliser son objectif (min. de R) en utilisant la stratégie donnée (min. de R_{emp}) ?

  Cas le plus favorable : il a vu une infinité de couples $(\boldsymbol{x}_i, y_i)$.

  Si il y a échec, il faut modifier soit l'apprenti, soit l'objectif.
- Historique : la consistance est un concept apprécié des statisticiens. Ils n'accepteront pas une théorie qui ne la garantisse pas.
- Vitesse de convergence à échantillon fini.

Qu'est-ce qu'un processus d'apprentissage consistant ?

- Le principe de la MRE est consistant pour les objectifs donnés et pour le probléme posé, si l'apprenti fait du mieux possible (dans la mesure de ses faibles moyens) quand la taille de l'échantillon tend vers l'infini.
- Soit $\{\boldsymbol{w}_\ell\}_{i=1}^{\infty}$ la suite des vect.a $\boldsymbol{w}$ pour lesquels le minimum de R_{emp} est atteint. Le principe de la MRE est consistant pour la fonction de perte L et pour la distrution $P(\boldsymbol{x}, y)$, si les suites $R_{\text{emp}}(\boldsymbol{w}_\ell)$ et $R(\boldsymbol{w}_\ell)$ convergent en probabilité vers la valeur minimale de R sur $f(\boldsymbol{x}, \boldsymbol{w}) \in \mathcal{H}$.

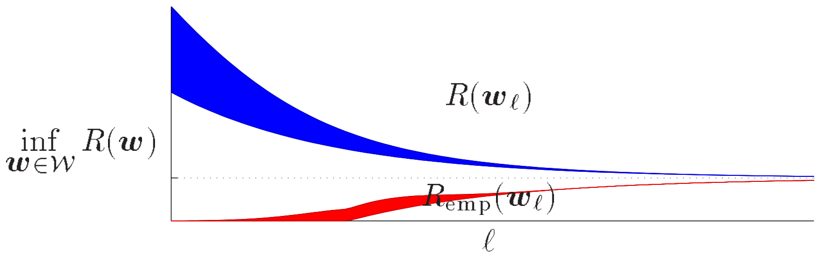
On retrouve le rôle central de L

Consistance (2)

$$R(\boldsymbol{w}_\ell) \xrightarrow{P} \inf_{\boldsymbol{w}\in\mathcal{W}} R(\boldsymbol{w}) \qquad (1)$$

$$R_{\text{emp}}(\boldsymbol{w}_\ell) \xrightarrow{P} \inf_{\boldsymbol{w}\in\mathcal{W}} R(\boldsymbol{w}) \qquad (2)$$

$$(1) \Leftrightarrow \forall\varepsilon, \forall\eta, \exists\ell_0 \;/\; \ell > \ell_0 \Rightarrow \mathbb{P}\left(|R(\boldsymbol{w}_\ell) - \inf_{\boldsymbol{w}\in\mathcal{W}} R(\boldsymbol{w})| > \varepsilon\right) < \eta$$



Le théorème de la théorie statistique de l'apprentissage (Vapnik & Chervonenkis 1989)

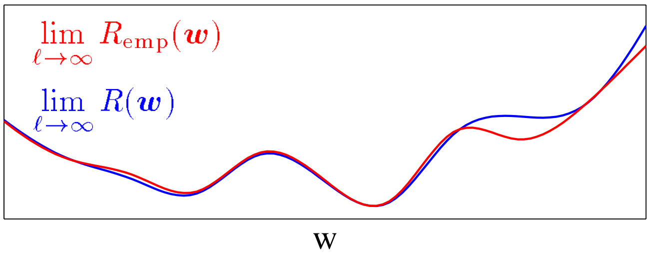
Si $A \leq R(\boldsymbol{w}) \leq B$,

alors le principe de la MRE est consistant si et seulement si R_{emp} converge uniformément vers R :

$$\lim_{\ell\to\infty} \mathbb{P}\left(\sup_{\boldsymbol{w}\in\mathcal{W}} (R(\boldsymbol{w}) - R_{\text{emp}}(\boldsymbol{w})) > \varepsilon \right) = 0, \quad \forall\varepsilon > 0$$

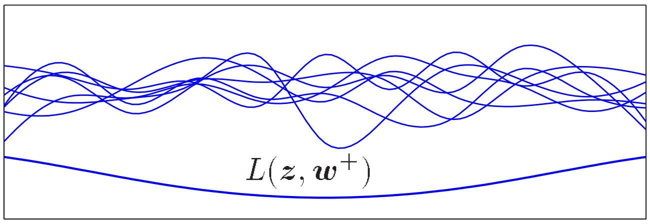
L'approche dans le pire des cas est donc la seule analyse possible pour la consistance du principe de MRE.

Ce résultat est choquant...



...mais s'explique par le concept de consistance non-triviale

Consistance triviale



z =(x,y)

Consistance non-triviale

Définition : Le principe de la MRE est non-trivialement consistant pour P, $\mathcal{H}$ et L sissi

$$\forall c \in \mathbb{R}, \mathcal{W}(c) = \{\boldsymbol{w} : R(\boldsymbol{w}) > c\}, \inf_{\boldsymbol{w} \in \mathcal{W}(c)} R_{\text{emp}}(\boldsymbol{w}_\ell) \xrightarrow{P} \inf_{\boldsymbol{w} \in \mathcal{W}(c)} R(\boldsymbol{w})$$

La consistance non-triviale est liée à des conditions sur la classe de fonctions $L(\boldsymbol{z}, \boldsymbol{w})$.

De la convergence uniforme à la complexité : Cas des fonctions indicatrices

La “diversité” de l'ensemble $L(\boldsymbol{z}, \boldsymbol{w})$, $w \in \mathcal{W}$ sur un échantillon $\{\boldsymbol{z}_i\}_{i=1}^{\ell}$ est donnée par le nombre de dichotomies réalisables $N^{\mathcal{W}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell)$ par $L(\boldsymbol{z}, \boldsymbol{w})$, $w \in \mathcal{W}$.

Ex : $f(x, w) = 1/(1 + exp(-(x - w)); \; y \in \{0, 1\}$

$$L(\boldsymbol{z}, w) = \begin{cases} 0 \text{ si } f(x, w) - y < 1/2 \\ 1 \text{ si } f(x, w) - y \geq 1/2 \end{cases}$$

$N^{\mathcal{W}}(\boldsymbol{z}_1, \boldsymbol{z}_2) = 4$, $N^{\mathcal{W}}(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_\ell) = 2\ell$

Toute la théorie est basée sur $N^{\mathcal{W}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell)$.

De la convergence uniforme à la complexité : la VC-entropie

Définition : $H^{\mathcal{W}}(\ell) = \mathbb{E} \ln N^{\mathcal{W}}(\boldsymbol{z}_1, \dots, \boldsymbol{z}_\ell)$

Théorème : (approx.) Le principe de la MRE est consistant si

$$\lim_{\ell \to \infty} \frac{H^{\mathcal{W}}(\ell)}{\ell} = 0$$

Si ce ratio ne tend pas vers 0, l'apprenti est capable d'apprendre par cœur sur tout ou partie de $(\mathcal{X} \times \mathcal{Y})^\ell$, d'où pas de convergence.

2. Consistance des processus d'apprentissage 17

Entropie

En thermodynamique l'entropie est une mesure du désordre moléculaire. $H = k \ln N$, k constante de Boltzman, et N nombre d'états microscopiques correspondant à un état macroscopique (température + flux de chaleur). Un système isolé évolue nécessairement et irreversiblement vers des états de plus en plus probables et de moins en moins ordonnés (équilibre).

Entropie de Shannon pour une v.a. discrète :
$H = \sum p_i \ln 1/p_i = \mathbb{E}[\ln(1/p)]$ mesure de l'incertitude moyenne de la variable aléatoire.

HEUDIASYC UMR 6599 - Université de Technologie de Compiègne CNRS

Bornes sur les vitesses de convergence des processus d'apprentissage

Principe : la majoration des déviations $R(\boldsymbol{w}) - R_{\text{emp}}(\boldsymbol{w})$ permet de donner une borne sur le risque R, connaissant R_{emp}.

But : contrôler la capité de généralisation de l'apprenti.

Les 3 types de bornes de la théorie :

1. dépendantes de $P(\boldsymbol{x}, y)$ et non-constructives
2. indépendantes de $P(\boldsymbol{x}, y)$ et non-constructives
3. indépendantes de $P(\boldsymbol{x}, y)$ et constructives

La seule (parfois) calculable : la dimension de Vapnik-Chervonenkis (VC dim)

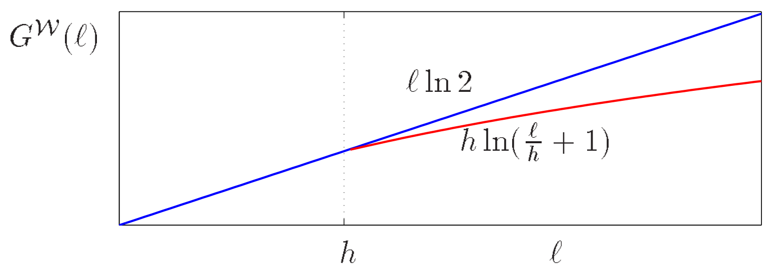
La dimension de Vapnik-Chervonenkis

Définition : La fonction de croissance $G^{\mathcal{W}}(\ell) = \ln \sup\limits_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_\ell} N^{\mathcal{W}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_\ell)$ est un majorant de $H^{\mathcal{W}}(\ell)$.

Théorème : toute fonction de croissance vérifie :

soit $G^{\mathcal{W}}(\ell) = \ell \ln 2$, soit $G^{\mathcal{W}}(\ell) \leq h \ln(\frac{\ell}{h} + 1)$.

h est défini par : $G^{\mathcal{W}}(h) = h \ln 2$, et $G^{\mathcal{W}}(h+1) < (h+1) \ln 2$



h est la VC dim de $L(\boldsymbol{z}, \boldsymbol{w})$, $\boldsymbol{w} \in \mathcal{W}$

Un exemple de borne pour les fonctions indicatrices

$$\mathbb{P}\left(R(\boldsymbol{w}) \leq R_{\text{emp}}(\boldsymbol{w}) + \sqrt{\frac{h(\ln(2\ell/h+1) - \ln(\eta/4))}{\ell}}\right) \geq 1-\eta$$

Remarques : dans l'inégalité, $h\ln(\frac{\ell}{h}+1)$ remplace $G^{\mathcal{W}}(\ell)$, qui remplace $H^{\mathcal{W}}_{\text{ann}}(\ell) = \ln(\mathbb{E}[N^{\mathcal{W}}(\boldsymbol{z}_1, \dots, \boldsymbol{z}_\ell)])$.

h est difficile à calculer, elle est souvent majorée.

Problème :Résultats garantis $\forall P(X,Y)$, mais à quel prix ?

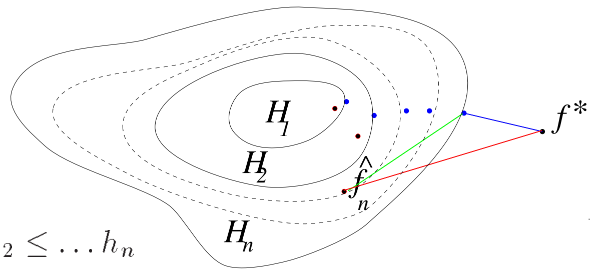
Majoration souvent grossière

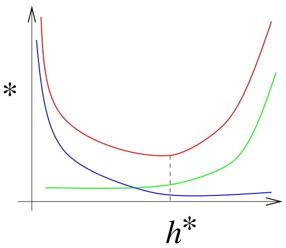
Les praticiens utilisent rarement la VC-dim

On ne peut pas critiquer la théorie dans le pire des cas, mais on peut critiquer les résultats dans le pire des cas, ou essayer de faire mieux.

Contrôle de la capacité de généralisation

Principe : former des structures emboîtées





$h_1 \leq h_2 \leq \dots h_n$

Algorithme

1. Minimisation de R_{emp} sur chaque structure $\to \boldsymbol{w}_\ell$.
2. Calcul $R_{\text{emp}}(\boldsymbol{w}_\ell \leq R(\boldsymbol{w}_\ell) + \Phi(h)$ pour chaque structure
3. Sélectionner la structure de risque minimum

C'est la minimisation du risque structurel