Eléments de Théorie de l'Apprentissage

Machine Learning

Cédric RICHARD
Université Nice Sophia Antipolis

Problème de l'apprentissage fonctionnel

Reconnaissance des formes

La connaissance d'un modèle probabiliste est remplacée par celle d'un ensemble d'apprentissage $\mathcal{A}_n$:

$$\mathcal{A}_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}.$$

L'élaboration d'une règle de décision consiste à rechercher une partition de l'espace des observations $\mathcal{X}$ qui soit optimale au sens du critère de performance choisi.

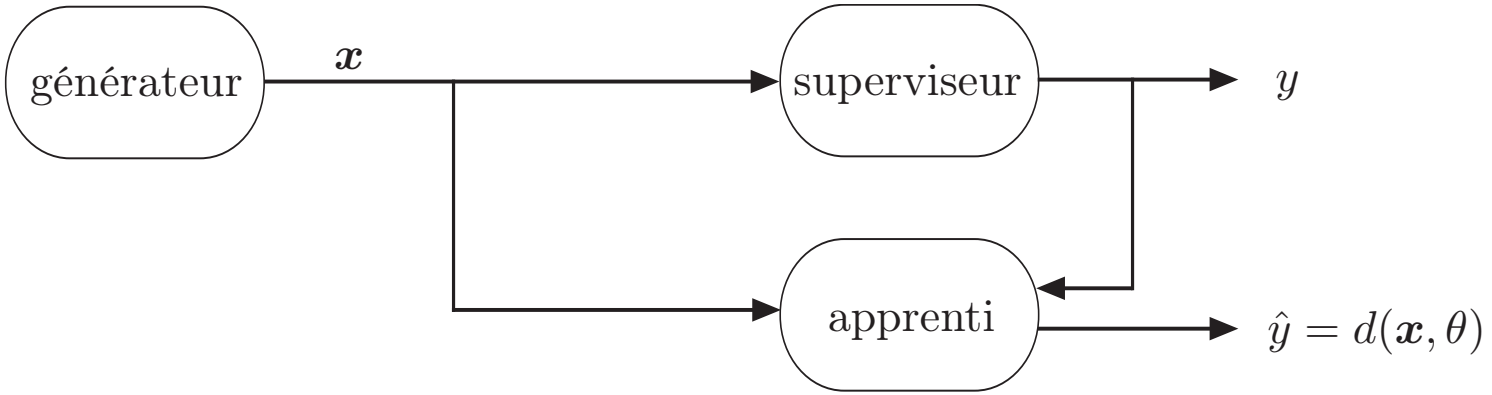
On distingue principalement deux approches possibles :

1. Choix préalable de la structure de la règle de décision, puis optimisation des paramètres caractéristiques selon le critère retenu.
2. Utilisation directe de l'ensemble d'apprentissage pour la prise de décision.

Problème de l'apprentissage fonctionnel

Le modèle d'apprentissage

Le modèle d'apprentissage comporte 3 éléments :



1. Générateur : $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^l$, des vecteurs aléatoires i.i.d.
2. Superviseur : $Y \in \mathcal{Y} \subset \mathbb{R}$, des variables aléatoires
3. Apprenti : représenté par $d(\boldsymbol{x}; \theta) \in \mathcal{D}$

Problème de l'apprentissage fonctionnel

Exemples d'apprentis

– **Polynômes de degré p**

$$d(\boldsymbol{x};\boldsymbol{a}) = \sum_{\substack{i_1,\ldots,i_l \in \mathbb{N} \\ i_1+\ldots+i_l \leq p}} a_{i_1,\ldots,i_l}\, x[1]^{i_1} \ldots x[l]^{i_l}$$

..., et autres décompositions sur une base de Fourier, de Haar, ...

– **Splines**

$$d(\boldsymbol{x};c) \in \mathcal{L}^2(\mathbb{R}^l) \text{ tel que } d' \in \mathcal{L}^2(\mathbb{R}^l), \|d'\|^2 \leq c$$

– **Nadaraya-Watson**

$$d(\boldsymbol{x};\sigma) = \frac{\sum_{i=1}^n y_i\, K_\sigma(\boldsymbol{x},\boldsymbol{x}_i)}{\sum_{i=1}^n K_\sigma(\boldsymbol{x},\boldsymbol{x}_i)}$$

– **MLP, RBF, ...**

$$d(\boldsymbol{x};\boldsymbol{a},\boldsymbol{\theta}) = \sum_k a_k\, g_k(\boldsymbol{x};\boldsymbol{\theta}_k)$$

Problème de l'apprentissage fonctionnel

Minimisation du risque

Objectif. Rechercher au sein de $\mathcal{D} = \{d(\boldsymbol{x}, \theta) \,:\, \theta \in \Theta\}$ une fonction réalisant la meilleure approximation de y au sens d'une fonctionnelle de risque de la forme

$$J(d) = \int Q(d(\boldsymbol{x}, \theta), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy,$$

où Q représente le coût associé à chaque couple $(\boldsymbol{x}, y)$.

Exemple de fonction coût : probabilité d'erreur. Lorsqu'il s'agit d'élaborer une structure de décision de probabilité d'erreur minimale, le risque s'exprime ainsi

$$P_e(d) = \int \mathbb{1}_{d(\boldsymbol{x},\theta)\neq y}\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy,$$

où $\mathbb{1}$ désigne la fonction indicatrice.

Problème de l'apprentissage fonctionnel

Autres exemples de coût

– **Coût quadratique**

$$Q(\boldsymbol{x}, y) = (y - d(\boldsymbol{x}; \theta))^2 \quad \rightarrow \quad d^*(\boldsymbol{x}; \theta) = \mathrm{E}(y \,|\, \boldsymbol{x})$$

– **Coût absolu**

$$Q(\boldsymbol{x}, y) = |y - d(\boldsymbol{x}; \theta)|$$

– **Entropie croisée**

$$Q(\boldsymbol{x}, y) = -y \log(d(\boldsymbol{x}; \theta)) - (1 - y) \log(1 - d(\boldsymbol{x}; \theta)) \quad \rightarrow \quad d^*(\boldsymbol{x}; \theta) = \mathrm{P}(y = 1 \,|\, \boldsymbol{x})$$

Problème de l'apprentissage fonctionnel

Minimisation du risque empirique

Principe MRE. Il s'agit de minimiser la fonctionnelle de risque

$$J(d) = \int Q(d(\boldsymbol{x};\theta), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy.$$

La densité $p(\boldsymbol{x}, y)$ étant inconnue, la minimisation de $J(d)$ se traduit par celle du risque empirique

$$J_{emp}(d) = \frac{1}{n} \sum_{k=1}^{n} Q(d(\boldsymbol{x}_k;\theta), y_k)$$

calculable sur les données constituant l'ensemble d'apprentissage $\mathcal{A}_n$.

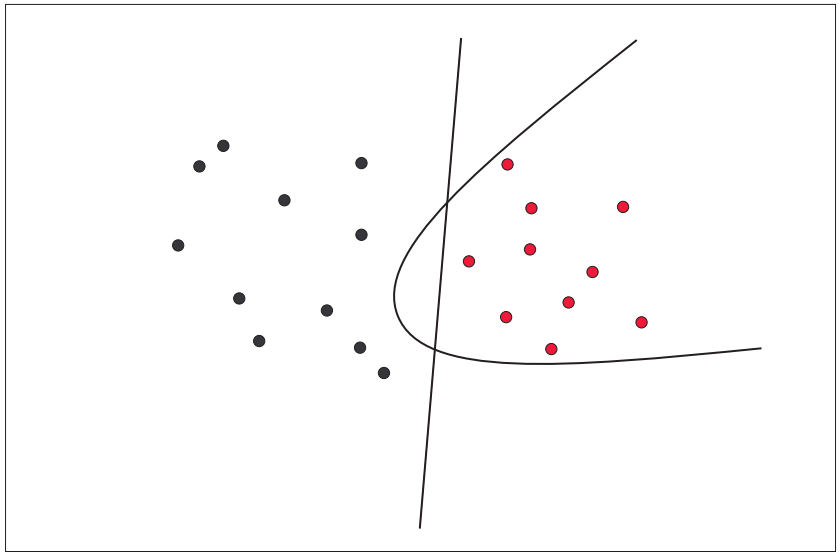
Probabilité d'erreur empirique. Le risque empirique associé à la probabilité d'erreur correspond au nombre d'erreurs d'affectation commises par $d(\boldsymbol{x};\theta)$ sur $\mathcal{A}_n$

$$P_{emp}(d) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{d(\boldsymbol{x}_k;\theta)\neq y_k}.$$

Problème de l'apprentissage fonctionnel

Exemple de mise en œuvre

Problème. Deux familles gaussiennes ω_0 et ω_1 dans $\mathbb{R}^2$, de moyennes et matrices de covariance distinctes, constituées de 10 échantillons chacune.



Quelle frontière choisir ?
Que dire de $\hat{P}_e(\text{linéaire}) = 5\%$ tandis que $\hat{P}_e(\text{quadratique}) = 9\%$?

Problème de l'apprentissage fonctionnel

Erreurs d'approximation et d'estimation

On note $d^* = \arg\min J(d)$ la règle de risque minimum, et $d_n^* = \arg\min_{d\in\mathcal{D}} J_{emp}(d)$ celle obtenue par minimisation du risque empirique sur $\mathcal{D}$ à partir de $\mathcal{A}_n$.

Erreur d'estimation. On définit celle-ci comme étant la différence de performance entre la meilleure règle de $\mathcal{D}$ et celle obtenue au terme de l'apprentissage :

$$J_{est} = J_e(d_n^*) - \inf_{d\in\mathcal{D}} J_e(d)$$

▷ *pertinence du critère empirique et performance de l'algorithme*

Erreur d'approximation. Elle est donnée par la différence de performance entre la règle optimum d^* et la meilleure au sein de $\mathcal{D}$:

$$J_{app} = \inf_{d\in\mathcal{D}} J_e(d) - J_e(d^*)$$

▷ *choix de la classe $\mathcal{D}$*

Problème de l'apprentissage fonctionnel

Erreur de modélisation

L'objectif de l'apprentissage est de minimiser l'erreur de modélisation, définie par :

$$J_{mod}(d_n^*) = J_e(d_n^*) - J_e(d^*).$$

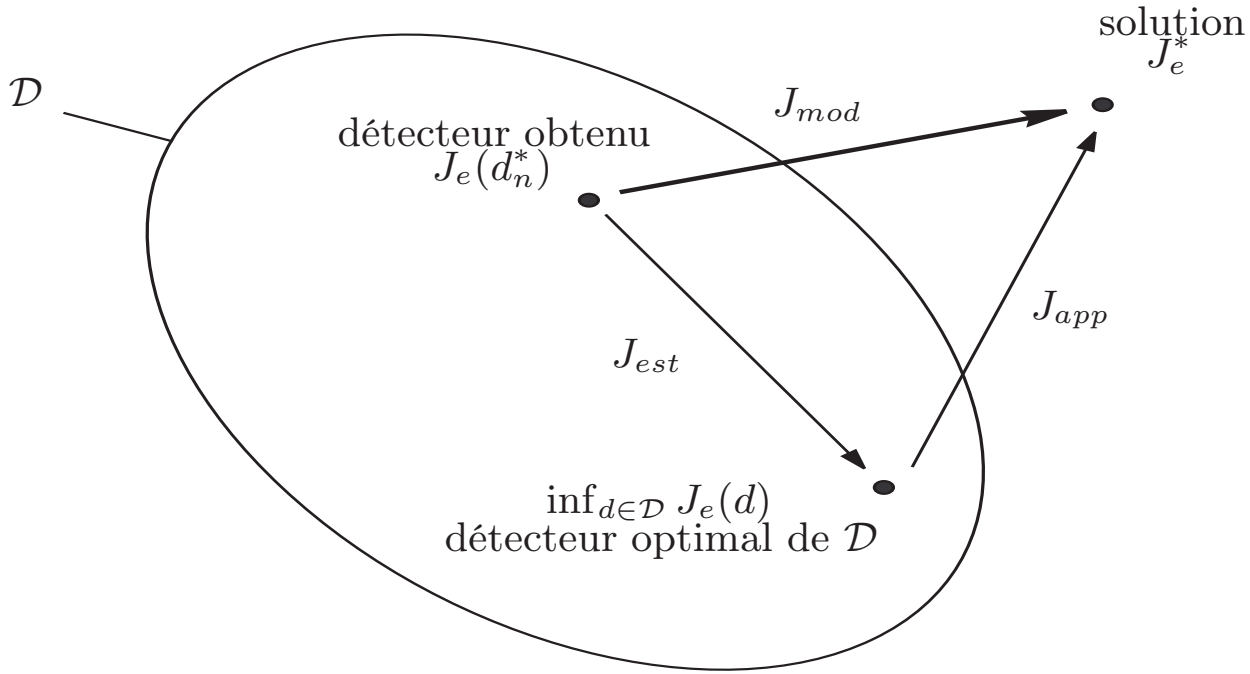
On distingue deux contributions de natures différentes dans cette erreur :

$$J_{mod}(d_n^*) = \underbrace{\left(J_e(d_n^*) - \inf_{d \in \mathcal{D}} J_e(d) \right)}_{J_{est}} + \underbrace{\left(\inf_{d \in \mathcal{D}} J_e(d) - J_e(d^*) \right)}_{J_{app}}.$$

La minimisation de J_{mod} repose sur la recherche d'un compromis entre ces deux termes antagonistes : l'augmentation du nombre de tests de $\mathcal{D}$ conduit à un accroissement de J_{est} tandis que J_{app} décroît, et inversement.

Problème de l'apprentissage fonctionnel

Erreur d'approximation, d'estimation et de modélisation



Problème de l'apprentissage fonctionnel

Questions

1. L'objectif est-il réalisable ?

→ *Consistance de la règle de décision*

→ *Consistance du principe d'induction*

→ *Vitesse de convergence*

2. : Si oui, comment en pratique ?

Consistance d'une règle de décision

Consistance et consistance forte

On peut espérer qu'il existe dans la classe $\mathcal{D}$ considérée, une suite $\{d_n^*(\boldsymbol{X};\theta)\}_{n>0}$ de détecteurs optimaux au sens du critère retenu telle que $P_e(d_n^*)$ puisse être rendue arbitrairement proche de P_e^* lorsque n tend vers l'infini.

Définition 1. *étant donnée une base $\mathcal{A}_n$, une suite $\{d_n^*(\boldsymbol{X};\theta)\}_{n>0}$ de détecteurs optimaux au sens d'un critère donné est dite consistante pour une loi $p(\boldsymbol{x}, y)$ si*

$$\lim_{n\to\infty} \mathrm{E}\{P_e(d_n^*;\mathcal{A}_n)\} = P_e^*.$$

On dit qu'elle est fortement consistante si

$$\lim_{n\to\infty} P_e(d_n^*;\mathcal{A}_n) = P_e^*$$

avec une probabilité égale à 1.

Consistance d'une règle de décision

Consistance universelle

On peut distinguer le cas où la propriété de consistance n'est vérifiée que pour une loi $p(\boldsymbol{x}, y)$ donnée, du cas où elle reste vraie indépendamment de celle-ci.

Définition 2. *La suite* $\{d_n^*(\boldsymbol{X}; \theta)\}_{n>0}$ *est dite universellement (fortement) consistante si elle est (fortement) consistante pour toute loi de probabilité* $p(\boldsymbol{x}, y)$.

Cette propriété a été observée pour la première fois en 1977 par Stone dans le cadre de la méthode des *k plus proches voisins*, à la condition que le paramètre k croisse moins vite que la taille n de la base d'apprentissage. Depuis, il a été démontré que d'autres règles de décision y satisfont :

- fonctions à noyaux réguliers
- certains détecteurs linéaires généralisés
- (...)

Consistance du principe d'induction

Définition

Le principe de MRE est consistant pour la fonction objectif choisie et le problème, si l'apprenti fait du mieux possible quand la taille de l'échantillon tend vers l'infini.

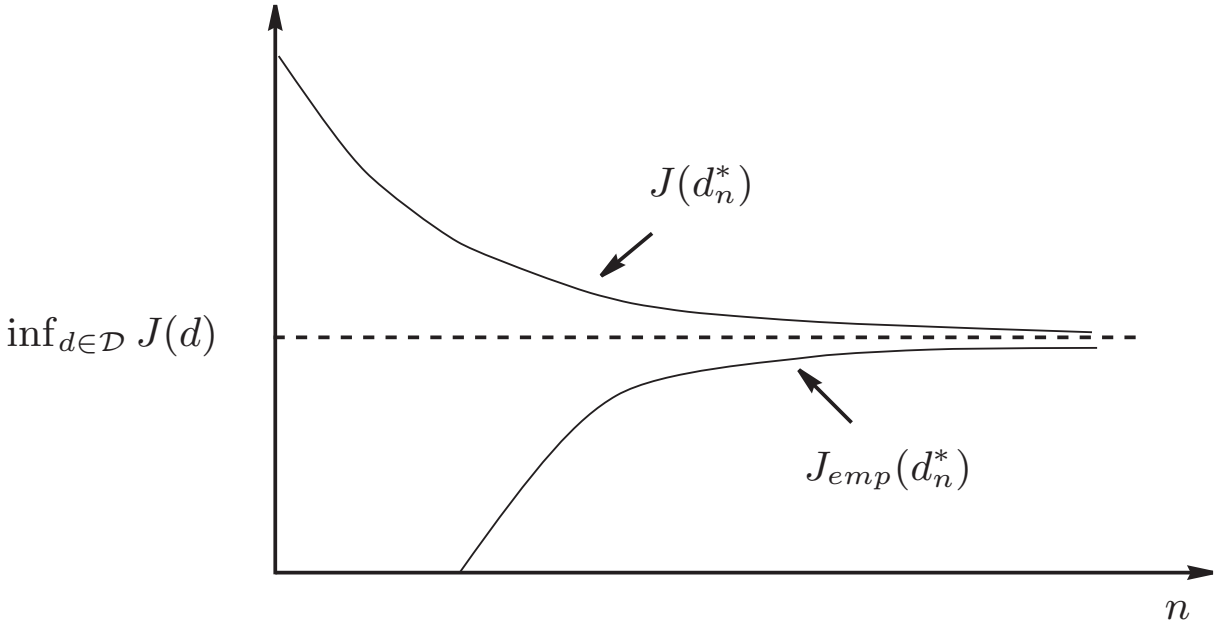
Définition 3. *Le principe de MRE est consistant pour un coût Q, une famille de fonctions $\mathcal{D} = \{d(\boldsymbol{x}; \theta) : \theta \in \Theta\}$ et une distribution $p(\boldsymbol{x}, y)$ si, appliqué à chaque taille n d'échantillon, il engendre une suite $\{d_n^*(\boldsymbol{x}; \theta) : \theta \in \Theta\}_{n>0}$ telle que*

$$J(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d)$$

$$J_{emp}(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d).$$

Consistance du principe d'induction

Illustration de la définition



$$J(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d)$$

$$J_{emp}(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d).$$

Consistance du principe d'induction

VC-dimension

Par soucis de clarté, on considère dans la suite de cette section que le coût Q prend la forme d'une fonction indicatrice, soit

$$Q(d(\boldsymbol{x};\theta);y) = \mathbb{1}_{d(\boldsymbol{x};\theta)\neq y} \triangleq \begin{cases} 0 & \text{si} \quad y = d(\boldsymbol{x};\theta) \\ 1 & \text{si} \quad y \neq d(\boldsymbol{x};\theta), \end{cases}$$

Définition 4. *La dimension de Vapnik-Chervonenkis d'une classe $\mathcal{D}$ donnée est définie par le plus grand nombre d'éléments $\boldsymbol{x}_k$ de l'espace des réalisations $\mathcal{X}$ dont les détecteurs de $\mathcal{D}$ peuvent réaliser toutes les dichotomies.*

Consistance du principe d'induction

Exemples de VC-dimension

Exemple 1. On considère la classe $\mathcal{D}$ des détecteurs linéaires opérant dans $\mathbb{R}^l$ définis par $d(\boldsymbol{x};\theta) = \text{sign}(\sum_{k=1}^{l} \theta_k\, x(k) + \theta_0)$, les paramètres θ_k étant réels et sign$(\cdot)$ désignant la fonction « signe ». On montre que

$$h_{\mathcal{D}} = l + 1$$

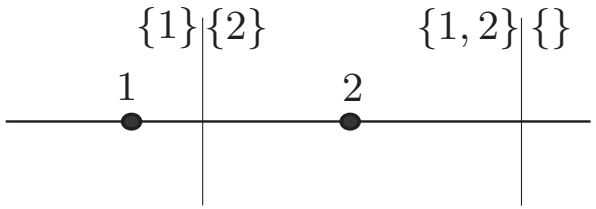
Exemple 2. On considère la classe des détecteurs
$\{d(x;\theta) = \text{sign}(\sin(\theta x)) \,:\, \theta \in \mathbb{R}\}$ opérant dans $\mathbb{R}$. Il est aisé de démontrer que

$$h_{\mathcal{D}} = +\infty$$

Consistance du principe d'induction

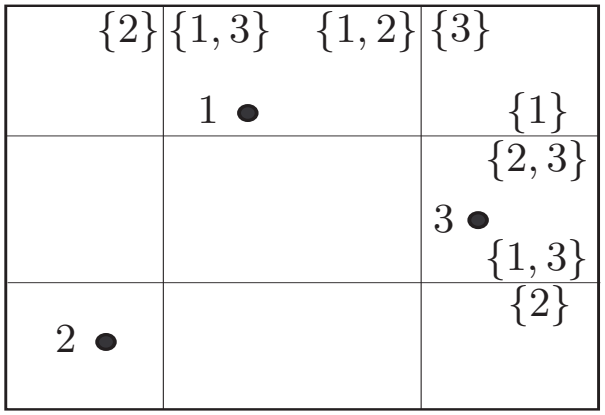
Caractérisation de la VC-dimension dans le cas linéaire

dans $\mathbb{R}$

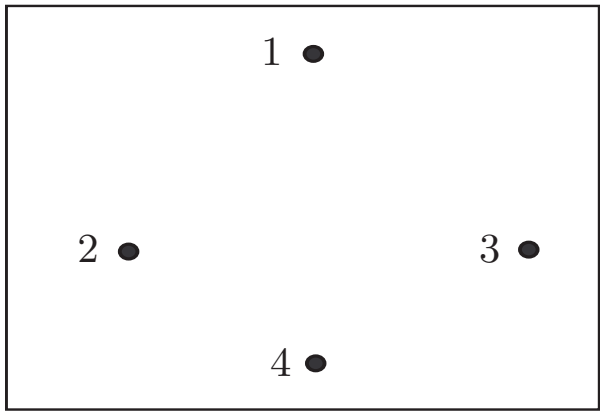


dans $\mathbb{R}$

dans $\mathbb{R}^2$



dans $\mathbb{R}^2$



Consistance du principe d'induction

Théorème fondamental

Théorème 1. *Pour que le principe MRE soit consistant indépendamment de la distribution de probabilité gouvernant les observations, il suffit que la classe de détecteurs considérée soit de VC-dimension h finie.*

Consistance du principe d'induction

Vitesse de convergence de P_{emp} vers P_e

les travaux précurseurs de Vapnik et Chervonenkis (1971) ont également apporté des enseignements quantitatifs relatifs à la vitesse de convergence de P_{emp} vers P_e.

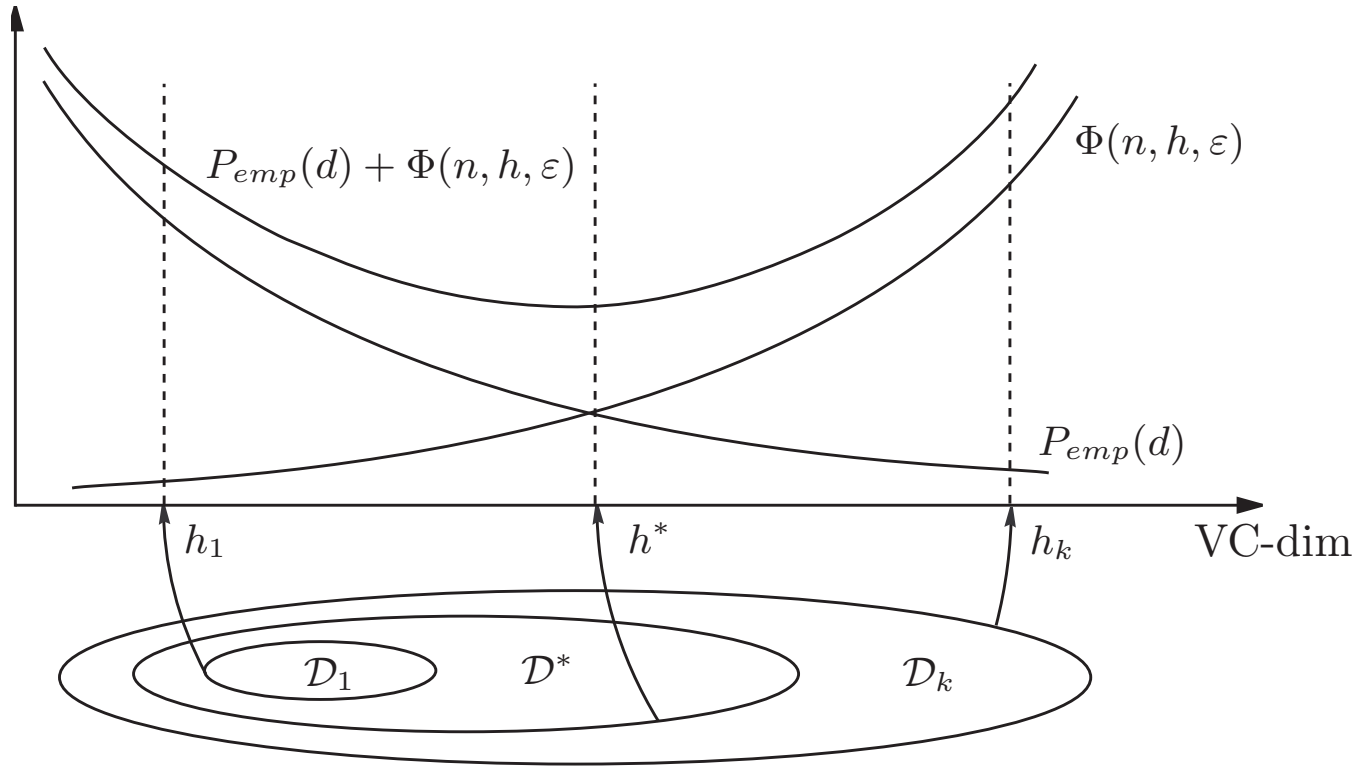
Inégalité de Vapnik-Chervonenkis. Avec une probabilité égale à $1 - \varepsilon$ au moins, on a :

$$P_e(d) \leq P_{emp}(d) + \sqrt{\frac{h \ln\left(\frac{2en}{h}\right) - \ln\frac{\varepsilon}{4}}{n}}.$$

Attention ! Majoration souvent grossière... mais indépendante de toute loi $p(\boldsymbol{x}, y)$.

Principe de minimisation du risque structurel

– MRS –



Principe de minimisation du risque structurel

– MRS –

Le principe de *minimisation du risque structurel* préconisé par Vapnik suppose la construction, au sein de la classe $\mathcal{D}$, d'une séquence de sous-ensembles imbriqués $\mathcal{D}_k$

$$\mathcal{D}_1 \subset \ldots \subset \mathcal{D}_k \subset \ldots \subset \mathcal{D}.$$

Cette structure étant établie, la phase d'apprentissage est menée en deux étapes :

1. Recherche du détecteur d'erreur empirique minimale dans chaque sous-ensemble $\mathcal{D}_k$:
$$d^*_{n,k} = \arg \min_{d \in \mathcal{D}_k} P_{emp}(d).$$
2. Sélection du détecteur présentant l'erreur garantie $P_{emp}(d^*_{n,k}) + \Phi(n, h_k, \varepsilon)$ la plus favorable :
$$d^*_n = \arg \min_{k \geq 1} \{P_{emp}(d^*_{n,k}) + \Phi(n, h_k, \varepsilon)\}.$$