Notes sur la dimension de Vapnik Chervonenkis

Maximilien

July 2020

Notions MIASHS fonctions, $\mathbb{R}^n$, calcul de dérivés, probabilité, espérance, combinatoire

1 Introduction

L'idée assez standard admise derrière l'apprentissage automatique est que "toutes les théories sont fausses, mais certaines sont utiles". On ne voit pas les mécanismes physiques qui décrivent la nature comme une vérité vraie, mais comme des approximations permettant de réaliser des prédictions utiles.

Le principe du rasoir d'Ockham nous donne une règle finalement assez intuitive afin de sélectionner une théorie parmi d'autres. Ce principe nous dit qu'il ne faut pas multiplier les hypothèses *ad-hoc*. Autrement dit, à pouvoir prédictif équivalent, il faut privilégier la théorie la plus simple. En effet, à partir d'un certain niveau de complexité, il sera toujours possible de trouver une théorie qui explique parfaitement les observations faites avec des règles complètement arbitraire. La particularité d'une telle théorie est son pouvoir prédictif mauvais voire nul. Ce raisonnement intuitif peut se formaliser via le *framework* des probabilités (Bayésiennes). D'autres nécessité comme le bsoin de pouvoir falsifier la théorie explique l'intérêt de ce principe [1].

Toutes ces idées se retrouvent en apprentissage automatique où la première étape d'optimisation d'optimisation peut être vue comme la recherche d'une théorie dont on souhaite un fort pouvoir prédictif. De fait, il convient de pouvoir mesure cette idée de "simplicité" de la théorie et la dimension de Vapnik Cherkonenkis (ou dimension VC) entre en jeu lorsque le cardinal de notre ensemble de théories est infini. Plus qu'une mesure intuitive de la taille d'un ensemble de théories, elle permet de fournir des "majorants de généralisation", c'est-à-dire de majorer la probabilité que notre optimisation permette de trouver une fonction avec un fort pouvoir prédictif.

Toutes ces difficultés apparaissent car les critères qui nous font choisir une théorie plutôt qu'une autre ne sont pas exactement ceux qu'on souhaite. On choisit une théorie à partir d'observations **passées**, mais on veut qu'elle fournisse des prédictions **futures** intéressantes.

1. cf. The Logic of Scientific Discovery, Karl Popper

2 Formalisation de l'apprentissage automatique

Soit deux variables aléatoires $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ où on pourrait par exemple voir $\mathcal{X}$ comme un ensemble de photos et $\mathcal{Y}$ un ensemble d'étiquettes sur ces photos (e.g. chien et chats).

On supposera $X \times Y \sim p$, où p ne serait pas connue. Notre objectif est de trouver une fonction f telle qu'on arrive prédire Y à partir de X en minimisant une certaine notion d'erreur.

Définition de l'erreur. La première étape consiste à définir l'erreur pour une réalisation. Soit $X \times Y \sim p$, une erreur élémentaire peut-être vue de deux manières différentes :

— $r : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$

— $r : \mathcal{Y} \times \mathbb{P}_{\mathcal{Y}} \mapsto \mathbb{R}^+$

Dans le premier cas, on peut voir l'erreur comme une "distance" entre deux éléments de $\mathcal{Y}$ et dans le second comme une "mesure" de l'adéquation entre une mesure de probabilité sur $\mathcal{Y}$ et un élément de $\mathcal{Y}$. Les moindres carrés sont une forme d'erreur où l'écart entre deux éléments de $\mathcal{Y} = \mathbb{R}$ est calculé comme l'écart quadratique de leur valeur :

$$\begin{aligned} r : \mathcal{Y} \times \mathcal{Y} &\mapsto \mathbb{R}^+ \\ \hat{y}, y &\to (\hat{y} - y)^2 \end{aligned}$$

De manière similaire et tout aussi standard, dans le cadre d'un problème de classification, nous avons l'entropie relative. Ici, $\mathcal{Y} = \{i\}_{i \leq K}$.

$$\begin{aligned} r : \mathbb{P}_{\mathcal{Y}} \times \mathcal{Y} &\mapsto \mathbb{R}^+ \\ \hat{y}, y &\to -\sum_k \delta_{k=y} \ln\big(\hat{y_k}\big), \end{aligned}$$

où $\hat{y}$ est associé à un point du K-simplexe interprété de la manière suivante : $\mathbb{P}(y = k|x) = \hat{y}(x)_k$.

L'objectif du *machine learning* est d'avoir les meilleures performances en espérance relativement à ces erreurs. On parlera de risque de généralisation lorsqu'on fera référence à l'erreur en espérance pour une fonction $h : \mathcal{X} \mapsto \mathcal{Y}/\mathbb{P}_{\mathcal{Y}}$:

$$R(h) = \mathbb{E}_{X \times Y \sim p}\big[r(h(X), Y)\big].$$

Soit $\mathcal{H}$ un ensemble de fonctions, l'objective du *machine learning* est de trouver la fonction f de $\mathcal{H}$ suivante :

$$f = \operatorname*{argmin}_{h \in \mathcal{H}} R(h).$$

Ce problème n'est malheureusement pas résoluble, ne serait-ce que parce que nous ne connaissons pas la loi p.

La stratégie mise en place en pratique et appellée *minimisation du risque empirique* (ERM) consiste à échantillonner selon p (i.e. collecter des données)

$\mathcal{D} = \{(x_i, y_i)\}_{i \leq N} \sim p^N$ (chaque couple est i.i.d.). Ainsi, le risque empirique suivant est un estimateur sans biais du risque de généralisation :

$$Re(h) = \frac{1}{N} \sum_{i=1}^{N} r(h(x_i), y_i),$$

et l'objectif se reformule de la manière suivante :

$$\hat{f} = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, Re(h).$$

La quantité $Re(\hat{f})$ n'est quant à elle plus sans biais. On sait qu'on va favoriser les fonctions particulièrement adaptées aux fluctuations aléatoires de notre jeu de données. La question soulevée ici est celle du *gap de généralisation*, c'est-à-dire :

$$\left| Re(\hat{f}) - R(\hat{f}) \right|.$$

Il semble intuitif que plus l'ensemble de fonctions considéré $\mathcal{H}$ sera grand, plus on risque de trouver une fonction arbitrairement "forte" sur $\mathcal{D}$ mais mauvaise en espérance. L'exemple paradigmatique de cette idée est la fonction mémoire :

$$f(x) = \begin{cases} y, & \text{si } (x, y) \in \mathcal{D} \\ \text{random}(\mathcal{Y}), & \text{sinon} \end{cases}$$

où random($\mathcal{Y}$) retourne un élément choisi aléatoirement de $\mathcal{Y}$. On voit bien que la fonction f ainsi construite ne nous intéresse pas.

La suite de ces notes se concentrera sur les cas où $|\mathcal{H}| = \infty$. On essaiera en particulier d'établir quelques majorants de généralisation relativement à la taille de l'ensemble de fonctions et à celle du jeu de données.

3 La dimension de Vapnik et Chervonenkis

La dimension VC définit une notion de taille d'un ensemble de fonctions du point de vu d'une tâche de *machine learning* et notamment de classification ($\mathcal{Y}$ discret non ordonné).

Par simplicité, restons dans le cas des fonctions $h : \mathcal{D} \mapsto \{0, 1\}$ (ou $\{-1, +1\}$). La première étape dans la construction d'une dimension VC est la définition suivante.

Définition 1 (Fonction de croissance (growth function))**.** *La fonction de croissance* $\Pi_{\mathcal{H}} : \mathbb{N} \mapsto \mathbb{N}$ *pour une classe de fonctions* $\mathcal{H}$ *est définie comme :*

$$\forall m \in \mathbb{N}, \Pi_{\mathcal{H}}(m) = \max_{\{x_1, \ldots, x_m\} \subseteq \mathcal{X}} |\{h(x_1), ..., h(x_m) : h \in \mathcal{H}\}| \tag{1}$$

On peut voir les fonctions $h \in \mathcal{H}$ comme des "étiqueteurs binairs" du jeu de données. Ainsi, $\Pi_{\mathcal{H}}(m)$ est le nombre de manières différentes qu'il y a d'étiqueter

un jeu de données de taille m avec des fonctions de $\mathcal{H}$. Cette quantité est majorée par 2^m qui est le nombre maximum d'étiquetage binaire d'un jeu de données de taille m.

La dimension VC d'un ensemble de fonction de $\mathcal{X}$ dans un ensemble binaire se construit comme suit :

Définition 2 (Dimension VC)**.** *La dimension VC d'un ensemble de fonctions $\mathcal{H}$ est le plus grand jeu de données (non aligné) qui peut être totalement séparé par $\mathcal{H}$:*

$$VCdim(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}. \tag{2}$$

On verra par la suite qu'on peut majorer le gap de généralisation d'un apprentissage par une formule qui dépend de $VCdim$. Avant cela, calculons la dimension VC de quelques exemples.

Calculer la dimension VC d'un ensemble de fonctions se passe en deux étapes :

1. Trouver des exemples en augmentant le nombre de point du jeu de données,
2. lorsqu'on ne trouve plus d'exemple, trouver une démonstration que la dimension est inférieure à cette quantité.

Exemple 1 (La fonction constante)**.** *Soit $\mathcal{H}$ une classe d'une seule fonction. La dimension VC est* 0.

Démonstration. Soit $x \in \mathcal{X}$, $y = 1$ et $f \in \mathcal{H}$. Si $f(x) \neq y$, alors f ne peut pas correctement classer l'élément. Sinon mettre $y = 0$. □

Exemple 2 (Les rectangles alignés avec les axes)**.** *Soit $\mathcal{X} = \mathbb{R}^2$ et $\mathcal{Y} = \{-1, +1\}$. On considère la classe de fonction qui à partir d'un rectangle quelconque aligné avec les axes retourne $+1$ si le point est contenu par le rectangle et -1 sinon. La dimension VC d'une telle classe est 4.*

Démonstration. Il est facile de trouver un exemple avec moins de 4 points.

La figure 1 illustre un jeu de données de taille 4 qui peut être "pulvérisé" par la classe $\mathcal{H}$.

Montrons que la dimension VC n'est pas 5. Soit $x_1, x_2, x_3, x_4, x_5 \in \mathcal{X}$. Soit le sous ensemble $\mathcal{E}$ de ces points qui regroupe le point le plus à gauche, le plus à droite, le plus haut et le plus bas. Puisque les points ne sont pas alignés, $|\mathcal{E}| = 4$. Associons à tous les points de $\mathcal{E}$ le label $+1$ et -1 aux autres. Soit le plus petit rectangle R contenant ces quatre points. Par définition, tous les autres rectangle permettant de retourner $+1$ pour les points de $\mathcal{E}$ contiennent R. Cependant, R contient également les 5 points puisque ces côtés sont définis par les extremums du jeu de données. Il ne peut donc pas les classer avec -1 et la dimension VC d'une telle classe ne peut pas être 5. On pourrait aller plus loin en soulignant qu'un rectangle est convexe, etc. □

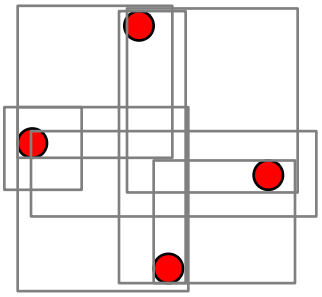

FIGURE 1 – 4 points dont toutes les combinaisons de classes peuvent être séparées par $\mathcal{H}$.

Exemple 3 (Séparation par seuil)**.** *Soit la classe de fonctions :*

$$\mathcal{H} = \{f_\theta : \theta \in \mathbb{R}\}$$

où

$$f_\theta(x) = \begin{cases} 1 \ si \ x > \theta \\ 0 \ sinon. \end{cases}$$

Alors, la dimension VC est 1.

Démonstration. 1. La dimension VC est au moins 1. En effet, soit $x \in \mathbb{R}$. Soit $y = 1$, alors la fonction f_θ avec $\theta = x - 1$ retrouve bien y. Soit $y = 0$ alors la fonction f_θ avec $\theta = x + 1$ retrouve bien y.

2. La dimension VC est inférieure à 2. Soit $x_1, x_2 \in \mathbb{R}$ tels que $x_1 < x_2$. Soit $y_1 = 1$ et $y_2 = 0$, alors $\nexists \theta \in \mathbb{R}$, $x_1 > \theta > x_2$ (évident).

□

Exemple 4 (Séparation sur $\mathbb{R}$)**.** *Soit la classe de fonctions :*

$$\mathcal{H} = \{f_{\theta,s} : \theta \in \mathbb{R}, s \in \{+1, -1\}\}$$

où

$$f_{\theta(x),s} = \begin{cases} 1 \ si \ x > \theta \ et \ s = +1 \\ 1 \ si \ x \leq \theta \ et \ s = -1 \\ 0 \ sinon. \end{cases}$$

La dimension VC d'une telle classe de fonctions est 2.

Démonstration. 1. La dimension est au moins 2. Soit $x_1, x_2 \in \mathbb{R}$ et $x_1 < x_2$. Soit $y_1 = 0$ et $y_2 = 1$. Alors le modèle $\theta = \frac{x_1+x_2}{2}$ et $s = +1$ est un bon classifieur. Si $y_1 = 1$ et $y_2 = 0$ alors le modèle $\theta = \frac{x_1+x_2}{2}$ et $s = -1$ est un bon classifieur.

2. La dimension ne peut pas être 3. Soit $x_1, x_2, x_3 \in \mathbb{R}$, $x_1 < x_2 < x_3$. Soit $y_1 = 0, y_2 = 1, y_3 = 0$. Si $\theta < x_1$ alors le modèle n'est pas un bon classifieur. Si $x_1 < \theta < x_2$ alors le classifieur n'est pas bon. Même chose si $x_2 < \theta < x_3$ et si $x_3 < \theta$.

□

Exemple 5 (Séparation par hyperplan)**.** *Soit $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{+, -\}$ et $\mathcal{H} = \{h : h(\mathbf{x}) = sign(\langle \theta, \mathbf{x} \rangle + b), x \in \mathcal{X}, \theta \in \mathbb{R}^2, b \in \mathbb{R}\}$. On observe de manière triviale qu'il est toujours possible de séparer deux points : $VCdim(\mathcal{H}) \geq 2$. En étudiant les différents cas où nous avons trois points non-alignés, on se rend compte qu'on peut également les séparer : $VCdim(\mathcal{H}) \geq 3$. Enfin, il est assez facile d'observer qu'avec quatre points, il y a des situations non séparables : $VCdim(\mathcal{H}) = 3$.*

De manière plus générale, pour $\mathcal{X} = \mathbb{R}^d$, nous avons $VCdim(\mathcal{H}) = d + 1$.

La preuve de ce dernier exemple est plus complexe.

Théorème 1. *Soit $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{+, -\}$ et $\mathcal{H}$ l'ensemble des classifieurs linéaires tels que définis précédemment. Alors,*

$$VCdim(\mathcal{H}) \geq d + 1$$

Démonstration. Soit $\mathcal{S} = \{x_0, ..., x_d\}$. Mettons x_0 de côté. Fixons $x_i = e_i$ (vu que les points ne sont pas alignées, il forment ainsi une base). Enfin, soit $y_i \in \{+, -\}$ choisis arbitrairement. Soit $\theta = (y_1, ..., y_d)$ et $b = \frac{y_0}{2}$. Alors $h(\mathbf{x}) = \langle \theta, \mathbf{x} \rangle + b$ sépare nécessairement l'ensemble $\mathcal{S}$. Le produit scalaire $\langle \cdot, \cdot \rangle$ est vu comme le produit scalaire canonique dans la nouvelle base. Ainsi, on sait que $\text{VCdim}(\mathcal{H}) \geq d + 1$. □

Théorème 2 (Théorème de Radon)**.** *Pour tout ensemble $\mathcal{S}$ de $d + 2$ vecteurs de $\mathbb{R}^d$, il est possible de trouver une partition $\mathcal{S}_1$, $\mathcal{S}_2$ telle que leur enveloppe convexe est en intersection.*

Démonstration. Soit $\mathcal{S} = \{x_1, ..., x_{d+2}\} \subset \mathbb{R}^d$. Soit $\alpha_1, ..., \alpha_{d+2} \in \mathbb{R}$ et le système d'équations suivant :

$$\sum_{i=1}^{d+2} \alpha_i x_i = 0 \text{ et } \sum_{i=1}^{d+2} \alpha_i = 0$$

En notant que $\sum_{i=1}^{d+2} \alpha_i x_i = 0$ peut se décomposer en d équations puisque $x_i \in \mathbb{R}^d$, on constate que nous avons $d + 1$ équations et $d + 2$ inconnues (les α_i). Le système admet donc une solution non nulle et appelons $\beta_1, ..., \beta_{d+2} \in \mathbb{R}$ cette solution. Puisque $\sum_{i=1}^{d+2} \beta_i = 0$, nécessairement les ensembles $I_1 = \{i \in [1, d + 2] : \beta_i > 0\}$ et $I_2 = \{i \in [1, d + 2] : \beta_i < 0\}$ sont non vides. Soit la partition $X_1 = \{x_i : i \in I_1\}$ et $X_2 = \{x_i : i \in I_2\}$. En partant du système, on observe que $\sum_{i \in I_1} \beta_i = -\sum_{i \in I2} \beta_i$. Soit $\beta = \sum_{i \in I_1} \beta_i$, on a directement par la première équation :

$$\sum_{i \in I_1} \frac{\beta_i}{\beta} x_i = -\sum_{i \in I_2} \frac{\beta_i}{\beta} x_i$$

En notant que $\sum_{i\in I_1}\frac{\beta_i}{\beta} = -\sum_{i\in I_2}\frac{\beta_i}{\beta} = 1$, $\frac{\beta_i}{\beta} \geq 0 \forall i \in I_1$ et $-\frac{\beta_i}{\beta} \geq 0 \forall i \in I_2$, par définition d'une enveloppe convexe, le point résultat de la combinaison linéaire se situe nécessairement à la fois dans l'enveloppe convexe de X_1 et dans celle de X_2. □

Théorème 3. *Soit $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{+, -\}$ et $\mathcal{H}$ l'ensemble des classifieurs linéaires tels que définis précédemment. Alors,*

$$VCdim(\mathcal{H}) = d + 1$$

Démonstration. Nous avons déjà démontré que $\text{VCdim}(\mathcal{H}) \geq d + 1$. Soit un ensemble de $d + 2$ points. S'il existe un hyperplan qui sépare deux partitions de cet ensemble, alors il sépare également leur enveloppe convexe. Le théorème de Radon, nous dit que pour une dimension d, avec $d + 2$ points, il existe toujours une partition telle que les enveloppes convexes sont en intersection. Donc $d + 2$ points ne peuvent pas toujours être séparables par un hyperplan. □

4 Majorant de généralisation

La dimension VC permet d'exprimer une notion de taille d'un ensemble de fonctions du point de vu *machine learning* qu'est la classification. En particulier, le théorème suivant majore le risque de généralisation (où nous notons $|\cdot| = VCdim(\cdot)$).

Théorème 4 (Majorant du risque de généralisation)**.** *Soit $\mathcal{H}$ un ensemble de fonctions $h : \mathcal{X} \mapsto \mathcal{Y} = \{+1, -1\}$. Soit $|\mathcal{H}|$ la dimension VC de cet ensemble. Soit un jeu de données $\mathcal{S}$ avec $|\mathcal{S}| = m$. Alors, $\forall \delta > 0$, avec probabilité $1 - \delta$, nous avons l'inégalité suivante :*

$$\forall h \in \mathcal{H}, R(h) \leq Re_{\mathcal{S}}(h) + \sqrt{\frac{2|\mathcal{H}| ln\frac{em}{|\mathcal{H}|}}{m}} + \sqrt{\frac{ln\frac{1}{\delta}}{2m}}$$

En termes d'ordre de grandeur, on obtient l'inégalité suivante :

$$\forall h \in \mathcal{H}, R(h) \leq Re_{\mathcal{S}}(h) + \mathcal{O}\left(\sqrt{\frac{ln(m/|\mathcal{H}|)}{(m/|\mathcal{H}|)}}\right)$$

Démonstration. Admis □

On observe qu'il suffit d'augmenter la taille de notre jeu de données ou de réduire celle de notre ensemble de fonction pour réduire l'écart entre le risque empirique et le risque de généralisation.

Les réseaux de neurones sont des exemples où la dimension VC ne permet pas de majorer de manière efficace le risque de généralisation.