Notes sur la notion de surapprentissage

Maximilien

June 2020

Notions MIASHS fonctions, probabilités, loi des grands nombres, estimateurs, théorème central limite, formule de Taylor.

1 Introduction

Intuitivement, on cherche une fonction $h : \mathcal{X} \mapsto \mathcal{Y}$ où $\mathcal{X} \subset \mathbb{R}^n$ représente par exemple l'ensemble des photos de chiens et de chats et $\mathcal{Y} = \{0, 1\}$ avec $0 =$ chien et $1 =$ chat.

Trouver/calculer cette fonction peut se faire de nombreuses manières différentes. En *machine learning* les différentes stratégies ont en commun de s'appuyer sur un jeu de données $\mathcal{D} = \{(x_i, y_i)\}_{i \leq N}$, $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ dit jeu d'apprentissage. La question clé soulevée par le *machine learning* est de déterminer à quel point la fonction qu'on aura calculé h aura réussi à déterminer les structures déterministes séparant les objets de nos différentes classes ou, à l'inverse, elle n'aura réussi à séparer les éléments de notre jeu de données uniquement à partir du bruit le composant, e.g. h reconnaît une photo de chien du jeu de données, plutôt que ce qui fait qu'un chien est un chien.

Ce document formalise une vision un peu plus théorique de cette problématique dans le cadre de ce qu'on appelle la "minimisation du risque empirique". La stratégie ERM est presque transversal au *machine learning* même si certains modèles comme ceux s'appuyant sur les k voisins n'en font pas parti. Cependant, même pour le KNN, il revient à trouver une bonne valeur de k et ce choix se fera par ERM.

2 Minimisation du risque empirique

Soient $X \in \mathcal{X}$ et $Y \in \mathcal{Y}$ deux variables aléatoires de loi jointe $X, Y \sim f_{\mathcal{D}}$. La loi $f_{\mathcal{D}}$ n'est évidemment pas connue et peut être vue comme le modèle réelle régissant les données dans $\mathcal{X}$ (e.g. photos de chiens et de chats) et leur label $\mathcal{Y}$.

2.1 Le risque élémentaire

L'étape suivante consiste à définir cette notion de risque. Ce dernier se construit en deux étapes dont la première est le risque élémentaire $r : \mathcal{Y} \times \mathcal{Y} \mapsto$

$\mathbb{R}^+$ qui, à partir de deux labels mesure un écart entre ces derniers. Le mot écart est choisi pour éviter celui de distance, cette dernière devant satisfaire certaines propriétés qui ne sont pas nécessairement satisfaites par notre risque élémentaire. Un exemple est l'écart quadratique $r(\hat{y}, y) = (\hat{y} - y)^2$. Le chapeau fait référence à la variable prédite par notre modèle. Cela souligne la non symétrie de cet écart. Un risque élémentaire utilisé en classification est l'entropie croisée $r(\hat{y}, y) = -\big(y\ln(\hat{y}) + (1-y)\ln(1-\hat{y})\big)$. Notons que dans la dernière formule, $\hat{y} \in]0;1[$ indique la "confiance" du modèle dans sa prédiction.

Dans la suite de ce document, nous utiliserons le risque suivant :

$$r(\hat{y}, y) = 1 - \delta_{\hat{y}y}$$

où δ_{ij} est le symbole de Kronecker et vaut 1 si $i = j$ et 0 sinon.

De manière transversale, ce risque élémentaire vaut 0 (ou du moins atteint sa valeur minimale) lorsque le modèle ne fait plus d'erreur.

2.2 Ensemble de fonctions, risque et minimisation

L'objectif de l'apprentissage est de trouver une fonction h qui minimiserait le risque élémentaire précédent "tout le temps". Nous allons formaliser ici ces idées.

Notons dans un premier temps $\mathcal{H}$ l'ensemble de toutes les fonctions qu'on s'autorise à considérer. Afin de généraliser la notion de risque élémentaire pour une fonction $h \in \mathcal{H}$, notons le risque dit de généralisation de la manière suivante :

$$R(h) = \mathbb{E}_{(X,Y)\sim f_{\mathcal{D}}}\big[r(h(X), Y)\big],\ h \in \mathcal{H} \tag{1}$$

Nous pouvons enfin définir notre objectif. Nous souhaitons trouver la fonction dans $\mathcal{H}$ telle que son risque de généralisation soit minimum.

$$f = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\, R(h) \tag{2}$$

Évidemment et malheureusement, ne serait-ce que parce que nous ne connaissons pas $f_{\mathcal{D}}$, nous ne pouvons pas utiliser cette stratégie.

2.3 Le risque empirique

Une stratégie alternative consiste à simuler selon $f_{\mathcal{D}}$. Dit autrement, il s'agit de construire un jeu de données $\mathcal{S} = \{(x_1, y_1), .., (x_N, y_N)\} \sim f_{\mathcal{D}}^N$.

Ce jeu de données nous permet d'évaluer le risque empiriquement en le moyennant sur chacun des éléments.

$$Re_{\mathcal{S}}(h) = \frac{1}{N}\sum_{i=1}^{N} r(h(x_i), y_i) \tag{3}$$

On peut bien évidemment associer cet estimateur à l'idée de loi des grands nombres ou bien de théorème central limite.

De la même manière que précédemment, nous allons utiliser $Re_{\mathcal{S}}$ afin de choisir une fonction dans $\mathcal{H}$:

$$\hat{f} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \, Re_{\mathcal{S}}(h) \tag{4}$$

Notons cependant ici que cet estimateur n'est plus sans biais.

3 Le gap de généralisation

Comme vu précédemment, dans les stratégies ERM, un jeu de données est utilisé afin d'estimer le risque associé à une fonction et l'objectif est de choisir la fonction qui le minimise. Cet stratégie n'est pas sans biais et il est probable que nous favorisions par ce procédé une fonction qui "s'adapte" aux données pour minimiser le risque plutôt qu'une fonction qui minimise le risque en espérance. Cette idée d'adaptation est ce qu'on appelle le *surapprentissage*. De manière plus formelle, on appelle le gap de généralisation l'écart en valeur absolue entre le risque en espérance et le risque empirique :

$$\text{gap}(h) = |Re_{\mathcal{S}}(h) - R(h)|. \tag{5}$$

En particulier, ce qui nous intéresse est le gap calculé pour la solution de la procédure d'optimisation $gap(\hat{f})$. On souhaite savoir si la fonction choisie fonctionnera bien générale ou bien n'a fait que s'adapter aux données. Évidemment, il suffit de collecter de nouvelles données afin d'avoir un estimateur sans biais des performances de notre fonction. Cependant, l'objectif ici est plutôt de déterminer s'il est crédible, en considérant $\mathcal{H}$ et N qu'on puisse effectivement avoir de bons résultats quitte à affiner certains éléments ou s'il faut revoir ces derniers.

4 Probablement approximativement correct

Un début de réponse est apporté par la théorie PAC dont nous allons illustrer quelques points ici. Cette théorie essaye de quantifier les éléments qui entrent en jeu dans le gap de généralisation en termes de $|\mathcal{H}|$ ainsi que de $|\mathcal{S}|$ notamment. Considérons en particulier, par simplicité, le cas où le risque minimum est 0 : $R(f) = 0$.

Définition 1 (PAC-learnable)**.** *Un ensemble de fonctions $\mathcal{H}$ est PAC-learnable s'il existe un algorithme $\mathcal{A}$ tels que $\forall \epsilon, \delta > 0$, $\forall f_{\mathcal{D}}$, et une fonction polynomiale g, alors pour $|\mathcal{S}| \geq g(1/\epsilon, 1/\delta, |\mathcal{H}|, f_{\mathcal{D}})$, nous avons pour $\hat{f} = \text{argmin}_{h \in \mathcal{H}} \, Re_{\mathcal{S}}(h)$:*

$$\mathbb{P}_{\mathcal{S} \sim f_{\mathcal{D}}}\big(R(\hat{f}) \leq \epsilon\big) \geq 1 - \delta \tag{6}$$

Dit autrement, un ensemble de fonctions est PAC-apprenable s'il existe un polynôme permettant de quantifier la taille du jeu de données qui permettraient de réduire avec forte probabilité le risque d'avoir un gap de généralisation trop grand.

En gardant à l'esprit que l'erreur minimale est 0, supposons que $|H|$ est de cardinal fini. Ce scénario se produit par exemple lorsqu'on fait de la sélection de modèles : on considère une quantité fini de paramètres et de modèles.

Théorème 1. *Soit $\mathcal{H}$ un ensemble de fonctions $h : \mathcal{X} \mapsto \mathcal{Y}$ de cardinal fini. Soit le risque défini comme $R(h) = \mathbb{P}_{x,y\sim f_{\mathcal{D}}}\Big(h(x) \neq y\Big)$. On note $\hat{f}_{\mathcal{S}} = \text{argmin}_{h\in\mathcal{H}}\, Re_{\mathcal{S}}(h)$ la solution de notre minimisation du risque empirique sur un échantillon i.i.d. $\mathcal{S} \sim {f_{\mathcal{D}}}^N$. Alors, $\forall \delta, \epsilon > 0$, ϵ, δ petits, l'affirmation suivante $\mathbb{P}_{f_{\mathcal{D}}}\big(\{\mathcal{S} : R(\hat{f}_{\mathcal{S}}) \leq \epsilon\}\big) \geq 1 - \delta$ est juste si :*

$$N \geq \frac{1}{\epsilon}\big(ln|\mathcal{H}| + ln\frac{1}{\delta}\big).$$

Une autre formulation peut en être déduite et nous permet de fournir une majoration de l'erreur de généralisation avec probabilité $1 - \delta$:

$$R(\hat{f}) \leq \frac{1}{N}\Big(ln|\mathcal{H}| + ln\frac{1}{\delta}\Big)$$

La dernière inégalité montre que plus le jeu de données sera grand, plus le risque sera avec forte probabilité proche de 0. Je rappelle encore une fois que le risque minimum atteignable est bien 0. Une formulation plus générale de cette idée nous dirait que plus le jeu de données est grand plus le risque se rapprocherait de son minimum dans $\mathcal{H}$.

Démonstration. Soit $\epsilon > 0$. Par hypothèse pour un échantillon $\mathcal{S}$, nous avons $Re_S(\hat{f}) = 0$. On souhaite minimiser la probabilité de tomber sur un jeu de données tel que la solution qui minimiserait le risque empirique serait mauvaise en espérance ($R(\hat{f}) > \epsilon$). Cet ensemble de jeu de données se note $\{S : R(\hat{f}_S) > \epsilon\}$.

Soit $\mathcal{H}_b = \{h \in \mathcal{H} : R(h) > \epsilon\}$ l'ensemble des mauvaises fonctions qui ont un risque en espérance supérieur à ϵ.

Soit $M = \{S : \exists h \in \mathcal{H}_b, Re_S(h) = 0\}$ l'ensemble des jeux de données qui pourraient résulter en une mauvaise solution (misleading examples). Plusieurs fonctions peuvent minimiser l'erreur sur un même jeu de données. On a évidemment :

$$\{S : R(\hat{f}_S) > \epsilon\} \subseteq M$$

La probabilité des jeux de données dont le résultat est une mauvaise fonction peut être majorée par la probabilité de M :

$$\mathbb{P}\big(\{S : R(\hat{f}_S) > \epsilon\}\big) \leq \mathbb{P}\big(M\big).$$

De plus,

$$M = \cup_{h\in\mathcal{H}_b}\{S : Re_S(h) = 0\},$$

Dit autrement, M est l'union pour chaque mauvaise fonction des jeux de données pour lesquelles cette fonction peut être une solution de la minimisation. On a donc,

$$\mathbb{P}\big(M\big) \leq \sum_{h\in\mathcal{H}_b} \mathbb{P}\big(\{S : Re_S(h) = 0\}\big)$$

Nous pouvons majorer la probabilité qu'un jeu de données soit bien séparée par une mauvaise fonction $h \in \mathcal{H}_b$:

$$\mathbb{P}\big(\{S : Re_S(h) = 0\}\big) = \prod_{i=1}^{N} \mathbb{P}\big(h(x_i) = y_i\big) = \prod_{i=1}^{N} 1 - R(h) \leq \prod_{i=1}^{N} (1-\epsilon) = (1-\epsilon)^N$$

En effet, par construction si $h \in \mathcal{H}_b$ alors $R(h) \geq \epsilon$.

On peut donc en conclure que la probabilité d'obtenir un jeu de données tel que la solution de minimisation du risque empirique ne marche pas en espérance est majorée de la manière suivante :

$$\mathbb{P}\big(\{S : R(\hat{f}_S) > \epsilon\}\big) \leq |\mathcal{H}_b|(1-\epsilon)^N \leq |\mathcal{H}|(1-\epsilon)^N$$

Cette équation donne un premier résultat intéressant. La probabilité d'erreur a un taux d'apprentissage exponentiel dans le sens où la taille du jeu de donnée réduit celle-ci exponentiellement vite. En réalité, à partir du moment où $|\mathcal{H}|$ est finie, le taux d'apprentissage est exponentiel !

En majorant la partie droite par δ, nous obtenons :

$$\begin{aligned} |\mathcal{H}|(1-\epsilon)^N &\leq \delta \\ \Leftrightarrow \ln|\mathcal{H}| + N\ln\,(1-\epsilon) &\leq \ln(\delta) \\ \Leftrightarrow N\,\ln(1-\epsilon) &\leq \ln(\delta) - \ln|\mathcal{H}| \\ \Leftrightarrow N &\geq \frac{1}{\ln(1-\epsilon)}\big(\ln(\delta) - \ln|\mathcal{H}|\big) \end{aligned}$$

Constatons par un développement de Taylor que $\ln(1-\epsilon) = \epsilon + o(\epsilon)$ et notons que $\ln(1-\epsilon) < \epsilon$ si $\epsilon > 0$. La formule de Taylor nous permet de constater que l'écart de majoration est petit pour ϵ petit. Nous obtenons ainsi :

$$N > \frac{1}{\epsilon}\big(\ln|\mathcal{H}| + \ln(\frac{1}{\delta})\big),$$

ce qui conclut la preuve. □

Bien entendu, un grand nombre de généralisations de ces idées existe, ne serait-ce qu'en considérant un cas où l'erreur minimale n'est pas 0. Évoquons également les cas où $|\mathcal{H}|$ n'est pas de cardinal fini comme c'est le cas de la plupart des modèles. Il convient alors d'utiliser d'autres "mesures" de la taille de l'ensemble comme la complexité de Rademacher ou encore la dimension Vapnik-Chervonenkis.