Introduction à la théorie de l'apprentissage statistique

Anass El yaagoubi
anass.el_yaagoubi@insa-rouen.fr

Valeria Petrov
valeria.petrov@insa-rouen.fr

July 6, 2018

Résumé

Ce document présente une introduction à la théorie de l'apprentissage statistique, nous commençons d'abord par introduire quelques inégalités de concentration très utiles dans de nombreux domaines tels que les statistiques et les processus aléatoires, puis nous formalisons le concept d'apprentissage statistique. L'étape suivante consiste à montrer la possibilité de garantir une bonne généralisation à savoir la convergence du risque empirique vers le vrai risque. Pour rédiger la dernière partie nous nous sommes inspiré du papier suivant [BBL04].

Table des matières

1 Introduction

La théorie de l'apprentissage est relativement jeune, nous pouvons retracer ses débuts à la fin des années cinquante. Elle est née comme une branche de l'intelligence artificielle, et c'est Arthur Samuel qui s'est intéressé alors à celle-ci et qui a utilisé pour la première fois le terme "Machine Learning" dans son papier "Some Studies in Machine Learning Using the game of Checkers". A partir des années soixante, Vladimir Vapnik et Alexey Chervonenkis ont développé l'approche statistique de l'apprentissage ainsi que le modèle d'apprentissage supervisé SVM (Machine a Vecteurs de Support). Aujourd'hui avec l'invention de nouvelles méthodes d'optimisation, et avec l'augmentation de la puissance de calcul des ordinateurs, les algorithmes que nous avons à notre disposition sont devenus très performants, ils sont même capables de rivaliser avec les êtres humains sur certaine tâches.

Le but de ce document sera de présenter les théories basiques de l'apprentissage statistique, qui permettent d'évaluer les capacités de prédictions de différents algorithmes automatiques. Nous essayerons de réutiliser au maximum les notations des papiers que nous citons afin de rester le plus claire possible.

2 Quelques inégalités de concentration

Des résultats comme la loi des grands nombres nous disent que les caractéristiques d'un échantillon aléatoire se rapproche des caractéristiques de la distribution d'origine au fur et à mesure que la taille de cet échantillon augmente. Une autres affirmation importante est le fait que le TCL nous dit que la moyenne empirique de variables aléatoires iid converge vers une loi Normale.

Ainsi, nous pouvons mesurer la probabilité que la moyenne empirique s'écarte de plus d'epsilon de la moyenne théorique. Mais ces résultats sont de nature asymptotique et donc pas facilement utilisables dans la pratique, nous n'avons jamais accès à des échantillons de taille infinie. Nous avons donc besoin de résultats supplémentaires.

Les inégalités de concentration donnent des résultats similaires mais avec des échantillons finis. Cependant les résultats de ces inégalités sont vrai avec haute probabilité, nous reviendrons sur cette notion de haute probabilité plus tard.

2.1 Inégalité de Markov

Soit X une variable aléatoire positive, alors $\forall\ \alpha > 0 : \mathbb{P}(X \geq \alpha) \leq \frac{E(X)}{\alpha}$

Cette inégalité nous donne une borne supérieure de la probabilité qu'une variable aléatoire réelle à valeurs positives soit supérieure ou égale à une constante positive α.

Preuve.

$$\begin{aligned} X &\geq \alpha \mathbb{1}_{X \geq \alpha} \\ \implies E(X) &\geq \alpha E(\mathbb{1}_{X \geq \alpha}) \\ or E(\mathbb{1}_{X \geq \alpha}) &= \mathbb{P}(X \geq \alpha) \\ \mathbb{P}(X \geq \alpha) &\leq \frac{E(X)}{\alpha} \end{aligned}$$

□

2.1.1 Corollaire

L'inégalité de Markov nous ramène au corrollaire suivant :

Soit ϕ une fonction strictement croissante sur R_+, alors $\{X \geq \alpha\} = \{\phi(X) \geq \phi(\alpha)\}$ c'est à dire $\mathbb{P}(X \geq \alpha) = \mathbb{P}(\phi(X) \geq \phi(\alpha))$ d'où :

$$\mathbb{P}(X \geq \alpha) \leq \frac{E(\phi(X))}{\phi(\alpha)}$$

2.2 Inégalité de Bienaymé-Tchebychev

Si $E(X)$ et $V(X)$ existent en prenant $|X - E(X)|$ comme variable aléatoire et en posant $\phi(x) = x^2$, on obtient l'inégalité suivante comme conséquence immédiate du corollaire précédent :

$$\mathbb{P}(|X - E(X)| \geq \alpha) \leq \frac{E(|X - E(X)|^2)}{\alpha^2}$$
$$\mathbb{P}(|X - E(X)| \geq \alpha) \leq \frac{V(X)}{\alpha^2}$$

Parfois l'inégalité de Bienaymé-Tchebychev peut être présentée sous la forme suivante en posant $\alpha = k\sigma$ et $V(X) = \sigma^2$:

$$\mathbb{P}(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$$

2.3 Inégalité de Chernoff

Cette inégalité est une conséquence directe du corollaire 2.1.1 En effet en prenant $\phi(x) = e^{sx}$ avec s positif, le corollaire nous donne le résultat suivant :

$$\mathbb{P}(X \geq \alpha) \leq \frac{E(e^{sX})}{e^{s\alpha}}$$

2.4 Inégalité de Hoeffding

Théorème 2.1 (Inégalité de Hoeffding)**.** *Soient X_1 ... X_n des variables aléatoires indépendantes centrées, si $\forall$ i on a que $a_i < X_i < b_i$ p.s. alors $\forall \epsilon > 0$:*

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i > \epsilon\Big) \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

Cette inégalité est fondamentale car c'est elle qui va nous permettre plus tard de trouver des bornes de généralisations pour des algorithmes d'apprentissage, pour la prouver nous allons d'abord prouver le lemme suivant.

2.4.1 Lemme de Hoeffding

Soit X une variable aléatoire centrée et bornée telle que $a < X < b$ p.s. alors $\forall$ s positif on a :

$$E(e^{sX}) \leq e^{\frac{s^2(b-a)^2}{8}}$$

Preuve.

$$\begin{aligned}
e^{sX} &\leq \frac{b - X}{b - a}e^{sa} + \frac{X - a}{b - a}e^{sb} && \text{par convexité de l'exponentielle} \\
E(e^{sX}) &\leq \frac{b}{b - a}e^{sa} - \frac{a}{b - a}e^{sb} && \text{en passant à l'espérance} \\
&= [\frac{b}{b - a} - \frac{a}{b - a}e^{s(b-a)}]e^{sa} && \text{par factorisation} \\
&= [1 - p + pe^u]e^{-up} && \text{en posant } u = s(b - a) \text{ et } p = -\frac{a}{b - a} \\
&= e^{ln(1-p+pe^u)-up} \\
&= e^{\phi(u)} && \text{où } \phi(u) = -up + ln(1 - p + pe^u)
\end{aligned}$$

Remarques :

$$
\begin{aligned}
\phi'(u) &= -p + \frac{pe^u}{1-p+pe^u} \\
\phi(0) &= \phi'(0) = 0 \\
\phi''(u) &= \frac{pe^u(1-p+pe^u) - pe^u pe^u}{(1-p+pe^u)^2} \\
&= \frac{pe^u(1-p)}{(1-p+pe^u)^2} \\
&= \frac{pe^u}{1-p+pe^u}(1 - \frac{pe^u}{1-p+pe^u}) \\
&= t(1-t) && \text{où } t = \frac{pe^u}{1-p+pe^u} \\
&\leq \frac{1}{4} && \forall t \in \mathbb{R} \\
\phi(u) &= \phi(0) + \phi'(0)u + \phi''(x)\frac{1}{2}u^2 && \text{par la formule de Taylor-Lagrange} \\
&\leq \frac{1}{8}u^2 \\
\phi(u) &\leq \frac{1}{8}u^2 \implies e^{\phi(u)} \leq e^{\frac{1}{8}u^2}
\end{aligned}
$$

Ainsi on a le résultat souhaité :

$$E(e^{sX}) \leq e^{\frac{1}{8}u^2} = e^{\frac{s^2(b-a)^2}{8}}$$

□

Preuve de l'inégalité de Hoeffding.

Soient X_1 ... X_n des variables aléatoires indépendantes, nous pouvons supposer sans perte de généralité que les X_i sont centrées car dans le cas contraire il suffit de prendre $Y_i = X_i - E(X_i)$ pour se ramener au cas centré, si $\forall$ i on a que $a_i < X_i < b_i$ p.s. alors $\forall \epsilon > 0$:

$$
\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} X_i > \epsilon) &\leq \frac{E(e^{s\sum_{i=1}^n X_i})}{e^{s\epsilon}} && \text{par l'inégalité de Chernoff} \\
&= \frac{E(\prod_{i=1}^n e^{sX_i})}{e^{s\epsilon}} \\
&= \frac{\prod_{i=1}^n E(e^{sX_i})}{e^{s\epsilon}} && \text{par indépendance des } X_i \\
&\leq \frac{\prod_{i=1}^n e^{\frac{s^2(b_i-a_i)^2}{8}}}{e^{s\epsilon}} && \text{par le Lemme de Hoeffding} \\
&= e^{-s\epsilon + s^2\sum_{i=1}^n \frac{(b_i-a_i)^2}{8}} \\
&= e^{P(s)} && \text{où } P(s) = As^2 - \epsilon s \text{ et } A = \sum_{i=1}^{n} \frac{(b_i-a_i)^2}{8}
\end{aligned}
$$

Ainsi, nous obtenons une majoration qui est vraie pour tout s positif, il est donc naturel de prendre le s qui minimise P, en effet ceci est possible car P est un polynôme

de degré deux avec A positif.

$$\begin{aligned} P'(s) &= -\epsilon + 2sA \\ P'(s^{opt}) = 0 &\iff s^{opt} = \frac{\epsilon}{2A} \\ P(s^{opt}) &= \frac{-\epsilon^2}{2A} + \frac{\epsilon^2}{(2A)^2}A \\ &= \frac{-\epsilon^2}{4A} \\ &= \frac{-2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2} \end{aligned}$$

Ainsi nous avons le résultat souhaité :

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i > \epsilon\Big) \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

□

2.5 Inégalité de Bennett

L'inégalité de Hoeffding ne donne pas les borne supérieur les plus performantes car elle n'utilise pas l'information sur la variance des variables aléatoires contrairement aux inégalités de Bennet et Bernstein.

Théorème 2.2 (Inégalité de Bennett)**.** *Soient $X_1 ... X_n$ des variables aléatoires indépendantes centrées ($E(X_i) = 0$), bornées ($|X_i| < M$) et de variance finie ($V(X_i) = \sigma^2 < \infty$), alors $\forall \epsilon > 0$:*

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i \geq \epsilon\Big) \leq e^{\frac{-n\sigma^2\phi(\frac{\epsilon M}{n\sigma^2})}{M^2}} \qquad \text{où } \phi(u) = (1+u)\log(1+u) - u$$

2.5.1 Lemme

$$E(e^{sX}) \leq e^{\frac{\sigma^2}{M^2}(e^{sM} - 1 - sM)}$$

Preuve du lemme.

Soit X une variable aléatoire centrée et bornée par M et de variance σ^2 alors :

$$
\begin{aligned}
E(e^{sX}) &= E\Big(\sum_{k=0}^{\infty} \frac{(sX)^k}{k!}\Big) && \text{par définition de l'exponentielle}\\
&= \sum_{k=0}^{\infty} \frac{s^k E(X^k)}{k!} && \text{par Théorème d'interversion série-intégrale}\\
&= 1 + \sum_{k=2}^{\infty} \frac{s^k E(X^k)}{k!} && \text{car X est centrée}\\
&= 1 + \sum_{k=2}^{\infty} \frac{s^k E(X^{k-2}X^2)}{k!} &&\\
&\leq 1 + \sum_{k=2}^{\infty} \frac{s^k \sigma^2 M^{k-2}}{k!} && \text{car } |X| < M \text{ et } E(X^2) = \sigma^2\\
&= 1 + \frac{\sigma^2}{M^2} \sum_{k=2}^{\infty} \frac{s^k M^k}{k!} && \text{par factorisation}\\
&= 1 + \frac{\sigma^2}{M^2}(e^{sM} - 1 - sM) && \text{par définition de l'exponentielle}\\
&\leq e^{\frac{\sigma^2}{M^2}(e^{sM}-1-sM)} && \text{car } \forall X, 1 + X \leq e^X
\end{aligned}
$$

□

Preuve de l'inégalité de Bennet.

$$
\begin{aligned}
\mathbb{P}\Big(\sum_{i=1}^{n} X_i \geq \epsilon\Big) &\leq \frac{E(e^{s\sum_{i=1}^{n} X_i})}{e^{s\epsilon}} && \text{Inégalité de Chernoff}\\
&\leq \frac{\prod_{i=1}^{n} E(e^{sX_i})}{e^{s\epsilon}} && \text{par indépendance}\\
&\leq \frac{\prod_{i=1}^{n} e^{\frac{\sigma^2}{M^2}(e^{sM}-1-sM)}}{e^{s\epsilon}} && \text{par le lemme précédent}\\
&= \frac{e^{\frac{n\sigma^2}{M^2}(e^{sM}-1-sM)}}{e^{s\epsilon}} &&\\
&= e^{\psi(s)} && \text{où } \psi(s) = -s\epsilon + \frac{n\sigma^2}{M^2}(e^{sM} - 1 - sM)
\end{aligned}
$$

Nous remarquons encore une fois que la majoration obtenue ne dépend pas de s, donc il est naturel de chercher le s qui donne la borne la plus petite possible, en effet la borne inférieure existe puisque ϕ est coercive :

$$
\begin{aligned}
\psi'(s) &= -\epsilon + \frac{n\sigma^2}{M^2}(Me^{sM} - M)\\
\psi'(s^{opt}) = 0 &\iff s^{opt} = \frac{1}{M} ln(\frac{\epsilon M}{n\sigma^2} + 1)\\
\implies \psi(s^{opt}) &= e^{-\epsilon \frac{1}{M} ln(\frac{\epsilon M}{n\sigma^2}+1) + \frac{n\sigma^2}{M^2}(\frac{\epsilon M}{n\sigma^2} - ln(\frac{\epsilon M}{n\sigma^2}+1))}\\
\implies e^{\psi(s^{opt})} &= e^{\frac{-n\sigma^2}{M^2}[\frac{\epsilon M}{n\sigma^2} ln(\frac{\epsilon M}{n\sigma^2}+1) - (\frac{\epsilon M}{n\sigma^2} - ln(\frac{\epsilon M}{n\sigma^2}+1))]}\\
&= e^{\frac{-n\sigma^2}{M^2}[(\frac{\epsilon M}{n\sigma^2}+1) ln(\frac{\epsilon M}{n\sigma^2}+1) - \frac{\epsilon M}{n\sigma^2}]}
\end{aligned}
$$

Ainsi, nous retrouvons le résultat souhaité :

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i \geq \epsilon\Big) \leq e^{\frac{-n\sigma^2}{M^2}\phi(\frac{\epsilon M}{n\sigma^2})} \qquad \text{où } \phi(u) = (u+1)ln(u+1) - u$$

□

2.6 Inégalité de Bernstein

Théorème 2.3 (Inégalité de Bernstein)**.** *Soit X_1 ...X_n des variables aléatoires indépendantes, bornées, centrées et de variance σ^2, alors $\forall \epsilon > 0$.*

$$\mathbb{P}\Big(\sum_{i=1}^{n} \frac{1}{n} X_i \geq t\Big) \leq e^{\frac{-nt^2}{2\sigma^2 + \frac{2Mt}{3}}}$$

L'inégalité de Bernstein, nous disent que la probabilité pour qu'une somme de variables aléatoires indépendantes soit à distance t de sa moyenne est exponentielle faible.

Preuve.

Soient :

$$\phi(x) = (1+x)ln(1+x) - x$$
$$\psi(x) = \frac{3x^2}{2(x+3)}.$$

Remarque :

$$\begin{aligned} \forall k \geq 0, \quad & \psi^{(k)}(0) \leq \phi^{(k)}(0) \\ \Longrightarrow \quad & \psi(x) \leq \phi(x) \end{aligned}$$

$$\begin{aligned} \mathbb{P}\Big(\sum_{i=1}^{n} X_i \geq \epsilon\Big) &\leq e^{\frac{-n\sigma^2}{M^2}\phi(\frac{\epsilon M}{n\sigma^2})} && \text{Inégalité de Bennet} \\ &\leq e^{\frac{-n\sigma^2}{M^2}\psi(\frac{\epsilon M}{n\sigma^2})} && \text{par l'inégalité précédente} \end{aligned}$$

Ainsi on obtient le résultat souhaité en prenant nt à la place d'epsilon.

□

2.7 Inégalité de McDiarmid

Soient $X_1, ..., X_n$ des variables aléatoires indépendantes qui sont à valeur dans $\mathcal{X}$ et $f : \mathcal{X}^n \to \mathbb{R}$ une fonction de $X_1, ..., X_n$.

Si $|f(x_1, ..., x_i, ..., x_n) - f(x_1, .., x_i', ..., x_n)| \leq c_i \ \forall i, \forall x_1, ..., x_n, x_i' \in \mathcal{X}$ alors:

$$\mathbb{P}(f - \mathbb{E}(f) \geq \epsilon) \leq exp\Big(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\Big)$$

Preuve de l'inégalité de McDiarmid. Dans la suite X_1^i dénotera $X_1, ...X_i$.

Posons $V_i = \mathbb{E}(f|X_1^i) - \mathbb{E}(f|X_1^{i-1})$

Il s'en suit alors que $\sum_{i=1}^{n} V_i = f - \mathbb{E}(f)$

Posons $L_i = \inf_{x_i} V_i$ et $U_i = \sup_{x_i} V_i$

Ainsi on a que $L_i \leq V_i \leq U_i$

Donc $U_i - L_i \leq c_i$ par hypothèse sur f

$$
\begin{aligned}
\mathbb{P}(f - \mathbb{E}(f) \geq \epsilon) &= \mathbb{P}\Big(\sum_{i=1}^{n} V_i \geq \epsilon\Big) && \text{par définition des } V_i \\
&= \mathbb{P}\Big(e^{s(\sum_{i=1}^{n} V_i)} \geq e^{s\epsilon}\Big) && \text{car l'exponentielle est bijective (s positif)} \\
&\leq \frac{\mathbb{E}\big(e^{s(\sum_{i=1}^{n} V_i)}\big)}{e^{s\epsilon}} && \text{inégalité de Markov} \\
&= \frac{\mathbb{E}\big(\prod_{i=1}^{n} e^{sV_i}\big)}{e^{s\epsilon}} \\
&= \frac{\mathbb{E}\mathbb{E}\big(\prod_{i=1}^{n} e^{sV_i} | X_1^{n-1}\big)}{e^{s\epsilon}} \\
&= \frac{\mathbb{E}\big(\prod_{i=1}^{n-1} e^{sV_i}\big)\mathbb{E}\big(e^{sV_n} | X_1^{n-1}\big)}{e^{s\epsilon}} \\
&\leq \frac{\mathbb{E}\big(\prod_{i=1}^{n-1} e^{sV_i}\big)e^{s^2 c_n^2/8}}{e^{s\epsilon}} && \text{par le lemme de Hoeffding} \\
&\vdots \\
&\leq \frac{e^{s^2 \sum_{i=1}^{n} \frac{c_i^2}{8}}}{e^{s\epsilon}} && \text{par le lemme de Hoeffding} \\
&= e^{-s\epsilon + s^2 \sum_{i=1}^{n} \frac{c_i^2}{8}}
\end{aligned}
$$

Finalement nous procédons de la même manière que pour démontrer Hoeffding, c'est-à-dire de choisir le s qui minimise la borne à droite. □

Remarque : En posant $f(x_1, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$, nous pouvons retrouver l'inégalité de Hoeffding comme un cas particulié de McDiarmid.

3 Théorie de l'apprentissage statistique

L'objectif principal de la théorie de l'apprentissage statistique est de fournir un cadre mathématique rigoureux, permettant l'étude de certains problèmes d'inférence. C'est-à-dire qu'à partir d'un ensemble de données le but est de trouver un modèle qui explique au mieux un certain phénomène. Ce modèle pourra ensuite être utilisé afin de faire des prédictions. Dans la suite, nous allons nous intéresser tout particulièrement aux problèmes de classification dans le cadre de l'apprentissage supervisé.

Nous aurons des données qui proviennent de $\mathcal{X}\mathrm{x}\mathcal{Y}$ où $\mathcal{X}$ est typiquement R^d est l'espace des entrées ("Input Space") et $\mathcal{Y}$ représente l'espace des labels, c'est-à-dire les classes aux quelles appartiennent les points de $\mathcal{X}$. Dans la suite $\mathcal{Y}$ désignera souvent {-1,1} c'est à dire appartenir ou non à une catégorie. L'objectif est donc de trouver à partir des données une fonction $g \in \mathcal{G} = \{g : \mathcal{X} \to \mathcal{Y}\}$ permettant de classifier tous les points de $\mathcal{X}$.

3.1 Formalisation du problème

Dans la suite nous verrons nos données $\{(X_i, Y_i) \in \mathcal{X}\mathrm{x}\mathcal{Y}\}_{i=1}^{n}$ comme une suite de variables aléatoires iid issues d'une distribution inconnue P.

Définition 3.1. *Risque*
Soit $g \in \mathcal{G}$, alors le risque R de g noté R(g) = P(g(X)≠Y)

Définition 3.2. *Risque empirique*
Soit $g \in \mathcal{G}$, alors le risque empirique R_n de g noté $R_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g(X_i)\neq Y_i}$

Afin de trouver le meilleur classificateur de nos données il semble judicieux de prendre celui qui minimise la probabilité d'erreur, il est donc naturel de prendre le risque R comme critère de sélection. Donc nous pouvons écrire notre problème sous la forme suivante:

$$g^{opt} = arg\min_{g\in\mathcal{G}} R(g)$$

Cette idée de minimiser le risque est très attrayante, cependant afin de mesurer le risque nous devons connaître la distribution P. Dans ce cas notre problème n'est pas bien compliqué car il suffirait de prendre le classificateur de Bayes,étant celui qui présente le plus petit risque:

$$\begin{aligned} t(x) &= signe(E(Y = 1|X = x)) \\ &= signe(P(Y = 1|X = x) - P(Y = -1|X = x)) \\ &= signe(2P(Y = 1|X = x) - 1) \end{aligned}$$

C'est pour cela que dans la pratique, on utilise plus tôt le risque empirique :

$$g_n = arg\min_{g\in\mathcal{G}} R_n(g)$$

Cette approche est appelée minimisation du risque empirique, mais il existe d'autres approches comme par exemple la minimisation du risque structurel ou les méthodes de régularisation.

Remarque : Si nous cherchons g dans une classe trop grande, il est possible de l'obtenir sous certaines hypothèses que $R_n(g) = 0$ mais $R(g) = 1$. Ainsi, notre classificateur prédit juste, seulement les données que nous avions dans l'ensemble d'apprentissage mais généralise très mal. Ceci est un cas extrême de ce que l'on appelle le sur-apprentissage. Il est donc primordiale de bien choisir $\mathcal{G}$.

3.2 Bornes de généralisation

Maintenant que nous avons à notre disposition une stratégie pour trouver le bon classificateur pour la minimisation du risque empirique, nous allons voir quel type de garanties nous allons pouvoir obtenir avec cette stratégie.

Un bon classificateur est celui qui minimise le risque, c'est-à-dire la probabilité de mal classifier de futures données provenant de $\mathcal{X}$. Mais le problème est que nous ne pouvons connaître R dué à son dèpendence de P. Nous avons donc besoin de lier le risque empirique R_n avec le vrai risque R, c'est de là que naît le besoin d'avoir des bornes de généralisation.

Nous pouvons écrire le vrai risque comme ceci :

$$R(g) = R_n(g) + R(g) - R_n(g)$$

Ainsi nous obtenons l'inégalité suivante :

$$R(g) \leq R_n(g) + |R(g) - R_n(g)|$$

Notons que $|R(g) - R_n(g)|$ représente l'écart entre le vrai risque et le risque empirique, il est donc nécessaire d'analyser cette quantité afin de savoir dans quels cas nous pouvons avoir $|R(g) - R_n(g)| > \epsilon$. En effet quand nous avons un petit risque empirique, si l'écart avec le vrai risque est trop grand, la stratégie qui consiste en la minimisation du risque empirique est complètement inutile car nous avons aucune garantie contre le sur-apprentissage.

3.2.1 Quelques notations

Pour un ensemble $\mathcal{G}$ donné, $\mathcal{F} = \{f : (x, y) \to \mathbb{1}_{g(x)\neq y} : g \in \mathcal{G}\}$. Remarquons que ces deux ensembles sont en bijection. Nous utiliserons Pf pour dénoter $E(f(x, y))$ et P_nf pour dénoter $\frac{1}{n}\sum_{i=1}^{n} f(x_i, y_i)$, aussi au lieux de (x_i, y_i) nous utiliserons z_i.

3.2.2 Déviation du risque empirique

Grâce à la notation introduite préalablement, nous pouvons réécrire l'écart entre le risque empirique et le vrai risque :

$$\begin{aligned} R(g) - R_n(g) &= E(f(z)) - \frac{1}{n}\sum_{i=1}^{n} f(z_i) \\ &= Pf - P_n f \end{aligned}$$

Rappel : Loi faible et loi forte des grands nombres

Soit $(X_n)_{n\in N^*}$ une suite de v.a. iid de moyenne μ et de variance σ^2.

Loi faible :

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) = 0$$

Loi forte :

$$\mathbb{P}(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu) = 1$$

Ces deux lois nous disent que la moyenne empirique converge presque sûrement vers la moyenne théorique quand les variables aléatoires sont iid et ont une variance finie. Dans notre cas, cela veut dire que presque sûrement nous avons :

$$\lim_{n\to\infty} P_n f \to Pf$$

Autrement dit plus l'échantillon est grand plus le risque empirique se rapproche du vrai risque. Ce résultat est rassurant mais reste asymptotique et en pratique il n'est pas très utile, de plus cela ne donne pas d'indications quant à la vitesse de convergence.

Cependant en supposant que nos données suivent une certaine loi de probabilité, en faisant jouer à $\frac{1}{n}(f(z_i) - E(f(z_i)))$ le rôle de X_i, l'inégalité de Hoeffding conduit au résultat suivant :

$$\mathbb{P}(|P_n f - Pf| > \epsilon) \leq 2\exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

En prenant la partie droit de l'inégalité égale à δ, en isolant ϵ, comme f est à valeur dans {0,1} on obtient :

$$\mathbb{P}\left[|P_n f - Pf| > \sqrt{\frac{ln(\frac{2}{\delta})}{2n}}\right] \leq \delta$$

Autrement dit, nous avons le résultat suivant avec probabilité au moins $1-\delta$:

$$|P_n f - Pf| \leq \sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$$

Puisque à chaque $f \in \mathcal{F}$ correspond un $g \in \mathcal{G}$, nous obtenons ainsi le résultat fondamental suivant avec probabilité au moins $1-\delta$:

$$R(g) \leq R_n(g) + \sqrt{\frac{ln(\frac{2}{\delta})}{2n}}$$

Remarquons qu'en utilisant de différentes inégalités de concentration nous obtiendrons des bornes supérieures de l'erreur de généralisation qui sont différentes, ainsi en appliquant l'inégalité de Bienaymé-Tchebychev à $|P_n f - Pf|$ on a:

$$\begin{aligned}\mathbb{P}(|P_n f - Pf| > \epsilon) &\leq \frac{1}{n}\frac{V(f(z_i))}{\epsilon^2} \\ &= \frac{1}{n}\frac{\sigma^2}{\epsilon^2}\end{aligned}$$

De manière analogue à ce qui précède, nous obtenons donc avec probabilité au moins $1-\delta$:

$$R(g) \leq R_n(g) + \frac{\sigma}{\sqrt{n\delta}}$$

3.3 Vitesse de convergence

Dans ce qui précède nous avons remarqué qu'en utilisant différentes inégalités de concentration nous avions obtenu des majorations de l'erreur de généralisation différentes. Pour les deux majorations obtenus précédemment nous allons analyser la vitesse de convergence du risque empirique vers le vrai risque.

Avec l'inégalité de Bienaymé-Tchebychev dans le pire des cas nous avons que:

$$\begin{aligned}|R(g) - R_n(g)| &= \frac{\sigma}{\sqrt{n\delta}} \\ \implies \frac{|R(g) - R_{n+1}(g)|}{|R(g) - R_n(g)|} &= \frac{\frac{\sigma}{\sqrt{(n+1)\delta}}}{\frac{\sigma}{\sqrt{n\delta}}} \\ &= \sqrt{\frac{n}{n+1}} \to 1\end{aligned}$$

Ce qui donne une convergence lente du risque empirique.

Avec l'inégalité de Hoeffding dans le pire des cas nous avons que :

$$\begin{aligned}|R(g) - R_n(g)| &= \sqrt{\frac{ln(\frac{2}{\delta})}{2n}} \\ \implies \frac{|R(g) - R_{n+1}(g)|}{|R(g) - R_n(g)|} &= \frac{\sqrt{\frac{ln(\frac{2}{\delta})}{2(n+1)}}}{\sqrt{\frac{ln(\frac{2}{\delta})}{2n}}} \\ &= \sqrt{\frac{n}{n+1}} \to 1\end{aligned}$$

Dans les deux cas nous avons les mêmes vitesses de convergence. Nous pouvons nous poser la question s'il est possible d'obtenir de meilleures vitesses de convergence en utilisant d'autres bornes de concentrations plus performantes.

Une autre interprétation que nous pouvons faire également avec les bornes obtenues concerne la taille de l'échantillon nécessaire afin de garantir une généralisation pour ϵ et δ donnés. Où ϵ représente l'écart entre le vrai risque et le risque empirique et δ la probabilité que cet écart soit plus grand que ϵ.

$$\begin{aligned}\epsilon &= \sqrt{\frac{ln(\frac{2}{\delta})}{2n}} \\ \implies n &= \frac{1}{2\epsilon^2} ln(\frac{2}{\delta})\end{aligned}$$

Nous remarquons que la taille de l'échantillon est polynomiale en $\frac{1}{\epsilon}$ et logarithmique en $\frac{1}{\delta}$. Ce qui veut dire que c'est beaucoup plus dure de garantir un ϵ petit pour un δ donné que l'inverse.

3.4 Borne uniforme pour des espaces d'hypothèses finis

Dans ce qui précède nous avons borné l'erreur de généralisation pour une fonction $f \in \mathcal{F}$ donnée. Or à priori nous ne savons pas quelle fonction sera retournée par notre algorithme d'apprentissage.

Si $\mathcal{F}$ est de taille finie et que $|\mathcal{F}| = N$ nous pouvons alors donner une borne de généralisation uniforme, c'est-à-dire une borne qui est valable pour toutes les fonctions de $\mathcal{F}$ ainsi :

$$\begin{aligned}
\mathbb{P}\Big(\exists\, f \in \mathcal{F}, P_n f - Pf > \epsilon\Big) &= \mathbb{P}\Big(\bigcup_{f\in\mathcal{F}} P_n f - Pf > \epsilon\Big) \\
&\leq \sum_{f\in\mathcal{F}} \mathbb{P}\Big(P_n f - Pf > \epsilon\Big) \\
&\leq N \exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)
\end{aligned}$$

Ainsi nous obtenons que :

$$R(g) \leq R_n(g) + \sqrt{\frac{ln(N) + ln(\frac{1}{\delta})}{2n}}$$

3.5 Borne uniforme pour des espaces d'hypothèses infinis

3.5.1 Cas dénombrable

Supposons que $\mathcal{F}$ est infini dénombrable. Dans ce cas nous pouvons associer une loi discrète p à $\mathcal{F}$ c'est à dire $\forall f \in \mathcal{F}$, $p(f) \in [0,1]$ et $\sum_{f\in\mathcal{F}} p(f) = 1$. L'idée maintenant est d'utiliser l'information concernant la vraisemblance qu'une fonction f soit retournée par notre algorithme afin d'améliorer les bornes supérieures, ainsi nous avons avec l'inégalité de Hoeffding :

$$\begin{aligned}
\mathbb{P}\left[P_n f - Pf > \sqrt{\frac{ln\frac{1}{\delta(f)}}{2n}}\right] &\leq \delta(f) \\
&= \delta p(f) \\
\implies \mathbb{P}\left[\exists\, f \in \mathcal{F}, P_n f - Pf > \sqrt{\frac{ln\frac{1}{\delta(f)}}{2n}}\right] &\leq \sum_{f\in\mathcal{F}} \delta p(f) \\
&= \delta \sum_{f\in\mathcal{F}} p(f) \\
&= \delta
\end{aligned}$$

Ainsi nous avons le résultat suivant avec probabilité au moins $1-\delta$:

$$\forall f \in \mathcal{F}, Pf \leq P_n f + \sqrt{\frac{ln\frac{1}{p(f)} + ln\frac{1}{\delta}}{2n}}$$

Remarquons que nous pouvons retrouver le résultat dans le cas fini de la section précédente en prenant une loi uniforme avec $|\mathcal{F}| = N$ et donc $p(f) = \frac{1}{N}$.

Dans la suite nous allons généraliser ce résultat au cas non-dénombrable mais pour cela nous avons besoin de définir deux nouveaux concepts à savoir fonction de croissance et $VCdimension$.

3.5.2 Fonction de croissance et VC dimension

Définition 3.3. *Fonction de croissance $S_{\mathcal{F}}$*
La fonction de croissance $S_{\mathcal{F}}$ compte le nombre maximum de façon différentes dont peuvent être classés n points par un ensemble de fonction $\mathcal{F}$ donné. D'une manière plus précise posons $\mathcal{F}_{z_1,..,z_n} = \{(f(z_1), ..., f(z_n)) : f \in \mathcal{F}\}$, nous avons :

$$S_{\mathcal{F}}(n) = \sup_{(z_1,...,z_n)} |\mathcal{F}_{z_1,..,z_n}|$$

Comme les f sont à valeur dans {0,1} la fonction de croissance sera toujours fini, c'est-à-dire que $S_{\mathcal{F}}(n) \leq 2^n$. Remarquons que $\forall n \in \mathbb{N}$, $S_{\mathcal{F}}(n) = S_{\mathcal{G}}(n)$.

Définition 3.4. *VC dimension*
Lorsque $S_{\mathcal{F}}(n) = 2^n$ alors il existe un ensemble de n points qui est classifié de toutes les manières possibles, disons alors que $\mathcal{F}$ brise ou "shatters" cet ensemble. La $VCdimension$ pour une classe de fonctions $\mathcal{F}$ est le plus grand n pour lequel $S_{\mathcal{F}}(n) = 2^n$.

Exemple : VC dimension des demis espaces de $\mathbb{R}^n$
Considérons l'ensemble des demis espaces de $\mathbb{R}^n$ comme ensemble d'hypothèses, c'est-à-dire $\mathcal{F} = \{signe(w^T x) | w \in \mathbb{R}^n\}$. Essayons donc de trouver la $VCdimension$ de cette classe de fonctions. Prenons $\{e_1, ..., e_n\}$ comme données d'apprentissage où e_i représente un vecteur de zéros avec un un à la i-ème position. Ainsi $signe(w^T x) = signe(w_i)$ nous avons ainsi c $S_{\mathcal{F}}(n) = 2^n$ et par conséquent $VCdimension$ de $\mathcal{F}$ supérieur ou égale à n. Prenons un ensemble $\{z_1, ..., z_{n+1}\}$ de $n+1$ vecteurs de $\mathbb{R}^n$, comme $\mathbb{R}^n$ est de dimension n la famille $\{z_1, ..., z_{n+1}\}$ est obligatoirement liée, c'est-à-dire que pour un certain i $z_i = \sum_{j \neq i} \alpha_j z_j$. Ainsi $signe(w^T z_i) = signe(w^T(\sum_{j \neq i} \alpha_j z_j)) = signe(\sum_{j \neq i} \alpha_j w^T z_j)$, ce qui veut dire que la classification du vecteur z_i dépend de la classification des autres vecteurs et par conséquent $S_{\mathcal{F}}(n+1) < 2^{n+1}$ et donc $VCdimension$ de $\mathcal{F} = n$.

3.5.3 Théorème Vapnik-Chervonenkis

Théorème 3.1 (Vapnik-Chervonenkis)**.** *Pour tout $\delta > 0$ nous avons le résultat suivant avec probabilité au moins $1 - \delta$:*

$$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + 2\sqrt{2\frac{ln[S_{\mathcal{F}}(2n)] + ln\frac{2}{\delta}}{n}}$$

Lemme : Symétrisation Pour tout $t > 0$ tel que $nt^2 \geq 2$,

$$\mathbb{P}\Big[\sup_{f \in \mathcal{F}}(P - P_n)f \geq t\Big] \leq 2\mathbb{P}\Big[\sup_{f \in \mathcal{F}}(P'_n - P_n)f \geq t/2\Big]$$

La preuve de ce lemme peut être retrouvée à la section 4.4 de [BBL04]. La preuve utilise quelques astuces et manipulations algébriques de fonctions indicatrices, elle utilise aussi le fait qu'une variable aléatoire à valeur dans $[0, 1]$ admet une variance inférieure à $1/4$. L'intérêt de ce lemme est qu'il permet de se débarrasser de l'espérance Pf en la remplaçant par une moyenne empirique en introduisant un ensemble de données fantôme.

Preuve.

$$\mathbb{P}\Big[\sup_{f\in\mathcal{F}}(P-P_n)f \geq t\Big] \leq 2\mathbb{P}\Big[\sup_{f\in\mathcal{F}}(P'_n-P_n)f \geq t/2\Big] \tag{1}$$

$$= 2\mathbb{P}\Big[\sup_{\mathcal{F}_{z_{1:n},z'_{1:n}}}(P'_n-P_n)f \geq t/2\Big] \tag{2}$$

$$\leq 2|\mathcal{F}_{z_{1:n},z'_{1:n}}|\mathbb{P}\Big[(P'_n-P_n)f \geq t/2\Big] \tag{3}$$

$$\leq 2\mathcal{S}_{\mathcal{F}}(2n)\mathbb{P}\Big[(P'_n-P_n)f \geq t/2\Big] \tag{4}$$

$$\leq 4\mathcal{S}_{\mathcal{F}}(2n)e^{-nt^2/8} \tag{5}$$

$$= \delta \tag{6}$$

$$\implies t = 2\sqrt{2\frac{ln[\mathcal{S}_{\mathcal{F}}(2n)] + ln\frac{4}{\delta}}{n}} \tag{7}$$

L'étape (2) est valide car le sup ne dépend que de la projection de f sur les données. A l'étape (3) nous faisons une borne d'union. A l'étape (4) nous utilisons la définition de la fonction de croissance et finalement à la dernière étape nous appliquons l'inégalité de Hoeffding deux fois. Par suite le résultat suivant est obtenu avec probabilité au moins $1-\delta$:

$$\forall f \in \mathcal{F}, Pf \leq P_n f + 2\sqrt{2\frac{ln[\mathcal{S}_{\mathcal{F}}(2n)] + ln\frac{4}{\delta}}{n}}$$

□

4 Conclusion

Nous avons donc commencé par présenter quelques inégalités de concentrations avec leur preuves, puis nous avons formaliser le concept d'apprentissage statistique grâce à la notion de borne de généralisation, tout d'abord dans le cas où l'espace d'hypothèse était fini puis nous avons généraliser au cas infini grâce à la notion de dimension VC. Comme le but de ce document était d'introduire le lecteur aux concepts de la théorie de l'apprentissage, nous avons omis un certains nombre de détails pourtant très intéressants ...

Ce document peut se développer(et sera avec haute probabilité) pour contenir et pour présenter d'autres concepts liés à l'apprentissage automatique, des notions telles que l'apprentissage ensembliste, l'analyse de la stabilité des algorithmes d'apprentissage statistique, l'apprentissage statistique du point de vue de la théorie de la complexité mais également une revue des principales méthodes d'optimisation.

References

[BBL04] O. Bousquet, S. Boucheron, and G. Lugosi. *Introduction to Statistical Learning Theory*, volume Lecture Notes in Artificial Intelligence 3176, pages 169–207. Springer, Heidelberg, Germany, 2004.

[BLB04] S. Boucheron, G. Lugosi, and O. Bousquet. *Concentration Inequalities*, volume Lecture Notes in Artificial Intelligence 3176, pages 208–240. Springer, Heidelberg, Germany, 2004.

[KH11] Sanjeev R. Kulkarni and Gilbert Harman. Statistical learning theory: a tutorial. 2011.

[Sri] Karthik Sridharan. A gentle introduction to concentration inequalities.

[KH11] [BBL04] [BLB04] [Sri]