Une théorie probabiliste de l'apprentissage supervisé de similarité pour l'optimisation en un point de la courbe ROC

Robin Vogel [1] & Stéphan Clémençon [2] & Aurélien Bellet [3]

[1] *IDEMIA, 11 Boulevard Gallieni, 92130 Issy-les-Moulineaux, robin.vogel@idemia.com*
[2] *Télécom ParisTech, 46 Rue Barrault, 75013 Paris, stephan.clemencon@telecom-paristech.fr*
[3] *Inria Lille, 40 avenue Halley, 59650 Villeneuve d'Ascq, aurelien.bellet@inria.fr*

Résumé. La performance de beaucoup de méthodes d'apprentissage statistique dépend du choix d'une métrique adéquate sur l'espace d'entrée. L'apprentissage de similarité (ou apprentissage de métrique) vise à construire une telle fonction à partir de données d'entraînement de manière à ce que les observations associées à la même (resp. à différentes) classe(s) soient aussi proches (resp. éloignées) que possible. Dans cet article, l'apprentissage de similarité est étudié en tant qu'ordonnancement biparti de paires d'observations, dont l'objectif est de ranger les éléments d'une base de données dans l'ordre décroissant de leur probabilité d'être dans la même classe qu'une donnée requête, en utilisant les scores de similarité. Un critère de performance naturel est alors l'optimisation en un point de la courbe ROC, qui consiste à maximiser le taux de vrais positifs sous un taux de faux positifs fixé. Nous étudions cette nouvelle perspective sur l'apprentissage de similarité avec une formulation probabiliste rigoureuse. La version empirique de ce problème induit un problème d'optimisation sous contrainte mettant en jeu des *U-statistiques*, pour lequel nous dérivons des vitesses d'apprentissage universelles ainsi que des vitesses rapides sous une hypothèse de bruit sur la distribution des données. Nous adressons aussi le problème de mise à l'échelle en analysant l'effet d'approximations basées sur des méthodes d'échantillonnage. Nos résultats théoriques sont illustrés par des expériences numériques.

Mots-clés. Apprentissage et classification, apprentissage de métrique, ranking.

Abstract. The performance of many machine learning techniques depends on the choice of an appropriate similarity or distance measure on the input space. Similarity learning (or metric learning) aims at building such a measure from training data so that observations with the same (resp. different) label are as close (resp. far) as possible. In this paper, similarity learning is investigated from the perspective of pairwise bipartite ranking, where the goal is to rank the elements of a database by decreasing order of the probability that they share the same label with some query data point, based on the similarity scores. A natural performance criterion in this setting is pointwise ROC optimization: maximize the true positive rate under a fixed false positive rate. We study this novel perspective on similarity learning through a rigorous probabilistic framework. The empirical version of the problem gives rise to a constrained optimization formulation involving U-statistics, for which we derive universal learning rates as well as faster rates under a noise assumption on the data distribution. We also address the large-scale setting by analyzing the effect of sampling-based approximations. Our theoretical results are supported by illustrative numerical experiments.

Keywords. Machine learning and classification, metric learning, ranking.

1 Introduction

Les fonctions de similarité jouent un rôle clé dans beaucoup d'algorithmes d'apprentissage statistique, pour des problèmes allant de la classification à la réduction de dimension en passant par le partitionnement de données. La performance de ces méthodes dépend fortement de la pertinence de la fonction de similarité concernant la tâche et les données considérées. Ce constat a motivé la recherche en apprentissage de

métrique et de similarité, qui consiste à apprendre automatiquement une fonction de distance ou similarité à partir de données, voir [Bellet et al., 2015] pour une revue de la littérature.

Dans cet article, nous étudions l'apprentissage de similarité en tant qu'ordonnancement biparti de paires d'observations, dont l'objectif est de ranger les éléments d'une base de données par ordre décroissant de la probabilité qu'ils soient dans la même classe qu'une donnée requête. Ce problème est motivé par des applications concrètes. Par exemple, l'identification biométrique cherche à vérifier l'identité d'un individu en comparant son information biométrique (comme une photo de son visage) avec une large base de données contenant celle de personnes autorisées. Sachant une fonction de similarité et un seuil, les éléments de la base de données sont rangés dans l'ordre décroissant de leur score de similarité avec la requête, et les éléments considérés correspondants sont ceux dont le score de similarité avec la requête dépasse le seuil. Les critères de performance pour ce problème sont généralement liés à la courbe ROC associée à la fonction de similarité, la courbe ROC représentant le taux de vrai positifs sachant le taux de faux positifs. Les approches précédentes cherchent à optimiser l'aire sous la courbe ROC empirique, cependant nous considérons dans ce travail l'optimisation en un point de la courbe ROC, qui consiste à maximiser le taux de vrais positifs sous un taux de faux positifs fixé. Cet objectif, écrit comme un problème d'optimisation contrainte, exprime naturellement les contraintes opérationnelles présentes dans beaucoup d'applications pratiques. Par exemple, les systèmes de vérification biométrique sont généralement paramétrés pour fonctionner à un taux de faux positifs fixé.

En plus de proposer un cadre probabiliste rigoureux pour étudier cette perspective nouvelle sur l'apprentissage de similarité, nous faisons les contributions suivantes. Dans un premier temps, des vitesses d'apprentissage universelles - c'est-à-dire sans hypothèses sur la distribution des données - de l'ordre de $O(1/\sqrt{n})$, avec n le nombre de données, sont démontrées pour l'apprentissage d'un problème empirique d'optimisation de la courbe ROC. Nous démontrons ensuite l'existence de vitesses rapides sous une hypothèse de bruit de même nature que la condition dite de Mammen-Tsybakov, voir [Mammen and Tsybakov, 1995a], en montrant une vitesse d'apprentissage de l'ordre de $O(n^{-(2+a)/4})$, avec $a \in (0, 1)$ un paramètre de bruit. De tels résultats n'existent pas, à notre connaissance, pour l'apprentissage de similarité contraint. Étant donné que les quantités manipulées ne sont pas des moyennes d'observations i.i.d. mais s'écrivent sous la forme de U-statistiques, nos résultats s'appuient sur des inégalités de concentration pour les U-processus. Les vitesses rapides ont été mises en évidence empiriquement par des simulations numériques, ce qui est rarement trouvé dans la littérature sur les vitesses d'apprentissage rapides.

Dans un second temps, nous adressons les problèmes de mise à l'échelle liés au très grand nombre de paires négatives dans ce problème quand le jeu de données et le nombre de classes est grand. En particulier, nous montrons qu'en utilisant une approximation du taux de faux positifs utilisant un nombre de paires sélectionnées aléatoirement de l'ordre de $O(n)$, appelée U-statistique incomplète, voir [Lee, 1990], nous préservons l'ordre en $O(1/\sqrt{n})$ de notre vitesse d'apprentissage universelle. Deux méthodes d'échantillonnage sont analysées, et nous présentons des conditions sur la distribution des données pour qu'une méthode soit plus précise qu'une autre.

2 Contexte et notations

L'indicatrice d'un évènement $\mathcal{E}$ sera ici notée $\mathbb{I}\{\mathcal{E}\}$, et la pseudo-inverse d'une fonction de répartition $F(u)$ sur $\mathbb{R}$ sera écrite $F^{-1}(t) = \inf\{v \in \mathbb{R} : \ F(v) \geq t\}$. Nous considérons le cadre de la classification à plusieurs classes. La variable aléatoire $Y \in \{1, \dots, K\}$ avec $K \geq 1$ dénote la classe d'une instance, et $X \in \mathcal{X} \subset \mathbb{R}^d$ représente les données caractéristiques d'une instance. La paire (X, Y) a pour distribution P et nous introduisons $p_k = \mathbb{P}\{Y = k\}$. L'apprentissage de similarité requiert un échantillon $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ composé de $n \geq 1$ copies indépendantes de (X, Y). Il consiste à construire une fonction de similarité $S : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ de manière à lorsque l'on considère deux paires indépendantes (X, Y) et (X', Y') tirées de P, plus la similarité $S(X, X')$ est grande, plus il est probable qu'elle soient associées à la même classe. L'ensemble des mesures de similarité est noté $\mathcal{S}$. La classe $\mathcal{S}^*$ des règles

de similarité optimales correspond naturellement à l'ensemble des transformées strictement croissantes T de la probabilité à posteriori η d'être dans la même classe pour une paire, $\eta(x, x') = \mathbb{P}\{Y = Y' \mid (X, X') = (x, x')\}$. Une fonction de similarité S définit un pré-ordre sur l'espace produit $\mathcal{X} \times \mathcal{X}$, et pour tout $x \in \mathcal{X}$, la fonction $s(x, .) : x' \mapsto s(x, x')$ définit un pré-ordre sur $\mathcal{X}$. Lorsque l'on cherche à optimiser un critère lié à la courbe ROC, il est naturel de voir le problème d'apprentissage de similarité comme un problème d'ordonnancement biparti sur l'espace produit $\mathcal{X} \times \mathcal{X}$ où les caractéristiques des données sont (X, X') et la classe binaire est définie par $2\mathbb{I}\{Y = Y'\} - 1$. La courbe ROC de S est le PP-plot $t \in \mathbb{R}^+ \mapsto (F_{S,-}(t), F_{S,+}(t))$, où, pour tout $t \geq 0$,

$$F_{S,-}(t) = \mathbb{P}\{S(X, X') > t \mid Z = -1\},$$
$$F_{S,+}(t) = \mathbb{P}\{S(X, X') > t \mid Z = +1\},$$

où les possibles sauts sont connectés par des segments. Par conséquent, elle peut être vue comme le graphe d'une fonction continue $\alpha \in (0, 1) \mapsto \mathrm{ROC}_S(\alpha)$, où $\mathrm{ROC}_S(\alpha) = F_{S,+} \circ F_{S,-}^{-1}(\alpha)$ en tout point $\alpha \in (0, 1)$ tel que $F_{S,-} \circ F_{S,-}^{-1}(\alpha) = \alpha$. La courbe ROC reflète la capacité de S à séparer les paires de la même classe et celle de différentes classes. Elle correspond au graphe de la puissance en fonction de l'erreur de type 1 du test statistique $\mathbb{I}\{S(X, X) > t\}$ quand l'hypothèse nulle est $\mathcal{H}_0 : Y \neq Y'$. Une mesure de similarité S_1 est dite plus précise qu'une autre S_2 si pour tout $\alpha \in (0, 1)$, $\mathrm{ROC}_{S_2}(\alpha) \leq \mathrm{ROC}_{S_1}(\alpha)$. Le lemme fondamental de Neyman-Pearson montre que les éléments de $\mathcal{S}^*$ sont optimaux pour cet ordre partiel sur $\mathcal{S}$. Définissons la fonction $S_\alpha^* : x, x' \mapsto \mathbb{I}\{x, x' \in R_\alpha^*\}$ où $R_\alpha^* = \{x, x' \in \mathcal{X}^2 \mid \eta(x, x') > Q_\alpha^*\}$ et Q_α^* est le $(1 - \alpha)$-quantile de $\eta(X, X') \mid Y \neq Y'$. Une autre conséquence du lemme fondamental de Neyman-Pearson est que la fonction S_α^* est optimale pour le problème d'optimisation de la courbe ROC en un point. Si l'on restreint les valeurs de nos fonctions de similarité aux fonctions bornées par 1, S_α^* est aussi la solution optimale au problème suivant:

$$\max_{S:\mathcal{X}^2 \to [0,1], \text{borélienne}} R^+(S) \quad \text{tel que} \quad R^-(S) \leq \alpha, \tag{1}$$

où $R^+(S) = \mathbb{E}[S(X, X') \mid Z = +1]$ est appelé *risque positif* et $R^-(S) = \mathbb{E}[S(X, X') \mid Z = -1]$ est appelé *risque négatif*. Le problème introduit par Eq. (1), restreint à une famille de fonctions de similarité $\mathcal{S}_0$ plus petite, décrit un problème plus général que l'optimisation de la courbe en un point. En effet, optimiser la courbe ROC pour le niveau α revient à maximiser sur $S \in \mathcal{S}_1, t \in \mathbb{R}$ la quantité $F_{S,+}(t)$ sous contrainte que $F_{S,-}(t) \leq \alpha$, ce qui est équivalent à considérer Eq. (1) restreint à $\mathcal{S}_0$, quand $\mathcal{S}_0$ est constitué des indicatrices des ensembles de sur-niveau stricts des éléments de $\mathcal{S}_1$, c'est-à-dire $\mathcal{S}_0 = \{x, x' \mapsto \mathbb{I}\{S(x, x') > t\} \mid S \in \mathcal{S}_1, t \in \mathbb{R}\}$.

3 Garanties statistiques de généralisation

L'optimisation en un point de la courbe ROC a été étudiée d'un point de vue statistique par [Scott and Nowak, 2005] et [Clémençon and Vayatis, 2010] dans le cas de la classification binaire. La différence essentielle du cas actuel repose sur le fait que les quantités dans Eq. (1) sont basées sur des paires, ce qui rend complexe sa version empirique. En effet, les estimateurs naturels du risque positif $R^+(S)$ et négatif $R^-(S)$ sur le sample $\mathcal{D}_n$ s'écrivent:

$$\hat{R}_n^+(S) = \frac{1}{n_+} \sum_{1 \leq i < j \leq n} S(X_i, X_j) \cdot \mathbb{I}\{Y_i = Y_j\}, \tag{2}$$

$$\hat{R}_n^-(S) = \frac{1}{n_-} \sum_{1 \leq i < j \leq n} S(X_i, X_j) \cdot \mathbb{I}\{Y_i \neq Y_j\}, \tag{3}$$

où $n+ = \sum_{1\leq i<j\leq n} \mathbb{I}\{Y_i = Y_j\} = n(n-1)/2 - n_-$. Il est important de noter que ces quantités ne sont pas des moyennes i.i.d., car plusieurs paires mettent en jeu chaque échantillon i.i.d. Cela rend caduque l'analyse conduite par [Clémençon and Vayatis, 2010, voir section 5] dans le cas de la classification binaire.

On peut cependant observer que $U_n^+(S) = 2n_+/(n(n-1))\hat{R}_n^+(S)$ et $U_n^-(S) = 2n_-/(n(n-1))\hat{R}_n^-(S)$ sont des U-statistiques de degré deux avec pour noyaux symétriques respectifs $K_+((x,y),(x',y')) = S(x,x')\cdot\mathbb{I}\{y=y'\}$ et $K_-((x,y),(x',y')) = S(x,x')\cdot\mathbb{I}\{y\neq y'\}$. Nous allons par conséquent être capables d'utiliser des astuces de représentation des U-statistiques pour en dériver des bornes de concentration pour les U-processus (collections de U-statistiques indexées par une classe de fonctions), sous des hypothèses de complexité.

Nous étudions donc les capacités de généralisation de solutions obtenues par la résolution de la version empirique du problème Eq. (1), où l'on restreint également le domaine des fonctions de similarités $\mathcal{S}$ à un sous-ensemble $\mathcal{S}_0$ de complexité contrôlée, par exemple par la VC-dimension. Dans notre cas, nous feront l'hypothèse que $\mathcal{S}_0$ est une classe VC-major de fonctions de VC-dimension finie V, ce qui signifie que les ensembles de sur-niveau strict des fonctions de $\mathcal{S}_0$, c'est-à-dire $\{\{x,x'\in\mathcal{X}^2 \mid S(x,x') > t\} \mid S\in\mathcal{S}_0, t\in\mathbb{R}\}$, forme une classe d'ensembles de VC-dimension finie V. La version empirique du problème Eq. (1) s'écrit:

$$\max_{S\in\mathcal{S}_0} \hat{R}_n^+(S) \quad \text{tel que} \quad \hat{R}_n^-(S) \leq \alpha + \Phi, \tag{4}$$

où Φ est un terme de tolérance positif qui sera de même ordre de grandeur que $\sup_{S\in\mathcal{S}_0}\left|\hat{R}_n^-(S) - R^-(S)\right|$. Qualitativement, la présence d'un paramètre $\Phi > 0$ se justifie par le fait qu'une solution S_1 de Eq. (1) restreint à $\mathcal{S}_0$ ne satisfait pas la contrainte de Eq. (4) avec grande probabilité si $\Phi = 0$. Dans ce cas, $R^-(S_1) = \alpha$ implique $\hat{R}_n^-(S_1) = \alpha + [\hat{R}_n(S_1) - R^-(S_1)]$, ce qui signifie que S_1 viole la contrainte d' Eq. (4) dès que $\hat{R}_n(S_1) > R^-(S_1)$.

Théorème 1. *Supposons que $\mathcal{S}_0$ est une classe VC-major de fonctions de VC-dimension $V < +\infty$, que $S(x,x') \leq 1$ pour tout $S\in\mathcal{S}_0$ et $x,x'\in\mathcal{X}\times\mathcal{X}$. Supposons aussi qu'il existe $\kappa\in(0,1)$ tel que $\kappa \leq \sum_{k=1}^K p_k^2 \leq 1-\kappa$. Pour tout $\delta\in(0,1)$ et $n>1$, définissons:*

$$\Phi_{n,\delta} = 2c\sqrt{\frac{V}{n}} + 2\sqrt{\frac{\log(2/\delta)}{n-1}},$$

où c est une constante universelle. Considérons une solution $\hat{S}_n$ du problème d'optimisation contrainte Eq. (4) avec $\Phi = \Phi_{n,\delta/2}$. Alors, pour tout $\delta\in(0,1)$, nous avons simultanément avec probabilité supérieure à $1-\delta$: pour tout $n\geq 1$,

$$\begin{aligned} R^+(\hat{S}_n) &\geq \sup_{S\in\mathcal{S}_0 : R^-(S)\leq\alpha} R^+(S) - C\Phi_{n,\delta/2}, \\ \text{et} \quad R^-(\hat{S}_n) &\leq \alpha + C\Phi_{n,\delta/2}, \end{aligned} \tag{5}$$

où C est une constante qui dépend seulement de κ.

Nous introduisons maintenant la condition de bruit nécéssaire à l'obtention des vitesses rapides.

Condition de bruit (NA). *Il existe $a\in(0,1)$ et $D>0$ tel que: $\forall t\geq 0$,*

$$\mathbb{P}\left\{|\eta(X,X') - Q_\alpha^*| \leq t\right\} \leq Dt^{\frac{a}{1-a}}.$$

Cette condition est similaire à celle introduite par [Mammen and Tsybakov, 1995b], sauf que le seuil 1/2 est remplacé par le quantile conditionnel Q_α^*. Sous cette hypothèse, nous pouvons démontrer le théorème suivant.

Théorème 2. *Supposons que les hypothèses du Théorème 1 sont vraies, que la condition NA est vraie et que la similarité optimale S^*_α appartient à $\mathcal{S}_0$. Fixons $\delta > 0$. Alors il existe une constante C', dépendant de $\delta, \kappa, Q^*_\alpha, a$ et D telle que, avec probabilité au moins $1-\delta$,*

$$\mathrm{ROC}_{S^*}(\alpha) - R^+(\hat{S}_n) \leq C' n^{-(2+a)/4},$$
$$\text{et} \quad R^-(\hat{S}_n) \leq \alpha + 2\Phi_{n,\delta/2}.$$

4 Mise à l'échelle à l'aide d'approximations par échantillonnage

Dans la partie précédente, nous avons analysé les vitesses d'apprentissage atteintes par un minimiseur du problème Eq. (4). Lorsque l'on aborde un problème de grande échelle, résoudre ce problème peut être computationnellement très couteux, à cause de la présence d'un très grand nombre de paires pour chaque somme $\hat{R}^+_n(S)$ et $\hat{R}^-_n(S)$, qui sont respectivement composées d'une moyenne sur $\sum_{k=1}^K n_k(n_k-1)/2$ et $\sum_{k<l} n_k n_l$ paires, avec n_k le nombre d'observations pour la classe $k \in \{1,\dots,K\}$. Nous sommes intéressés par le cas où l'on a un grand nombre de classes avec peu d'observations par classe, car il est proche du cas usuel en identification biométrique. Dans ce contexte, le nombre de paires négatives augmentant quadratiquement avec nombre d'observations, il est naturel de réduire drastiquement le nombre de négatifs. Une approche simple est de remplacer l'estimateur empirique du risque négatif $\hat{R}^-_n(S)$ par l'approximation suivante:

$$\bar{R}^-_B(S) := \frac{1}{B} \sum_{(i,j)\in\mathcal{P}_B} S(X_i, X_j),$$

où $\mathcal{P}_B$ est un ensemble de cardinal B construit en échantillonnant avec remise dans l'ensemble des paires négatives $\Lambda_{\mathcal{P}} = \{(i,j) \mid i,j \in \{1,\dots,n\}; Y_i \neq Y_j\}$. $\bar{R}^-_B(S)$ est appelée une version incomplète composée de B paires de la U-statistique de degré 2 $\hat{R}^-_n(S)$. Nous présentons une seconde stratégie d'échantillonnage, qui consiste à séléctionner des K-tuples contenant un échantillon aléatoire de chaque classe. Cela correspond à l'approximation suivante:

$$\widetilde{R}_B(S) := \frac{1}{B} \sum_{(i_1,\dots i_K)\in\mathcal{T}_B} h_S(X_{i_1},\dots,X_{i_K}),$$

et $h_S(X_1,\dots,X_K) = n_-^{-1} \sum_{k<l} n_k n_l S(X_k, X_l)$ et $\mathcal{T}_B$ est un ensemble de cardinal B construit en échantillonnant avec remplacement dans l'ensemble des K-tuples $\Lambda_{\mathcal{T}} = \{(i_1,\dots,i_K) \mid i_k \in \{1,\dots,n_k\}; k = 1,\dots,K\}$. $\bar{R}^-_B(S)$ et $\widetilde{R}_B(S)$ sont tous deux des estimateurs non biaisés de $\hat{R}^-_n(S)$, mais leurs variances sont différentes et une approximation peut être meilleure qu'une autre dans certains régimes.

Proposition 1. *Soit n_0 le nombre de paires échantillonnées dans chaque type d'échantillonnage, et dénotons $V_n = Var\left(\hat{R}^-_n(S)\right)$. Quand $n_0/n \to 0$, $n \to \infty$ et pour tout $k \in \{1,\dots,K\}$, $n_k/n \to p_k > 0$, nous avons:*

$$Var\left(\widetilde{R}^-_B(S)\right) - V_n \sim \frac{K(K-1)}{2n_0} Var\left(h_S(X^{(1)},\dots,X^{(K)})\right),$$
$$Var\left(\bar{R}^-_B(S)\right) - V_n \sim \frac{1}{n_0} Var\left(S(X,X') \mid Y \neq Y'\right),$$

où $X^{(k)}$ dénote $X|Y=k$ pour tout $k \in \{1,\dots,K\}$.

Ce résultat montre que dans certains régimes, un des schémas d'échantillonnage est meilleur que l'autre. Les résultats de [Clémençon et al., 2016] permettent de généraliser l'ordre des garanties données par le Théorème 1 au problème où $\hat{R}^-_n(S)$ est remplacé par $\bar{R}^-_B(S)$ ou $\widetilde{R}^-_B(S)$.

5 Conclusion

Nous avons introduit un cadre probabiliste rigoureux pour étudier l'apprentissage de similarité sous la perspective nouvelle de l'ordonnancement de paires d'observations et l'optimisation en un point de la courbe ROC. Nous avons dérivé des garanties statistiques sur la généralisation et étudié l'impact de l'utilisation d'approximations basées sur du sous-échantillonnage. Des expériences illustratives ont été conduites pour illustrer nos vitesses rapides de convergence, ainsi que la viabilité de l'apprentissage avec des estimateurs incomplets.

References

[Bellet et al., 2015] Bellet, A., Habrard, A., and Sebban, M. (2015). *Metric Learning*. Morgan & Claypool Publishers.

[Clémençon et al., 2016] Clémençon, S., Colin, I., and Bellet, A. (2016). Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics. *Journal of Machine Learning Research*, 17(76):1–36.

[Clémençon and Vayatis, 2010] Clémençon, S. and Vayatis, N. (2010). Overlaying classifiers: A practical approach to optimal scoring. *Constructive Approximation*, 32(3):619–648.

[Lee, 1990] Lee, A. J. (1990). *U-statistics: Theory and practice*. Marcel Dekker, Inc., New York.

[Mammen and Tsybakov, 1995a] Mammen, E. and Tsybakov, A. B. (1995a). Asympotical minimax recovery of the sets with smooth boundaries. *The Annals of Statistics*, 23(2):502–524.

[Mammen and Tsybakov, 1995b] Mammen, E. and Tsybakov, A. B. (1995b). Asymptotical minimax recovery of sets with smooth boundaries. *Annals of Statistics*, 23(2):502–524.

[Scott and Nowak, 2005] Scott, C. and Nowak, R. (2005). A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819.

31/05/2018

Une théorie probabiliste de l'apprentissage supervisé de similarité pour l'optimisation en un point de la courbe ROC

Journées de Statistique 2018

Robin Vogel, Stéphan Clémençon et Aurélien Bellet.

IDEMIA

TELECOM ParisTech

Une théorie probabiliste de l’apprentissage supervisé de similarité pour l’optimisation en un point de la courbe ROC

1. Contexte et motivation,
2. Formalisation,
3. Garanties de généralisation pour l’ERM,
4. Mise à l’échelle par échantillonnage.
5. Perspective

Contexte et motivation (1/2)

L'authentification biométrique valide l'identité d'un individu par ses attributs biologiques.

Une mesure enregistrée x (ex: photo de passeport) est comparée à une mesure capturée x' (ex: photo à l'aéroport) pour prendre une décision quant à leur correspondance.

La décision est faite par le seuillage d'une similarité S entre deux mesures.

Soit un seuil $t \in \mathbb{R}$, si $S(x, x') > t$, alors x et x' sont considérées correspondantes.

Deux types d'erreur différents peuvent être commises:

I - Accepter à tort une identité : faux positif (FP),

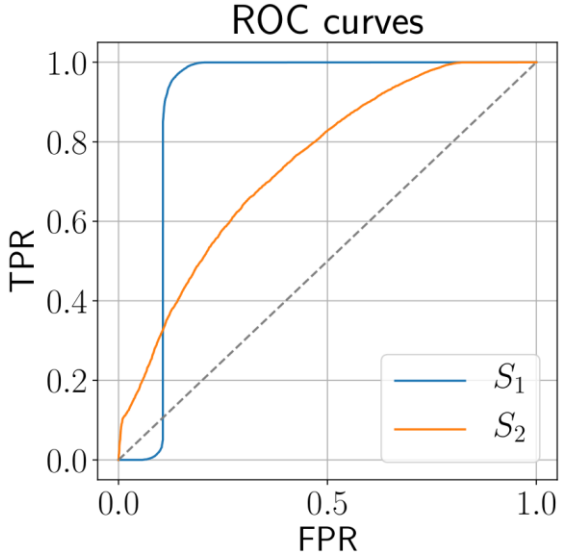
II - Rejeter à tort une identité : faux négatif (FN).

L'erreur de type I est généralement plus critique que l'autre.

La courbe ROC résume les erreurs pour chaque seuil.

Représente le taux de vrais positifs (1-FNR) en fonction du taux de faux positifs (FPR).

Permet de comparer deux similarités.



Contexte et motivation (2/2)

Le seuil t est alors choisi et fixé pour garantir un certain FPR, généralement très petit.

Des exemples raisonnables de FPR ciblés peuvent être $\mathrm{FPR} = 10^{-3}$ ou $\mathrm{FPR} = 10^{-6}$.

La fonction de similarité S est apprise à partir de données, en optimisant un critère.

Il est optimisé sur une base de données $(X_i, Y_i)_{i=1}^{n}$,
avec X_i une mesure (ex: visage) et Y_i son identité associée.

Les critères n'optimisent souvent pas le TPR à FPR fixé.

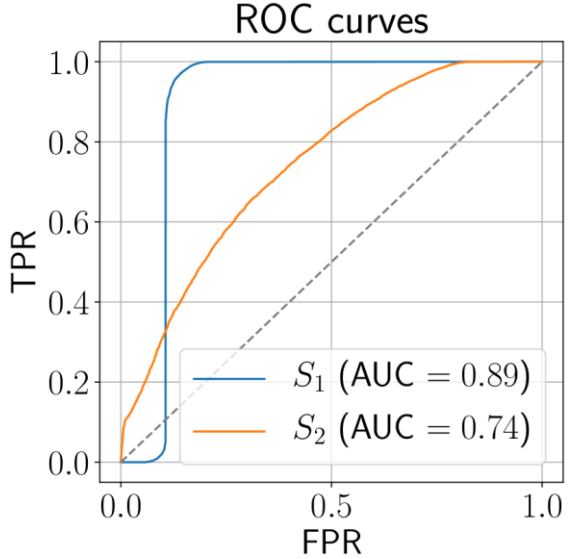
Des exemples de critères sont:

- l'aire sous la courbe ROC (AUC),
- l'erreur de classification.

Nous étudions donc l'optimisation du TPR à FPR fixé.

Ce problème s'écrit, sur une classe $\mathcal{S}_0$ de fonctions de similarité:

$$(\mathrm{P}_\alpha) \qquad \max_{S \in \mathcal{S}_0, t \in \mathbb{R}} \mathrm{TPR}_S(t), \ \text{s.c}\ \mathrm{FPR}_S(t) \leq \alpha.$$



Formalisation

Nos données $(X_i, Y_i)_{i=1}^n$ sont similaires à celles de la classification.

$(X_i, Y_i)_{i=1}^n$ **sont des copies i.i.d. de** (X, Y)**, avec** $X \in \mathbb{R}^d, Y \in \{1, \dots, K\}$.

Nous étudions un problème plus général que l'optimisation en un point de la courbe ROC:

$$(\mathrm{T}_\alpha) \qquad \max_{S \in \mathcal{S}} R^+(S), \text{ s.c. } R^-(S) \leq \alpha, \qquad \rightarrow \text{solution } S^*.$$

où $R^+(S) = \mathbb{E}[\, S(X, X') \mid Y = Y' \,]$ et $R^-(S) = \mathbb{E}[\, S(X, X') \mid Y \neq Y']$.

Choisir $\mathcal{S} = \{\mathbb{I}\{S(x, x') > t\} \mid S \in \mathcal{S}_0, t \in \mathbb{R}\}$ pour (T_α) nous ramène au problème (P_α).

Les estimateurs naturels des quantités $R^+(S)$ et $R^-(S)$ sont :

$$R_n^+(S) = \frac{1}{n_+} \sum_{i<j} S(X_i, X_j) \cdot \mathbb{I}\{Y_i = Y_j\} \text{ et } R_n^-(S) = \frac{1}{n_-} \sum_{i<j} S(X_i, X_j) \cdot \mathbb{I}\{Y_i \neq Y_j\},$$

où n_+ et n_- sont les nombres de paires positives et négatives.

La version empirique (E_α) de (T_α) s'écrit:

$$(\mathrm{E}_\alpha) \qquad \max_{S \in \mathcal{S}} R_n^+(S), \text{ s.c. } R_n^-(S) \leq \alpha + \Phi, \qquad \rightarrow \text{solution } \hat{S}_n.$$

où Φ tolère les variations de R_n^- autour de se moyenne.

Garanties de généralisation pour l'ERM

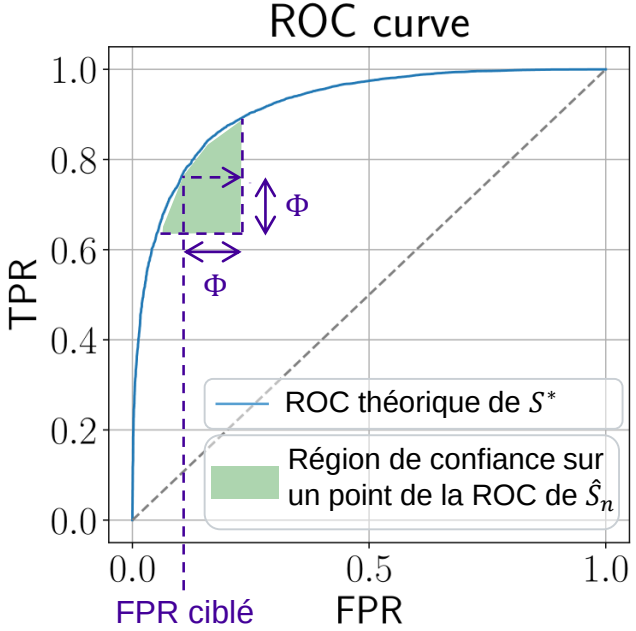
Nos résultats garantissent localement la proximité entre la ROC de $\widehat{S}_n$ et celle de S^*.

Théorème. *Supposons que $\mathcal{S}$ est une classe VC-major de VC dimension $V < +\infty$, que $0 \leq S \leq 1$ pour tout $S \in \mathcal{S}$ et que $p = \mathbb{P}\{Y = Y'\}$ ne s'approche pas de zéro. Soit $\delta \in (0,1)$, $n \geq 1 + 4p^{-2}\log(3/\delta)$,*

$$\Phi_{n,\delta} = 2cp^{-1}\sqrt{\frac{V}{n}} + 2p^{-1}(1+p^{-1})\sqrt{\frac{\log(3/\delta)}{n-1}},$$

alors avec probabilité $\geq 1-\delta$,

$$R^+(\hat{S}_n) \geq R^+(S^*) - \Phi_{n,\delta} \text{ et } R^-(\hat{S}_n) \leq \alpha + \Phi_{n,\delta}.$$



L'ordre en n de la borne sur R^+ peut être amélioré.

- **Une analyse de la variance de l'excès de risque sur R^+ amène à des ordres entre $n^{-1/2}$ et $n^{-3/4}$.**
- **Conséquences de (Mammen & Tsybakov, 1995).**

Garanties de généralisation (Eléments de preuve)

R_n^+ et R_n^- sont des ratios de U-statistiques:

Definition. *Soit $V_1, \ldots, V_n$ des v.a. i.i.d. dans un espace mesurable $\mathcal{X}$, K une fonction $\mathcal{X} \times \mathcal{X} \to [0,1]$, alors $U_n = (1/n(n-1)) \sum_{i \neq j} K(V_i, V_j)$ est appelée U-statistique de degré 2 de noyau K. C'est l'estimateur non biaisé de $\mathbb{E}[K(V_1, V_2)]$ de plus petite variance.*

Des résultats classiques de concentration existent pour les U-statistiques.

Soit U_n une U-statistique de noyau borné par 1, avec probabilité > $1 - \delta$,

$$\left|U_n - \mathbb{E}[U_n]\right| \leq \sqrt{\frac{\log(2/\delta)}{n-1}}. \qquad \textbf{(Hoeffding, 1963)}$$

La minimisation de critères liés au meilleur test de niveau α a été étudiée.

Ce travail nous fournit la propriété suivante:

$$\mathbb{P}\left\{\left(R^+(S^*) - R^+(\hat{S}_n) \geq \Phi\right) \cap \left(R^-(\hat{S}_n) \leq \alpha + \Phi\right)\right\}, \qquad \textbf{(Scott and Nowak, 2005)}$$

$$\geq 1 - \mathbb{P}\left\{\sup_{S \in \mathcal{S}} |R_n^+(S) - R^+(S)| > \Phi\right\} - \mathbb{P}\left\{\sup_{S \in \mathcal{S}} |R_n^-(S) - R^-(S)| > \Phi\right\}.$$

Mise à l'échelle par échantillonnage

Généralement, nous avons beaucoup de classes avec peu d'observations par classe.

K grand, $n_k = \sum_{i=1}^{n} \mathbb{I}\{ Y_i = k \}$ petit, pour tout $k \in \{1, \dots, K\}$.

Le nombre de paires moyenné pour calculer R_n^- est alors quadratique en n.

Exemple: base de données LFW: 2×10^5 paires positives pour 9×10^7 paires négatives.

Pour résoudre $(\mathbf{E}_\alpha)$, on peut estimer R^- avec moins de paires.

Nous introduisons un estimateur $\widetilde{R}_B^-$, qui moyenne B paires sélectionnées aléatoirement.

Si B est de l'ordre de n, l'ordre en n des garanties de généralisation n'est pas affecté.

Théorème. *Supposons que $\mathcal{S}$ est une classe VC-major de VC dimension $V < +\infty$, que $0 \leq S \leq 1$ pour tout $S \in \mathcal{S}$. Pour tout $\delta > 0$, avec probabilité $\geq 1 - \delta$,*

$$\sup_{S \in \mathcal{S}} \left| \tilde{R}_B^-(S) - R_n^-(S) \right| \leq \sqrt{2 \frac{V \log(1 + n^2) + \log(2/\delta)}{B}}.$$

(Clémençon & Colin & Bellet , 2016)

Perspective

Nous avons trouvé des garanties pour l'optimisation en un point de la courbe ROC.

Ainsi que présenté une stratégie pour sa mise à l'échelle.

Nous aimerions nous approcher de la solution de (P_α): $\max_{S \in \mathcal{S}_0, t \in \mathbb{R}} \mathrm{TPR}_S(t), \text{ s.c } \mathrm{FPR}_S(t) \leq \alpha.$

Optimiser l'AUC ou faire de la classification n'optimise pas notre critère.

Approche: Résoudre (E_α) pour une famille $\mathcal{S}$.

➔ **Dans quelle mesure S^* optimise bien la ROC en un point ?**

Si $\mathcal{S} = \{\mathbb{I}\{S(x, x') > t\} \mid S \in \mathcal{S}_0, t \in \mathbb{R}\}$, parfaitement.

Si $\mathcal{S}$ est constitué de fonctions linéaires, probablement assez peu.

➔ **Comment s'approcher de $\widehat{S}_n$?**

Si $\mathcal{S}$ est constitué des fonctions linéaires, c'est très facile.

Merci !

Références

- Hoeffding, Wassily. Probability Inequalities for Sums of Bounded Random Variables. Journal of the American Statistical Association, 58(301):13-30, 1963.

- Scott, Clayton and Nowak, Robert. A Neyman-Pearson approach to statistical learning. IEEE Transactions on Information Theory, 51(11):3806-3819, 2005.

- Clémençon, Stephan, Colin, Igor and Bellet, Aurélien. Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics. Journal of Machine Learning Research, 17(76):1-36, 2016.

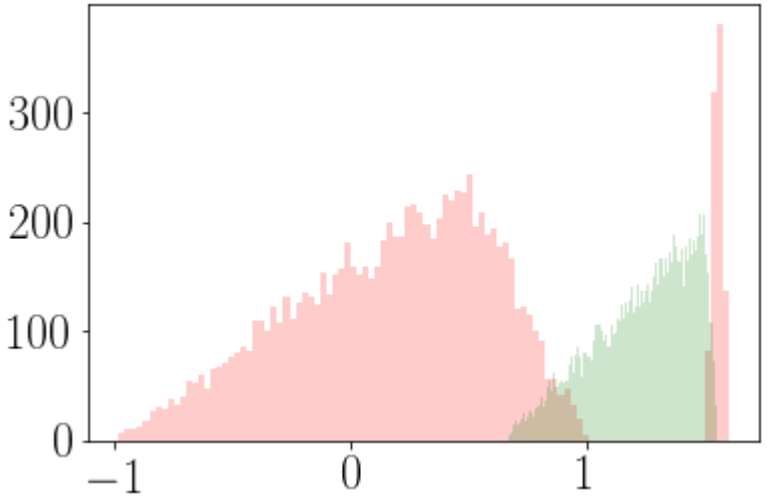
Our paper:

- Vogel, Robin, Clémençon, Stéphan and Bellet, Aurélien. A Probabilistic Theory of Supervised Similarity Learning for Pointwise ROC Curve Optimization. To appear in ICML 2018.

Appendice

Distributions de scores des courbes ROC présentées en introduction.

S_1



S_2