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

Journée de la Chaire Machine Learning for Big Data 2018

Robin Vogel, en thèse CIFRE avec Stéphan Clémençon, Anne Sabourin et Aurélien Bellet.

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:

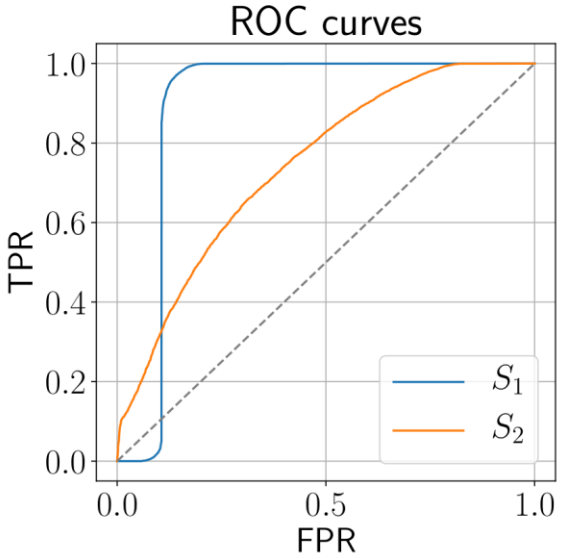
I - Accepter à tort une identité : faux positif (FP),

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 $\mathrm{FPR} = 10^{-6}$.

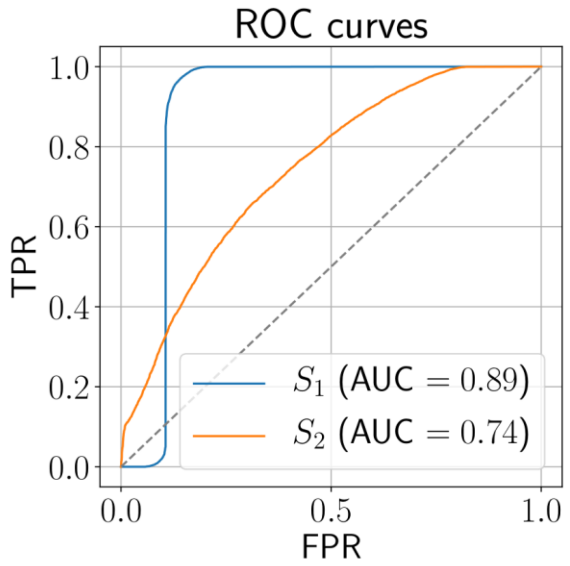
La fonction de similarité S est apprise à partir de données, en optimisant un critère.

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:

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

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:

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

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

Garanties de généralisation pour l'ERM

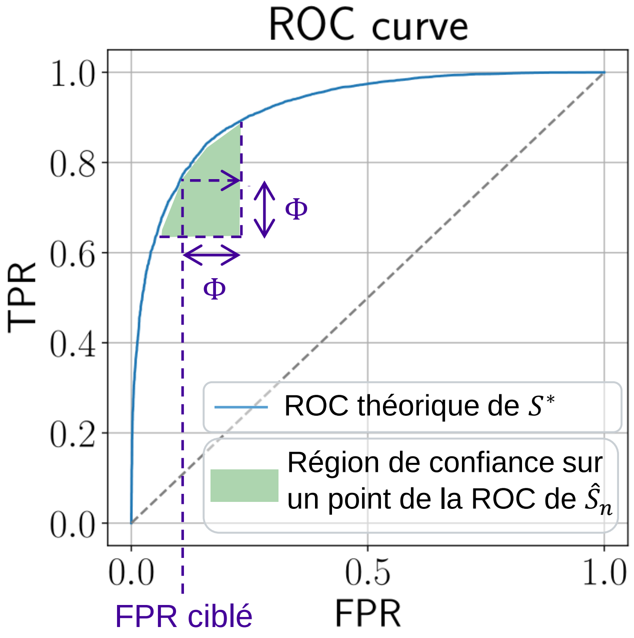
Nos résultats garantissent localement la proximité entre la ROC de $\hat{S}_n$ et celle de S^*.

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 $\mathbb{P}\{Y = Y'\}$ ne s'approche pas de zéro. Soit $\delta \in (0,1)$, $n \geq 1$,*

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

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 > $1 - \delta$,

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

(Hoeffding, 1963)

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

Ce travail nous fournit la propriété suivante:

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

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

(Scott and Nowak, 2005)

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: 9×10^7 paires négatives pour 2×10^5 paires positives.

Pour résoudre (E_α), 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.

En pratique on apprend une similarité à l'aide de coûts ne correspondant pas à ce problème.

Optimiser l'AUC ne correspond pas à notre critère.

Certaines approches résolvent un problème plus général (ex: classification), pour en déduire une similarité.

➔ contre le principe de Vapnik.

Dans quelle mesure $\widehat{S}_n$ optimise bien la ROC en un point, pour quelle classe de fonctions $\mathcal{S}$?

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

Dans un premier temps, on peut adresser le cas de l'ordonnancement bipartite (B_α):

$$(\mathrm{B}_\alpha) \quad \max_{S \in \mathcal{S}} R_n^+(S), \text{ s.c. } R_n^-(S) \leq \alpha,$$

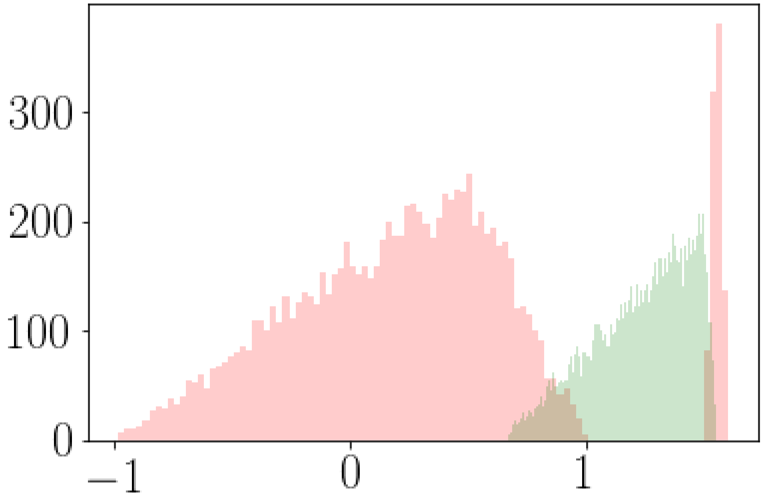
$$\text{où } R_n^+(S) = \frac{1}{n_+} \sum_{i=1}^{n} S(X_i) \cdot \mathbb{I}\{Y_i = +1\} \text{ et } R_n^-(S) = \frac{1}{n_-} \sum_{i=1}^{n} S(X_i) \cdot \mathbb{I}\{Y_i = -1\}.$$

Merci !

Appendice

Distributions de scores des courbes ROC présentées en introduction.

S_1



S_2