Agrégation d'hold outs

Guillaume Maillard[1] & Sylvain Arlot[2] & Matthieu Lerasle[3]

[1] *Département de Mathématiques, Bât. 307, Faculté des Sciences d'Orsay, Université Paris-Sud, F-91405 Orsay Cedex, France, mail: guillaume.maillard@u-psud.fr*
[2] *idem, mail: sylvain.arlot@u-psud.fr*
[3] *idem, mail: matthieu.lerasle@u-psud.fr*

Résumé. La validation croisée est souvent utilisée pour sélectionner une règle d'apprentissage parmis une famille, souvent paramètrée (sélection d'hyperparamètres). L'article que nous présentons étudie une méthode voisine, appelée agrégation d'hold-out (Agghoo), qui mélange validation croisée et agrégation; des liens peuvent aussi être établis avec le bagging. Nous obtenons les premières garanties théoriques sur Agghoo, ce qui assure que l'on peut l'utiliser sans risque: au pire, les performances d'Agghoo sont celles du hold-out, à constante près. Pour le hold-out, des inégalités oracle étaient connues dans le cas des pertes bornées, comme en classification binaire. Cette approche semble pouvoir être étendue, sous de bonnes hypothèses, à la plupart des problèmes de minimisation de risque. Sous des hypothèses faibles, nous obtenons notamment une inégalité d'oracle concernant le choix du paramètre de pénalisation des SVM à perte Lipschitz. Dans toutes ces situations, Agghoo vérifie donc une inégalité d'oracle. Cependant, des simulations suggèrent que le comportement réel est souvent bien meilleur que ce que la théorie peut démontrer pour l'instant. En particlier, l'agrégation conduit à une amélioration significative que les bornes théoriques actuelles venant du hold-out sont incapables d'expliquer. En conséquence, l'agrégation d'hold-out semble donc bien être compétitive en pratique, lorsqu'on la compare à la validation croisée.

Mots-clés. Sélection d'hyperparamètre, Validation Croisée, Agrégation, SVM

Abstract. Cross-validation is widely used for selecting among a family of learning rules. This paper studies a related method, called aggregated hold-out (Agghoo), which mixes cross-validation with aggregation; Agghoo can also be related to bagging. We provide the first theoretical guarantees on Agghoo, ensuring that one can use it safely: at worse, Agghoo performs like the hold-out, up to a constant factor. For the hold-out, oracle inequalities were known in the case of bounded losses, as in binary classification. The approach can be extended, with appropriate hypotheses, to most classical risk-minimization problems. Under weak hypotheses, we obtain an oracle inequality for the penalty parameter of SVM with a Lipschitz loss. In all these settings, Agghoo verifies an oracle inequality. However, simulation studies suggest that real performance is often much better than what theory can currently prove. In particular, there is a large gain from aggregation that current bounds derived from the hold-out are incapable of capturing. As a result, Aggregated hold-out appears to be competitive with standard cross-validation in practice.

Keywords. Hyperparameter selection, Cross-Validation, Agregation, SVM ...

1 Définition de l'agrégation d'hold out

Commençons par introduire quelques notations. Soit un problème de minimisation de risque en prédiction caractérisé par des espaces $\mathcal{X}, \mathcal{Y}$ et une fonction de perte $\gamma(u, y)$. Un prédicteur t est une fonction mesurable de $\mathcal{X}$ dans $\mathcal{Y}$. Son risque est défini par $\mathcal{L}(t) = \mathbb{E}[\gamma(t(X), Y)]$ où X, Y suit une loi inconnue P. On mesurera la performance d'un prédicteur à l'aide de l'excès de risque $\ell(s,t) = \mathcal{L}(t) - \inf_{t' : prédicteur} \mathcal{L}(t')$. $D_n = (X_i, Y_i)_{1 \leq i \leq n} \sim P^{\otimes n}$ dénotera un $n-$ échantillon, supposé i.i.d de loi P. Pour un sous ensemble $T \subset \{1, \ldots, n\}$, le sous échantillon correspondant sera $D_n^T = (X_j, Y_j)_{j \in T}$. En apprentissage statistique, les prédicteurs doivent être construits à partir d'un échantillon selon une *règle d'apprentissage*, notée $\mathcal{A}$ qui sera donc une fonction $\mathcal{A} : \cup_{k \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^k \to \{\text{prédicteurs}\}$ où $\{prédicteurs\}$ dénote l'ensemble des fonctions mesurables de $\mathcal{X}$ dans $\mathcal{Y}$.

Les algorithmes d'apprentissage correspondent souvent à des familles $(\mathcal{A}_m)_{m \in \mathcal{M}}$ de règles d'apprentissage paramétrées par un ou plusieurs *hyperparamètres* m laissés au choix de l'utilisateur. Par exemple, $m = k$ en classification $k-$NN ou $m = \lambda$, paramètre de régularisation du Lasso, du Ridge ou des SVMs..

La procédure hold-out, bien connue en machine learning, consiste à diviser les données en un échantillon d'entraînement et un échantillon test. Les estimateurs sont entraînés sur le premier sous-échantillon et leur performance est évaluée sur l'échantillon test grâce à la fonction de perte γ. La validation croisée consiste à moyenner les résultats obtenus sur plusieurs découpages en échantillons d'entraînement et échantillon test.

Définition 1 *Soit $D_n = (X_i, Y_i)_{1 \leq i \leq n}$ un $n-$ échantillon. Soit $T \subset \{1, \ldots, n\}$. Pour toute règle d'apprentissage $\mathcal{A}$,*

$$HO_T(\mathcal{A}) = \frac{1}{|T^c|} \sum_{j \notin T} \gamma\big(\mathcal{A}(D_n^T)(X_j), Y_j\big).$$

$1 \leq p \leq n$, $\mathcal{T} \subset \{T \subset \{1, ..., n\} : |T| = p\}$

$$CV_{\mathcal{T}}(\mathcal{A}) = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} HO_T(\mathcal{A})$$

Si l'on a une famille paramétrée de règles d'apprentissage, ceci fournit une méthode de sélection d'hyperparamètres - l'idée étant d'optimiser en m l'estimateur hold-out / VC du risque.

Définition 2 *Soit $\mathcal{M}$ un ensemble d'hyperparamètres.*

- *Un* prédicteur hold out *est donné par* $\widehat{f}_T^{\text{ho}} = \mathcal{A}_{\hat{m}_T^{ho}}(D_n^T)$ *où* $\hat{m}_T^{ho} = \text{argmin}_{m\in\mathcal{M}} HO_T(\mathcal{A}_m)$.
- *Un* prédicteur VC *est donné par:* $\widehat{f}_{\mathcal{T}}^{CV} = \mathcal{A}_{\hat{m}_{\mathcal{T}}^{CV}}(D_n)$ *où* $\hat{m}_{\mathcal{T}}^{CV} = \text{argmin}_{m\in\mathcal{M}} CV_{\mathcal{T}}(\mathcal{A}_m)$

En pratique, la validation croisée est préférée au hold-out car plus stable. La procédure Agghoo est un autre moyen de stabiliser le hold-out, en agrégeant les prédicteurs obtenus par la procédure hold-out appliquée plusieurs fois avec des échantillons d'entraînement différents $T \in \mathcal{T}$. Formellement, le prédicteur Agghoo $\widehat{f}_{\tau,V}^{\text{ag}}$ est obtenu en les moyennant.

Définition 3 *Supposons que* $\mathcal{Y}$ *est convexe (régression). Soient* $T_1, ..., T_V$ *i.i.d de loi* $\mathcal{U}(\{T \subset \{1, .., n\} : |T| = \lfloor \tau n \rfloor\})$.

$$\widehat{f}_{\tau,V}^{\text{ag}} = \frac{1}{V}\sum_{i=1}^{V} \widehat{f}_{T_i}^{\text{ho}}$$

Les paramètres sont V *et* τ.

Comparé à la validation croisée, l'ordre entre minimisation de l'estimateur du risque et agrégation est inversé: au lieu de moyenner les estimateurs de risque avant de sélectionner l'hyperparamètre, l'étape de sélection est faite d'abord, produisant des prédicteurs hold-out $\widehat{f}_{T_j}^{\text{ho}}$ qui sont ensuite moyennés.

A notre connaissance, Agghoo n'a encore fait l'objet d'aucune publication. Les méthodes les plus proches que nous ayons trouvé sont la"K-fold averaging cross-validation" (ACV) proposée par Jun et Hu (2016) pour la régression linéaire, et "efficient K-fold cross-validation" (EKCV) proposée par Jun(2016) pour la régression en grande dimension. La différence principale avec Agghoo est que ACV et EKCV calculent la moyenne des paramètres sélectionnés tandis qu'Agghoo moyennise les règles. Ceci donne des résultats complètement différents quand les règles d'apprentissage ne sont pas linéaires en leurs paramètres. Moyenniser les prédicteurs plutôt que les hyperparamètres semble plus naturel: en effet, pour un même problème, plusieurs paramétrisations différentes peuvent s'appliquer - par exemple, en régression, pénalisation ridge ou restriction à des boules de rayon R - de plus la convexité du risque par rapport à l'hyperparamètre n'est typiquement pas garantie.

2 Résultats théoriques

Si le risque est convexe, on peut montrer grâce à une inégalité de convexité que le risque d'Agghoo est toujours inférieur à celui du hold-out. Une inégalité oracle pour Agghoo peut donc se déduire d'une inégalité d'oracle pour le hold-out. Concernant le hold-out, Arlot et Celisse (2010) donne un état de l'art des résultats disponibles à cette date. La littérature existante s'est concentrée sur des cas où le contraste est borné, en posant par

exemple $\mathcal{Y} = [-M; M]$ au lieu de $\mathcal{Y} = \mathbb{R}$ en régression. Nous avons réussi à relâcher cette hypothèse dans le cas des méthodes à noyaux.

Etant donnée une fonction de contraste c et un RKHS $\mathcal{H}$ associé au noyau K, considérons les méthodes à noyaux:

$$\mathcal{A}_\lambda(D_{n_t}) = \operatorname*{argmin}_{t\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n_t} c(t(X_i), Y_i) + \lambda \|t\|^2_{\mathcal{H}}\,. \tag{1}$$

Par exemple, $c(u, y) = (1 - uy)_+$ donne les SVM habituelles, $c(u, y) = (|u - y| - \varepsilon)_+$ correspond à la $\varepsilon-$régression. Supposons que le noyau est borné: $\sup K(x, x) < \infty$, que $\gamma(t, (x, y)) = c(t(x), y) = |t(x) - y|$, cas correspondant à la régression médiane ou $\varepsilon-$régression avec $\varepsilon = 0$. Alors on a:

Théorème 4 *Supposons que pour presque tout $x \in \mathcal{X}$ et pour tout $u \in \mathbb{R}$,*

$$\big|\mathbb{P}(Y \le u|X = x) - \mathbb{P}(Y \le s(x)|X = x)\big| \ge \frac{|u - s(x)|}{\delta}\mathbb{I}_{|u-s(x)|\le\frac{\delta}{4}}.$$

Soient $\Lambda \subset \mathbb{R}_+$ et $\lambda_m = \min\Lambda$. Soit $\tau \in [0; 1]$, $n_t = \lfloor \tau n \rfloor$ et $n_v = n - n_t$. Alors pour tout $\theta \in (0; 1]$, l'application de la procédure Agghoo à la famille $(\mathcal{A}_\lambda)_{\lambda\in\Lambda}$ donne:

$$\begin{aligned}(1-\theta)\mathbb{E}[\ell(s, \widehat{f}^{\mathrm{ag}}_{\tau,V})] \le (1+\theta)\mathbb{E}\big[\min_{\lambda\in\Lambda}\ell(s, \mathcal{A}_\lambda(D_{n_t}))\big] \\ + \frac{a\delta}{n_v} + \frac{C_3}{\lambda_m}\left[\frac{\log^2(n_v|\Lambda|)}{\theta^3 n_v^2} + \frac{\log^{\frac{3}{2}}(n_v|\Lambda|)}{\theta n_v\sqrt{n_t}}\right],\end{aligned} \tag{2}$$

où C_3 ne dépend que de K et a est une constante absolue.

Si le terme de reste est négligeable par rapport à l'oracle $\mathbb{E}\big[\min_{\lambda\in\Lambda}\ell(s, \mathcal{A}_\lambda(D_{n_t}))\big]$ quand $n \to +\infty$, alors Agghoo fait aussi bien, à constante près, que le meilleur choix de λ. Cela dépend de λ_m, donc de la fenêtre dans laquelle le hold-out s'autorise à choisir λ. Il faut donc que cette fenêtre contienne le λ optimal. Sous des hypothèses semblables à celles du théorème 4, Steinwart et Christmann (2008) montrent la consistance des estimateurs à noyaux $\mathcal{A}_{\lambda_n}(D_n)$ sous l'hypothèse $\lambda_n^2 n \to +\infty$. Prenons θ, τ fixés, indépendants de n. En prenant donc $\lambda_m = \frac{b}{\sqrt{n}}$ pour une constante b, on a donc un terme de reste d'ordre $\frac{\log^{\frac{3}{2}}(n|\Lambda|)}{n}$, ce qui est bien négligeable par rapport aux vitesses de convergence non-paramétriques.

Dans le cas particulier du noyau gaussien et sous des hypothèses sur X, en régression bornée, Eberts et Steinwart (2013) montrent la convergence avec vitesse approchant $\frac{2\alpha}{2\alpha+d}$ sur la classe de Sobolev $W_2^\alpha(\mathbb{R}^d)$, pour $\lambda_n \sim \frac{1}{n^\beta}$ quand $\beta < 1$ et β arbitrairement proche de 1. En prenant donc $\lambda_m = \frac{1}{n}$, le terme de reste de l'équation (2) est alors d'ordre $\frac{\log^{\frac{3}{2}}(n|\Lambda|)}{\sqrt{n}}$, ce qui est négligeable si $\frac{2\alpha}{2\alpha+d} < \frac{1}{2}$.

3 Résultats empiriques

La théorie ne permet pas à ce stade d'évaluer l'impact de l'agrégation: on observe que l'inégalité d'oracle (équation ..) ne fait pas intervenir le paramètre V.Pour pallier à ce manque, nous avons réalisé des simulations.

Considérons les méthodes à noyau décrites dans la section 3 avec un noyau gaussien $K(x,y) = e^{-2(x-y)^2}$ et $c(u,y) = (|u-y| - 0.25)_+$ ($\varepsilon-$régression). 500 données i.i.d sont générées suivant la loi de (X,Y), où $X \sim \mathcal{N}(0,\pi)$ et $Y = e^{\cos(X)} + \xi$ avec $\xi \sim \mathcal{N}\left(0, \frac{1}{2}\right)$ indépendant de X. Pour la sélection de λ, on utilise la grille $\Lambda = \left\{ \frac{2^{j-1}}{500\tau n} | 1 \leq j \leq 17 \right\}$. Le risque est évalué en utilisant un échantillon supplémentaire de 500 données, et 1000 répétitions sont réalisées afin de calculer la moyenne.

Les résultats sont représentés sur la Figure 1. Le risque moyen de Agghoo y est comparé au risque moyen de la validation croisée monte carlo qui a les mêmes paramètres τ et V. τ est représenté en abscisse, et l'excès de risque moyen en ordonnée.

On observe que le risque d'Agghoo diminue de façon marquée avec V et que pour $\tau \in [0,4; 0,6]$, il est même inférieur à celui de l'oracle (meilleure valeur de λ) - chose que ne peut évidemment réaliser la validation croisée.

Cela illustre la bonne performance de cette méthode en pratique.

Bibliographie

Jun,Y. et Hu,J. (2015). A $K-$fold averaging cross-validation procedure, *Journal of Non-parametric Statistics*,27,pp. 167-179

Jun,Y. (2016). Efficient Tuning Parameter Selection by Cross-Validated Score in High Dimensions, *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering*, 10, pp. 19-25

Arlot,S. et Celisse,A. (2010). A survey of cross-validation procedures for model selection, *Statistics Surveys*, 4, pp. 40-79

Steinwart,I. et Christmann, A. (2008), *Support Vector Machines*, Springer Science and Business Media

Eberts,M. et Steinwart,I. (2013). Optimal regression rates for SVMs using Gaussian kernels, *Electronic Journal of Statistics*, 7, pp. 1-42

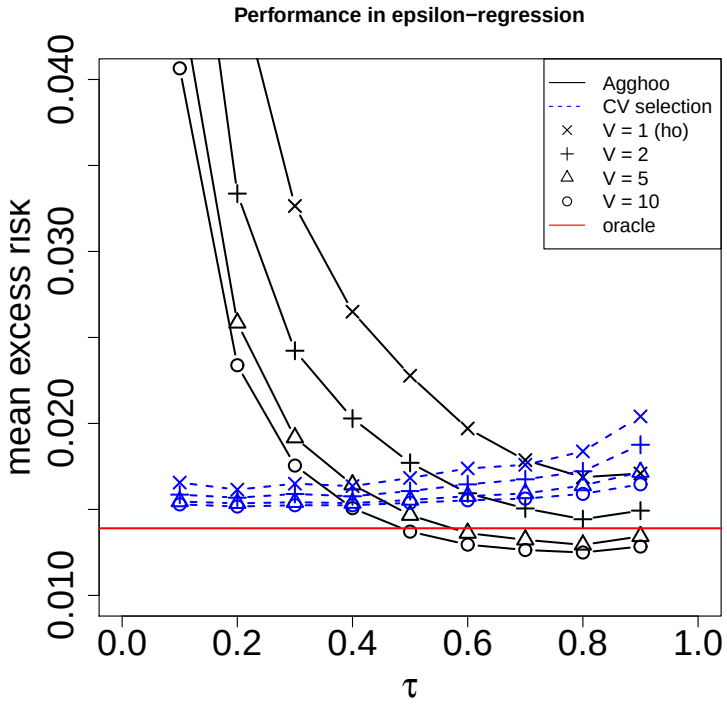

Figure 1: Risque en $\varepsilon-$régression