Agrégation d'hold-out

Guillaume Maillard, M.Lerasle, S.Arlot

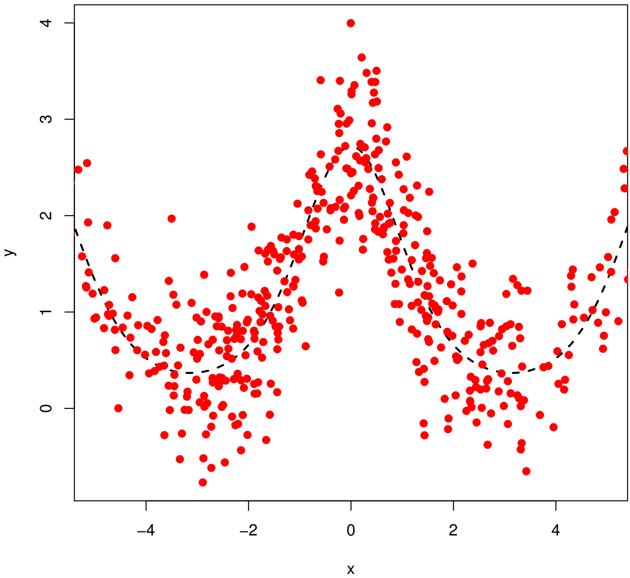
Université du Luxembourg

May 25, 2021

Plan

- Presentation of Agghoo
- Theoretical performance
- Simulations in regression
- Classification

1.Presentation of Agghoo: Regression



$\mathcal{X}, \mathcal{Y}$ are measurable spaces, P a distribution on $\mathcal{X} \times \mathcal{Y}$, and $\gamma : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a measurable *contrast function*. For ex:

- $\mathcal{Y} = \{-1; 1\}$ and $\gamma(u, y) = \mathbb{I}_{u \neq y}$ in classification
- $\mathcal{Y} = \mathbb{R}$ and $\gamma(u, y) = |y - u|$ in median regression.

Definition

- A *predictor* is a measurable function $t : \mathcal{X} \to \mathcal{Y}$.
- The *excess risk* of a predictor f is defined by

$$\ell(s, t) = \mathbb{E}\left[\gamma(f(X), Y)\right] - \inf_{g \text{ predictor}} \mathbb{E}\left[\gamma(g(X), Y)\right]$$

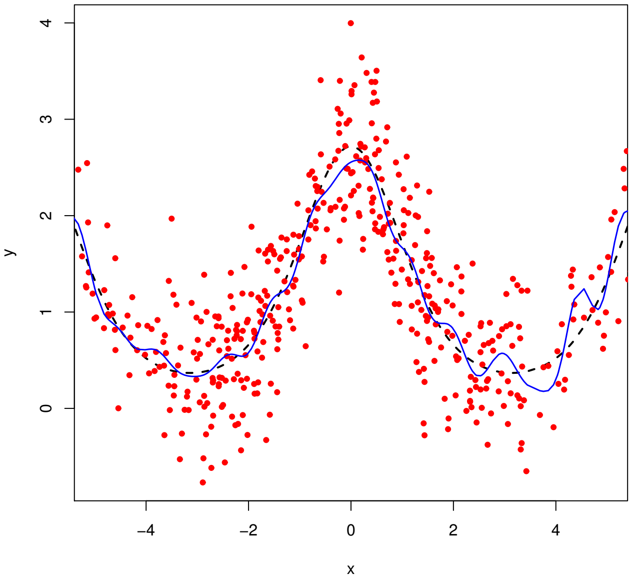
1.Presentation of Agghoo: notations

Fix an integer n. Denote

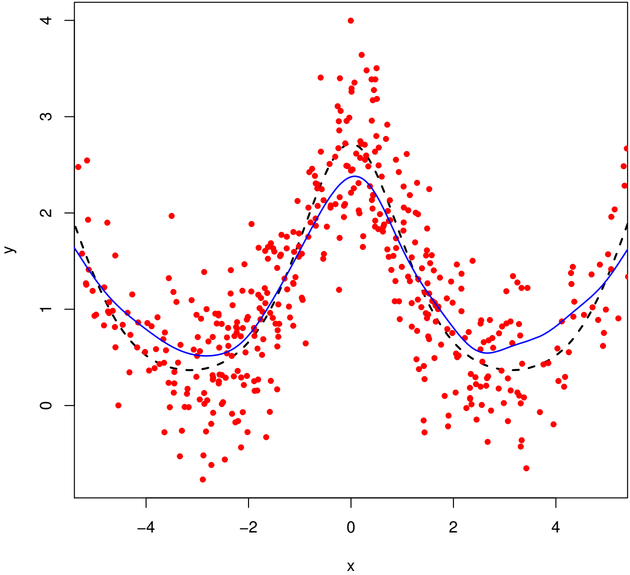
Definition

- $D_n = (X_i, Y_i)_{1 \leq i \leq n} \sim P^{\otimes n}$ a *sample* of size n.
- For a sample D_n and $T \subset \{1, \ldots, n\}$, $D_n^T = (X_j, Y_j)_{j \in T}$.
- $\mathcal{A} : \cup_{k \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^k \to$ predictors a *learning rule*.

1.Presentation of Agghoo: example estimator



1.Presentation of Agghoo: example estimator



1.Presentation of Agghoo: hyperparameter selection

- Given learning rules $(\mathcal{A}_m)_{m\in\mathcal{M}}$ and a sample D_n, choose parameter m.
- Optimal choice m_* realizes $\inf_{m\in\mathcal{M}} \ell(s, \mathcal{A}_m(D_n))$.

Examples:

- $m = k$ in $k-$NN for classification.
- regularization parameter, $m = \lambda$ (Lasso, Ridge, SVM...)

1.Presentation of Agghoo: CV risk estimation

Definition

Fix a sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$. Let $T \subset \{1, \dots, n\}$. For any learning rule $\mathcal{A}$,

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

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

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

1.Presentation of Agghoo: CV selection

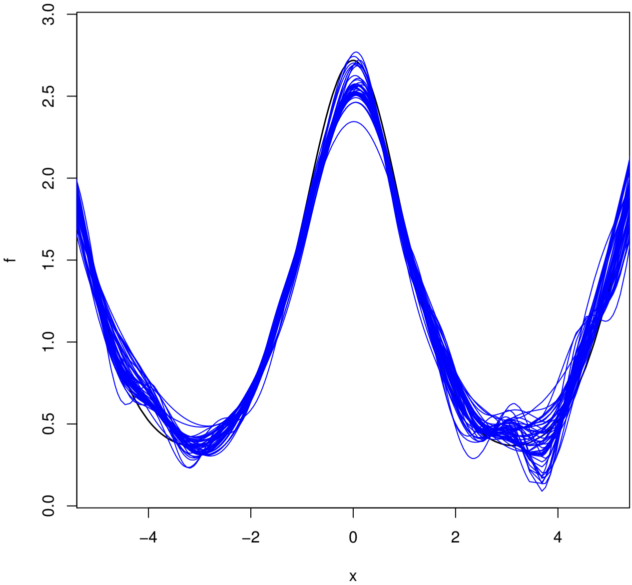
Idea: estimate $\ell(s, \mathcal{A}_m(D))$ and optimize in m.

Definition

Let $\mathcal{M}$ be a set of hyperparameters.

- A *hold out predictor* is $\widehat{f}_T^{\mathrm{ho}} = \mathcal{A}_{\hat{m}_T^{ho}}(D_n^T)$ where $\hat{m}_T^{ho} = \mathrm{argmin}_{m \in \mathcal{M}} \mathrm{HO}_T(\mathcal{A}_m)$.
- A *CV predictor* is defined by $\widehat{f}_T^{CV} = \mathcal{A}_{\hat{m}_T^{CV}}(D_n)$ where $\hat{m}_T^{CV} = \mathrm{argmin}_{m \in \mathcal{M}} CV_{\mathcal{T}}(\mathcal{A}_m)$

1.Presentation of Agghoo: variability of the hold out



Idea: *aggregate* several hold out estimators.

Definition

Assume $\mathcal{Y}$ is convex (regression...).
$T_1, ..., T_V$ i.i.d $\sim \mathcal{U}(\{T \subset \{1, .., n\} : |T| = \lfloor \tau n \rfloor\})$.

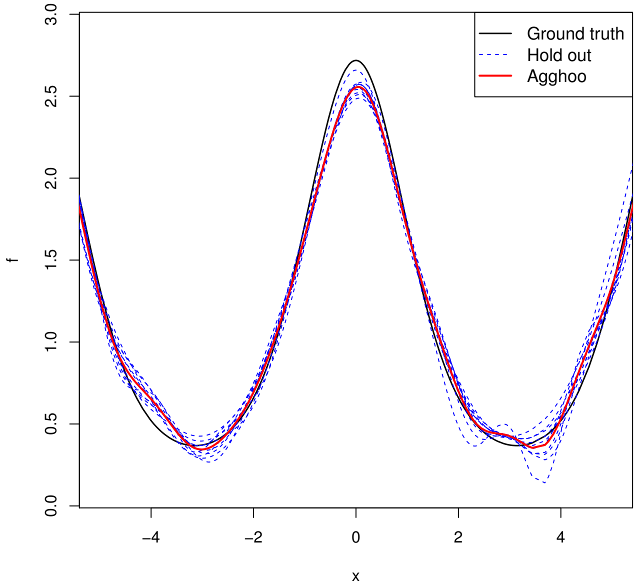
$$\widehat{f}^{\mathrm{ag}}_{\tau,V} = \frac{1}{V}\sum_{i=1}^{V} \widehat{f}^{\mathrm{ho}}_{T_i}$$

Parameters are V and τ.

V-fold CV $\Longrightarrow$ V-fold aggregation.

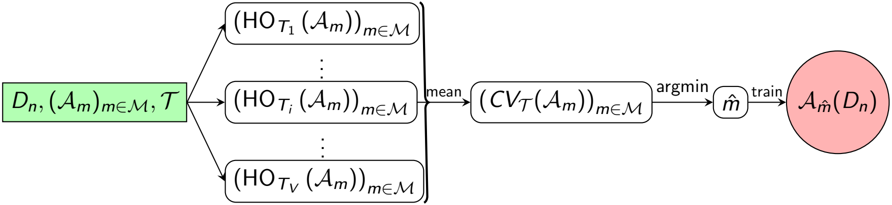
$CV_{\mathcal{T}} \Longrightarrow \widehat{f}^{\mathrm{ag}}_{\mathcal{T}}$.

1.Presentation of Agghoo: an example

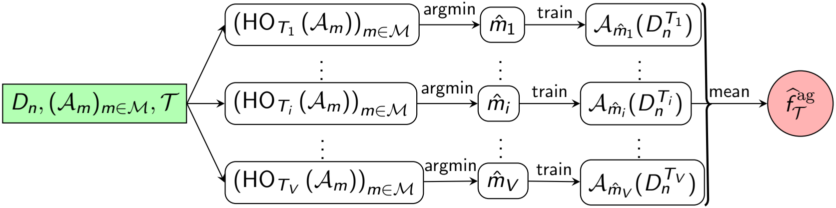


1.Presentation of Agghoo: comparison with cross-validation

Cross-validation:



Agghoo:



2.Theoretical performance: comparison with hold out

Agghoo is safe:

- $\widehat{f}^{\mathrm{ag}}_{\mathcal{T},1} = \widehat{f}^{\mathrm{ho}}_{T_1}$ is just the hold out.
- $\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T},V})\right] \leq \mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1})\right]$ if γ is convex.
- For any $k \geq 1$, $\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\mathcal{T},V})^k\right] \leq \mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ho}}_{T_1})^k\right]$ if γ is convex.

Oracle inequalities for the hold-out imply oracle inequalities for Agghoo.

2.Theoretical performance: oracle inequalities

Aim is to show that

$$\mathbb{E}\left[\ell(s, \widehat{f}^{\mathrm{ag}}_{\tau,V})\right] \leq C_n \mathbb{E}\left[\inf_{m\in\mathcal{M}} \ell(s, \mathcal{A}_m(D_n^T))\right] + r_n$$

where ideally,

- $C_n \to 1$, r_n negligible.
- C_n, r_n only depend on $(\mathcal{A}_m)_{m\in\mathcal{M}}$ through $|\mathcal{M}|$(generality).
- Few assumptions are made on P.

2.Theoretical performance: known results on the hold out

- Generality in classification. Margin adaptivity (G.Blanchard, P.Massart, 2006)
- Least squares regression with bounded Y and linear models (F.Navarro and A.Saumard, 2017)
- Abstract theorems - how do you check the assumptions?
- Specific settings (ex: regressograms, S.Arlot, 2008)

See 2010 survey by S.Arlot and A.Celisse for more references.

A new setting: penalization hyperparameters.

2.Theoretical performance: an RKHS setting

For any sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$, let

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

Penalized empirical risk minimization where

- The model is an RKHS $\mathcal{H}$
- The penalty is $\lambda \|\cdot\|_{\mathcal{H}}^2$
- c is convex and Lipschitz continuous in its first argument

Examples: $c(u, y) = (1 - uy)_+$ (SVM), $c(u, y) = (|u - y| - \varepsilon)_+$ (SVMR).

2.Theoretical performance: an oracle inequality

Now assume that:

- The kernel K is bounded.
- $c(u, y) = \gamma(u, y) = |u - y|$

Theorem

Assume that for a.e. $x \in \mathcal{X}$, for all $u \in \mathbb{R}$,

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

Let $\Lambda \subset \mathbb{R}_+$ and $\lambda_m = \min \Lambda$. Then for all $\theta \in (0; 1]$,

$$(1-\theta)\mathbb{E}[\ell(s, \widehat{f}^{\mathrm{ag}}_{\tau,V})] \le (1+\theta)\mathbb{E}\left[\min_{\lambda \in \Lambda} \ell(s, \mathcal{A}_\lambda(D_{n_t}))\right] + \frac{12\delta}{n_v} + \frac{C_3}{\lambda_m}\left[\frac{\log^2 n_v}{\theta^3 n_v^2} + \frac{\log^{\frac{3}{2}} |\Lambda|}{\theta n_v \sqrt{n_t}}\right]$$

where $n_t = \lfloor \tau n \rfloor$ and $n_v = n - n_t$.

Size of λ_m?

- *Universal consistency* for $\lambda_n \to 0$, $\lambda_n >> \frac{1}{\sqrt{n}}$ (Steinwart and Christmann, Theorem 9.6)
- Gaussian kernel K, $\lambda_n \approx \frac{1}{n}$ minimax-optimal (Eberts and Steinwart, 2013)
- In practice: libsvm, default $\lambda_n \approx \frac{1}{n}$.

State-of-the art on the hold-out I

Conditionnally on $D_n^{T_1}$, $t_m = \mathcal{A}_m(D_n^{T_1})$ deterministic function. Let

$$\hat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} P_n \gamma(t_m, \cdot) \text{ (hold-out)}.$$

Definition

Let w be a function. If $x \mapsto \frac{w(x)}{x}$ non-increasing, for any $\lambda > 0$, let $\delta(w, \lambda)$ solve $w(x) = \lambda x^2$.

State-of-the-art on the hold-out II

Theorem (Massart, St-Flour lecture notes, 2003)

Assume that

$$\mathrm{Var}_P\left(\gamma(t_m,\cdot)-\gamma(t_{m'},\cdot)\right) \le \left(w\left(\sqrt{\ell(s,t_m)}\right)+w\left(\sqrt{\ell(s,t_{m'})}\right)\right)^2$$

$$\left\|\gamma(t_m,\cdot)-\gamma(t_{m'},\cdot)\right\|_\infty \le B$$

where w non-decreasing, $x \mapsto \frac{w(x)}{x}$ non-increasing. With probability $\geq 1-e^{-x}$, for any $\theta \in [0;1]$,

$$(1-\theta)\ell(s,t_{\widehat{m}}) \le (1+\theta)\min_{m\in\mathcal{M}} \ell(s,t_m) + c_1 \frac{\log|\mathcal{M}|+x}{\theta}\delta(w,\sqrt{n})^2 + c_2(\log|\mathcal{M}|+x)\frac{B}{n}.$$

Weaker assumptions

- w non-decreasing, $\frac{w(x)}{x}$ non-increasing $\to$ w non-decreasing.
- $$\begin{aligned} \delta(w,\lambda) \quad & \text{solves } w(\delta) = \lambda\delta^2 \to \\ \delta(w,\lambda) \quad & = \inf\left\{\delta \geq 0 : \forall u \geq \delta, w(u) \leq \lambda u^2\right\}. \end{aligned}$$
- $$\begin{aligned} \|\gamma(t_m,\cdot) - \gamma(t_{m'},\cdot)\|_\infty \quad & \leq B \to (H_\infty) : \\ \|\gamma(t_m,\cdot) - \gamma(t_{m'},\cdot)\|_\infty \quad & \leq \left(w_1(\sqrt{\ell(s,t_m)}) + w_1(\sqrt{\ell(s,t_{m'})})\right)^2 \end{aligned}$$

General result on the hold-out

With probability $\geq 1 - e^{-x}$, for any $\theta \in [0;1]$,

$$(1-\theta)\ell(s, t_{\hat{m}}) \leq (1+\theta) \min_{m \in \mathcal{M}} \ell(s, t_m) + c_1 r_1(\theta, x + \log|\mathcal{M}|)$$
$$+ c_2 r_2(\theta, x + \log|\mathcal{M}|) \text{ où}$$

$\lVert\cdot\rVert_\infty$	w	$r_1(\theta, y)$	$r_2(\theta, y)$
borné B	$\frac{w(u)}{u} \downarrow$	$\frac{y}{\theta}\delta^2(w, \sqrt{n})$	$\frac{By}{n}$
(H_∞)	$\frac{w(u)}{u} \downarrow$	$\frac{y}{\theta}\delta^2(w, \sqrt{n})$	$\frac{y^2}{\theta}\delta^2(w_1, n)$
(H_∞)	Toute $w \uparrow$	$\theta\delta^2\left(w, \frac{\theta}{2}\sqrt{\frac{n}{y}}\right)$	$\theta^2\delta^2\left(w_1, \frac{\theta^2}{4}\frac{n}{y}\right)$

Bounds for different ws can be combined.

Application to kernel methods

Let $\lambda < \mu$, $\gamma(t,(x,y)) = c(t(x),y)$ (to simplify), $\hat{t}_\lambda = \mathcal{A}_\lambda(D_n^T)$.

- $\left\|\gamma(\hat{t}_\mu,\cdot) - \gamma(\hat{t}_\lambda,\cdot)\right\|_\infty \leq \left\|\hat{t}_\lambda - \hat{t}_\mu\right\|_\infty$ (γ 1-Lipschitz).
- $\left\|\hat{t}_\lambda - \hat{t}_\mu\right\|_\infty \leq \kappa \left\|\hat{t}_\lambda - \hat{t}_\mu\right\|_{\mathcal{H}}$ (RKHS, bounded kernel)
- $\left\|\hat{t}_\lambda - \hat{t}_\mu\right\|_{\mathcal{H}}^2 \leq \left\|\hat{t}_\lambda\right\|_{\mathcal{H}}^2 - \left\|\hat{t}_\mu\right\|_{\mathcal{H}}^2$ ($\mathcal{H}$ Hilbert, c convex)
- By definition of $\hat{t}_\lambda$,

$$\begin{aligned}
\lambda\left(\|\hat{t}_\lambda\|_{\mathcal{H}}^2 - \|\hat{t}_\mu\|_{\mathcal{H}}^2\right) &\leq P_n^T\gamma(\hat{t}_\mu,\cdot) - P_{n_t}\gamma(\hat{t}_\lambda,\cdot) \\
&\leq \ell(s,\hat{t}_\mu) + (P_n^T - P)\left[\gamma(\hat{t}_\mu,\cdot) - \gamma(\hat{t}_\lambda,\cdot)\right].
\end{aligned}$$

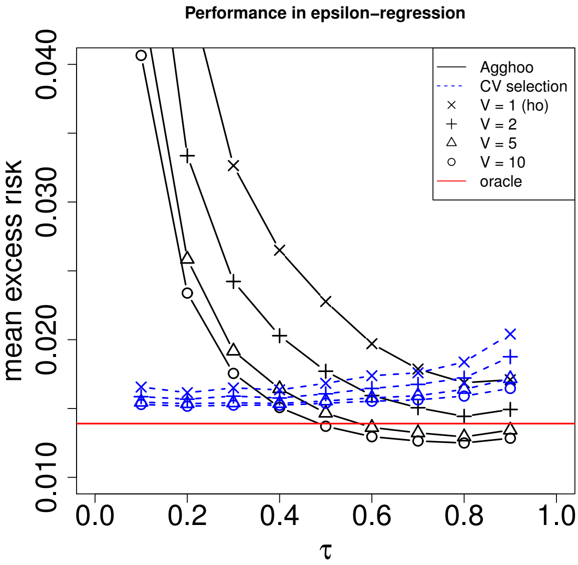
From theory to practice

Questions not answered by theory:

- What is the effect of aggregation?
- Is Agghoo competitive with cross-validation?
- How should τ, V be chosen?

3. Simulations in regression: description

- Kernel rules $(\mathcal{A}_\lambda)_{\lambda \in \Lambda}$ with Gaussian kernel $k(x,y) = e^{-2(x-y)^2}$, cost $c(u,y) = (|u-y| - 0.25)_+$ and grid $\Lambda = \left\{ \frac{2^{j-1}}{500\tau n} | 1 \leq j \leq 17 \right\}$
- $Y = e^{\cos(X)} + \xi$ where $X \sim \mathcal{N}(0,\pi)$, $\xi \sim \mathcal{N}\left(0, \frac{1}{2}\right)$.
- I.i.d sample of size $n = 500$.
- Risk evaluation using $\gamma(u,y) = |u-y|$.


Performance in epsilon–regression
mean excess risk
0.010
0.020
0.030
0.040
0.0
0.2
0.4
0.6
0.8
1.0
τ
Agghoo
CV selection
V = 1 (ho)
V = 2
V = 5
V = 10
oracle

What can be done if $\mathcal{Y}$ is not convex? $\mathcal{Y} = \{0; 1\}$, $\gamma(u, y) = \mathbb{I}_{u \neq y}$.

Definition

Majority vote: Let $(f_i)_{i=1..V}$ be predictors.

$$\mathrm{maj}\left(f_1, \ldots, f_V\right) = x \to \operatorname*{argmax}_{y \in \{0;1\}} \left|\{i : f_i(x) = y\}\right|$$

Majhoo: $\widehat{f}_{T_1}^{\mathrm{ho}}, \ldots, \widehat{f}_{T_v}^{\mathrm{ho}} \to \mathrm{maj}\left(\widehat{f}_{T_1}^{\mathrm{ho}}, \ldots, \widehat{f}_{T_v}^{\mathrm{ho}}\right)$.

Comparison with hold-out: a factor 2 is lost:

$$\forall k \geq 1, \mathbb{E}\left[\ell(s, \widehat{f}_T^{\mathrm{maj}})^k\right] \leq 2^k \mathbb{E}\left[\ell(s, \widehat{f}_{T_1}^{\mathrm{ho}})^k\right].$$

Using Massart (2007) yields for example:

Theorem

Let $\eta(x) = \mathbb{P}[Y = 1 | X = x]$ Suppose that

$$\forall h \in [0;1], \mathbb{P}[|2\eta(X) - 1| \leq h] \leq ch^\beta$$

(margin hypothesis). Then

$$\mathbb{E}[\ell(s, \widehat{f}_{\tau,V}^{\mathrm{ag}})] \leq 3\mathbb{E}\left[\inf_{m \in \mathcal{M}} \ell(s, \mathcal{A}(D_n^{T_1}))\right] + \frac{29c^{\frac{1}{\beta+2}} \log(e|\mathcal{M}|)}{(1-\tau)^{\frac{\beta+1}{\beta+2}} n^{\frac{\beta+1}{\beta+2}}}$$

4.Classification: simulation description

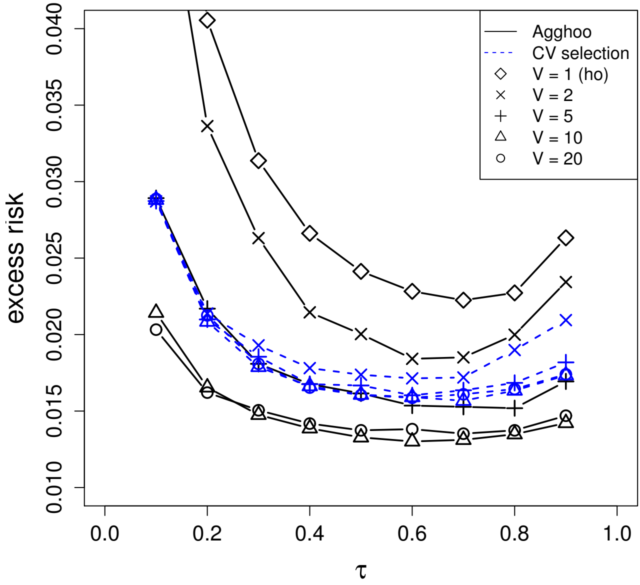
- Binary classification with $0-1$ loss $\gamma(u,y) = \mathbb{I}_{u \neq y}$.
- k-NN rules for $k \in \{k | 1 \leq k \leq n, k \text{ odd.}\}$
- I.i.d sample, $n = 500$, with distribution

$$X \sim \mathcal{U}([0;1]^2)$$

$$\mathbb{P}(Y = 1|X) = \sigma\left(\frac{g(X) - b}{\lambda}\right) \quad \text{where} \quad \sigma(u) = \frac{1}{1 + e^{-u}},$$

$$g(u,v) = e^{-(u^2+v)^3} + u^2 + v^2$$

4.Classification: simulation results



Conclusion

- Agghoo is an alternative to cross-validation.
- It verifies an oracle inequality in regularized kernel regression and classification.
- Simulations show that Agghoo can perform better than cross-validation in regularized kernel regression and classification.

Article reference:

G. Maillard, S. Arlot and M. Lerasle, Aggregated Hold-out, *Journal of Machine Learning Research*, (2021) v.22 , p.1 - 55

Preprint can be found at

`https://hal.archives-ouvertes.fr/hal-02273193`