Cross validation for rare events

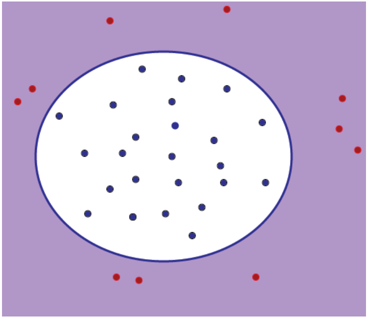
Anass Aghbalou

Telecom Paristech

June 15, 2022

Outline

Extreme value theory



Goal: Modeling extreme events.
Applications: Risk management, insurance, environment, etc.

Motivation

- Cross validation (CV) is widely used for risk estimation/hyper parameter tuning.
- Many empirical evidences advocating the use of CV.
- Numerous theoretical guarantees insuring the consistency of cross validation in multiple frameworks :
    1. ERM algorithms.
    2. Stable learners (*e.g.* SVM, t-nearest-neighbors, LASSO,...)

However Failure/inefficiency of cross validation in other frameworks such as : Regression (Shao 1997), Classification (Yang 2006), Density estimation (Arlot 2008; Arlot and Lerasle 2016).

Motivation

- Cross validation (CV) is widely used for risk estimation/hyper parameter tuning.
- Many empirical evidences advocating the use of CV.
- Numerous theoretical guarantees insuring the consistency of cross validation in multiple frameworks :
  1. ERM algorithms.
  2. Stable learners (*e.g.* SVM, t-nearest-neighbors, LASSO,...)

Our goal: Exploring the question of possible theoretical guarantees/pitfalls for CV estimates in rare regions .

.

Problem settings

- Supervised classification $O = (X, Y)$. The set of classifiers $g \in \mathcal{G}$ has a finite Vapnik-Chervonenkis dimension .
- Choice of classifier: Dataset $\mathcal{D}_n = (O_1, O_2, \ldots, O_n) \in \mathcal{Z}^n$, decision rule (algorithm) $\Psi : \mathcal{Z}^n \to \mathcal{G}$.
- Evaluation: Positive and bounded cost function c, risk of classifier $\mathcal{R}(g) = \mathbb{E}\big[c(g, O)\big]$.

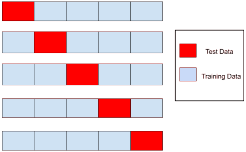
Cross validation

- $\widehat{\mathcal{R}}(g, S) = \frac{1}{n_S} \sum_{i \in S} c(g, O_i)$. Empirical risk on a sample $S \subset \{1, 2, \ldots, n\}$.

- CV estimate of an algorithm Ψ

$$\widehat{\mathcal{R}}_{\mathrm{CV}}(\Psi, V_{1:K}) = \frac{1}{K} \sum_{j=1}^{K} \widehat{\mathcal{R}}\big[\Psi(T_j), V_j\big].$$

$T_j = \{1, 2, \ldots, n\} \setminus V_j$. K number of folds, $K = n$ for *l-o-o*.



Cross validation for extreme regions

- Extreme region $(\mid \|X\| \geq t_\alpha)$, t_α such as $\mathbb{P}(\|X\| \geq t_\alpha) = \alpha \to 0$ and $\alpha n \to \infty$. Typically $\alpha = \frac{1}{\sqrt{n}}$.
- Extreme true risk $\mathcal{R}_\alpha(g) = \mathbb{E}\big[c(g, O) \mid \|X\| \geq t_\alpha\big]$.
- Extreme empirical risk.
  $\widehat{\mathcal{R}}(g, S) \to \widehat{\mathcal{R}}_\alpha(g, S) = \frac{1}{n_S \alpha} \sum_{i \in S} c(g, O_i) \mathbb{1}_{\{\|X_i\| > \|X_{(\lfloor \alpha n \rfloor)}\}}$.
- Extreme CV estimate.

$$\widehat{\mathcal{R}}_{\mathrm{CV}}(\Psi, V_{1:K}) \to \widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi, V_{1:K}) = \frac{1}{K} \sum_{j=1}^{K} \widehat{\mathcal{R}}_\alpha\big[\Psi(T_j), V_j\big].$$

- Assumption : Ψ_α is an ERM i.e $\Psi_\alpha(S) = \arg\min_{g \in \mathcal{G}} \widehat{\mathcal{R}}_\alpha(g, S)$.

Cross validation - state of the art

- CV estimate of an algorithm Ψ

$$\widehat{\mathcal{R}}_{\mathrm{CV}}(\Psi, V_{1:K}) = \frac{1}{K}\sum_{j=1}^{K}\widehat{\mathcal{R}}\big[\Psi(T_j), V_j\big].$$

$T_j = \{1, 2, \dots, n\} \setminus V_j$. K the number of folds, $K = n$ for the *l-o-o*.

- Assumption : Ψ is an ERM i.e $\Psi(S) = \arg\min_{g\in\mathcal{G}} \widehat{\mathcal{R}}(g, S)$.

Cross validation - state of the art

- CV estimate of an algorithm Ψ

$$\widehat{\mathcal{R}}_{\mathrm{CV}}(\Psi, V_{1:K}) = \frac{1}{K}\sum_{j=1}^{K} \widehat{\mathcal{R}}\big[\Psi(T_j), V_j\big].$$

$T_j = \{1, 2, \ldots, n\} \setminus V_j$. K the number of folds, $K = n$ for the *l-o-o*.

- Assumption : Ψ is an ERM i.e $\Psi(S) = \arg\min_{g\in\mathcal{G}} \widehat{\mathcal{R}}(g, S)$.

Exponential bound for K-fold (Cornec 2017): For any $\delta > 0$ one has with probability 1-δ,

$$|\widehat{\mathcal{R}}_{\mathrm{Kfold}}(\Psi, V_{1:K}) - \mathcal{R}\big[\Psi([n])\big]| \leq M \log\frac{1}{\delta}\sqrt{\frac{\mathcal{V}_{\mathcal{G}} K}{n}}.$$

Where $M > 0$ is universal constant.

Cross validation - state of the art

- CV estimate of an algorithm Ψ

$$\widehat{\mathcal{R}}_{\mathrm{CV}}(\Psi, V_{1:K}) = \frac{1}{K}\sum_{j=1}^{K}\widehat{\mathcal{R}}\big[\Psi(T_j), V_j\big].$$

$T_j = \{1, 2, \dots, n\} \setminus V_j$. K the number of folds, $K = n$ for the *l-o-o*.

- Assumption : Ψ is an ERM i.e $\Psi(S) = \arg\min_{g\in\mathcal{G}} \widehat{\mathcal{R}}(g, S)$.

Polynomial bound for l-o-o (Kearns 1999): For any $\delta > 0$ one has with probability 1-δ,

$$|\widehat{\mathcal{R}}_{\mathrm{loo}}(\Psi, V_{1:n}) - \mathcal{R}\big[\Psi([n])\big]| \leq \frac{M}{\delta}\sqrt{\frac{\mathcal{V}_{\mathcal{G}}}{n}}.$$

Where $M > 0$ is universal constant.

Technical difficulty

- To sum up, existing results insures that

$$\mathrm{Err} = \left|\widehat{\mathcal{R}}_{\mathrm{CV}} - \mathcal{R}\right| = \mathcal{O}\left(\sqrt{\frac{1}{n}}\right).$$

- Dividing both terms by the normalization constant α yields,

$$\mathrm{Err}_{\alpha} = \left|\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{\alpha}\right| = \mathcal{O}\left(\frac{1}{\alpha\sqrt{n}}\right).$$

$\rightarrow$ Vacuous bound when $\alpha \geq \sqrt{1/n}$.

Bernstein inequality extension

Proposition

Let $f : \mathcal{Z}^n \to \mathbb{R}$ be some measurable function , let $Z = f(O_1, O_2, \dots, O_n)$ and define for $l \in [n]$: Then we have

$$\mathbb{P}(Z - \mathbb{E}(Z) > t) \leq \exp\left(\frac{-t^2}{2(\sigma^2 + Dt/3)}\right).$$

- σ^2 reflects the variance of Z.
- D reflects maximal deviations on Z

Exponential bounds for CV schemes

Goal: Estimating $\mathcal{R}_\alpha\big[\Psi([n])\big] = \mathbb{E}\big[c(\Psi([n]), O) \mid \|X\| \geq t_\alpha\big]$.

Error decomposition

$$\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\big(\Psi_\alpha([n])\big)\right| \leq \mathrm{D}_{\mathrm{t}_\alpha} + \mathrm{D}_{\mathrm{cv}} + \mathrm{Bias},$$

- Quantile estimation error: $\mathrm{D}_{\mathrm{t}_\alpha}$.
- CV Deviations: D_{cv}.
- CV Bias: Bias.

Exponential bounds for K-fold

Error decomposition

$$\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\big(\Psi_\alpha(S_n)\big)\right| \leq \mathrm{D}_{\mathrm{t}_\alpha} + \mathrm{D}_{\mathrm{cv}} + \mathrm{Bias},$$

Controling terms

$$\begin{cases} \mathrm{D}_{\mathrm{t}_\alpha} = \mathcal{O}(\log(1/\delta)\sqrt{\frac{1}{n\alpha}}). \\ \mathrm{D}_{\mathrm{cv}} = \mathcal{O}(\log(1/\delta)\sqrt{\frac{\mathcal{V}_{\mathcal{G}}}{n_V\alpha}}). \to \textit{Dominant term}. \\ \mathrm{Bias} = \mathcal{O}(\log(1/\delta)\sqrt{\frac{\mathcal{V}_{\mathcal{G}}}{n_T\alpha}}). \end{cases}$$

Exponential bounds for K-fold

Controling terms

$$\begin{cases} \mathrm{D_{t_\alpha}} = \mathcal{O}(\log(1/\delta)\sqrt{\frac{1}{n\alpha}}). \\ \mathrm{D_{cv}} = \mathcal{O}(\log(1/\delta)\sqrt{\frac{\mathcal{V}_\mathcal{G}}{n_V\alpha}}). \to \textit{Dominant term}. \\ \mathrm{Bias} = \mathcal{O}(\log(1/\delta)\sqrt{\frac{\mathcal{V}_\mathcal{G}}{n_T\alpha}}). \end{cases}$$

K-fold consistency

$$n_V = n/K \implies \left|\widehat{\mathcal{R}}_{\mathrm{Kfold},\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\big(\Psi_\alpha(S_n)\big)\right| = \mathcal{O}(\log(1/\delta)\sqrt{\frac{\mathcal{V}_\mathcal{G}K}{n\alpha}})$$

Exponential bounds for K-fold

Controling terms

$$\begin{cases} \mathrm{D}_{\mathrm{t}_\alpha} = \mathcal{O}(\log(1/\delta)\sqrt{\frac{1}{n\alpha}}). \\ \mathrm{D}_{\mathrm{cv}} = \mathcal{O}(\log(1/\delta)\sqrt{\frac{\mathcal{V}_{\mathcal{G}}}{n_V\alpha}}). \rightarrow \textit{Dominant term}. \\ \mathrm{Bias} = \mathcal{O}(\log(1/\delta)\sqrt{\frac{\mathcal{V}_{\mathcal{G}}}{n_T\alpha}}). \end{cases}$$

l-o-o CV

$$n_V = 1 \implies \textit{Trivial Bound !}$$

Polynomial bounds for *l-p-o* CV

Error decomposition

$$\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\big(\Psi_\alpha([n])\big)\right| \leq \underbrace{\mathrm{D}_{\mathrm{t}_\alpha}}_{\propto\sqrt{\frac{1}{n}}} + \underbrace{\mathrm{D}_{\mathrm{cv}}}_{\propto\sqrt{\frac{1}{n_V}}} + \underbrace{\mathrm{Bias}}_{\propto\sqrt{\frac{1}{n_T}}} .$$

Lemma

For all $t > 0$, one has

$$\mathbb{P}(\mathrm{D}_{\mathrm{cv}} \geq t) \leq \frac{M}{t}\sqrt{\frac{\mathcal{V}_{\mathcal{G}}}{n\alpha}}.$$

For some universal constant $M > 0$.

Polynomial bounds for *l-o-o* CV

l-o-o CV consistency

With probability $1-\delta$, one has

$$|\widehat{\mathcal{R}}_{\mathrm{loo},\alpha}(\Psi, V_{1:n}) - \mathcal{R}_\alpha\left[\Psi([n])\right]| \leq \frac{C}{\delta}\sqrt{\frac{\mathcal{V}_{\mathcal{G}}}{n\alpha}}$$

For some universal constant $C > 0$.

Applications

- Model selection : choosing the optimal penality parameter for RERM.
- Feature selection.
- Imbalanced classification.

Numerical illustration

- Toy example: simulated data, dimension 1, student distribution, threshold classifier, Hamming loss.
- $n = 2.10^4$, $\alpha \in [1\%, 20\%]$.
- Average absolute error of the K-fold (K = 10) and upper quantile atlevel 0.90, logarithmic scale, over 10^4 experiments.

References I

S. Arlot. V-fold cross-validation improved: V-fold penalization. 40 pages, plus a separate technical appendix., Feb. 2008. URL https://hal.archives-ouvertes.fr/hal-00239182.

S. Arlot and M. Lerasle. Choice of v for v-fold cross-validation in least-squares density estimation. *Journal of Machine Learning Research*, 17(208):1–50, 2016. URL http://jmlr.org/papers/v17/14-296.html.

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–242, 1997. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24306073.

Y. Yang. Comparing learning methods for classification. *Statistica Sinica*, 16(2):635–657, 2006. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24307562.