Cross-validation Confidence Intervals for Test Error

Lester Mackey

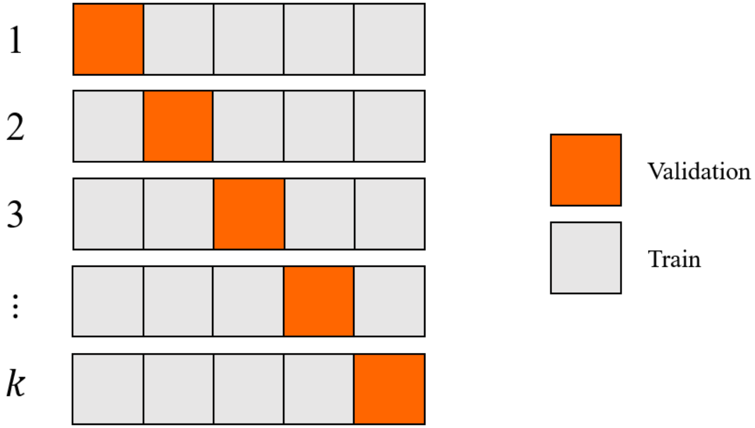
Microsoft Research New England

August 9, 2021

Joint work with **Pierre Bayle** (Princeton University), **Alexandre Bayle** (Harvard University), and **Lucas Janson** (Harvard University).

How good is my learning algorithm?

Cross-validation (CV) [Stone, 1974, Geisser, 1975]



- Divide data into k validation sets
- Fit k prediction rules, each with one validation set held out
- Evaluate each prediction rule on its held-out set
- Average the k error estimates

Pros: Unbiased for test error & lower variance than single train-test split

High-stakes Applications

Need: Test error confidence intervals to quantify uncertainty

Prediction of cancer outcome with microarrays: a multiple random validation strategy

Stefan Michiels, Serge Koscielny, Catherine Hill

Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study

Romain Pirracchio, Maya L Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, Mark J van der Laan

Problem: CV distribution is complex & existing intervals often invalid

"The widely used approach of basing confidence intervals on an independent binomial assumption of the leave-one-out cross-validation errors results in serious under-coverage of the true prediction error."

Calculating Confidence Intervals for Prediction Error in Microarray Classification Using Resampling

Wenyu Jiang, Sudhir Varma and Richard Simon

Is algorithm A actually better than algorithm B?

Need: Trustworthy hypothesis tests of error improvement

Problem: Standard tests (like the cross-validated t-test [Dietterich, 1998], the repeated train-validation t-test [Nadeau and Bengio, 2003], and the 5×2-fold CV test [Dietterich, 1998]) do not appropriately account for dependence and have no correctness guarantees

Our Contributions

ALGORITHMIC STABILITY

CV CENTRAL LIMIT THEOREM
+
CONSISTENT VARIANCE ESTIMATOR

CV CONFIDENCE INTERVALS FOR TEST ERROR

CV TESTS FOR ALGORITHM IMPROVEMENT

Problem Setup

Given

- **Datapoints $Z_1, \ldots, Z_n$**
  - Often each $Z_i = (X_i, Y_i)$ with covariates X_i and response Y_i
  - For any vector B of indices, Z_B denotes the corresponding vector of datapoints
- **Loss function $h_n(Z_i, Z_B)$**: error when training on Z_B and testing on Z_i
  - **Regression:** $h_n(Z_i, Z_B) = (Y_i - \hat{f}(X_i; Z_B))^2$ for $\hat{f}(\cdot; Z_B)$ trained on Z_B
  - **Classification:** $h_n(Z_i, Z_B) = \mathbb{1}[Y_i \neq \hat{f}(X_i; Z_B)]$
  - **Algorithm comparison:** $h_n(Z_i, Z_B) = \mathbb{1}[Y_i \neq \hat{f}_1(X_i; Z_B)] - \mathbb{1}[Y_i \neq \hat{f}_2(X_i; Z_B)]$
- **Validation sets $\{B'_j\}_{j=1}^k$** and associated **training sets $\{B_j\}_{j=1}^k$**
  - Validation sets partition datapoint indices $\{1, \ldots, n\}$ into k folds; k can grow with n

Goal: Characterize the distribution of **cross-validation error**

$$\hat{R}_n \triangleq \frac{1}{n} \sum_{j=1}^k \sum_{i \in B'_j} h_n(Z_i, Z_{B_j})$$

Why CV Error?

Cross-validation error: $\boldsymbol{\hat{R}_n \triangleq \frac{1}{n}\sum_{j=1}^{k}\sum_{i\in B_j'} h_n(Z_i, Z_{B_j})}$

- Unbiased estimate of k-fold test error, a common inferential target [Blum, Kalai, and Langford, 1999, Dudoit and van der Laan, 2005, Kale, Kumar, and Vassilvitskii, 2011, Kumar, Lokshtanov, Vassilvitskii, and Vattani, 2013, Austern and Zhou, 2020]
- Lower variance than single train-validation split

k**-fold test error:** $\boldsymbol{R_n \triangleq \frac{1}{n}\sum_{j=1}^{k}\sum_{i\in B_j'} \mathbb{E}[h_n(Z_i, Z_{B_j}) \mid Z_{B_j}]}$

- Average test error of the k prediction rules $\hat{f}(\cdot; Z_{B_j})$

Goal: Establish a central limit theorem for $\hat{R}_n - R_n$

Algorithmic Stability

Stability

How much does prediction performance change when one training point changes?

- Uniform stability [Bousquet and Elisseeff, 2002]: worst-case change in loss h_n
- Mean-square stability [Kale, Kumar, and Vassilvitskii, 2011]: mean-square change in loss h_n
- **Loss stability** [Kumar, Lokshtanov, Vassilvitskii, and Vattani, 2013]
  - Mean-square change in loss *difference* $h_n(Z_0, Z_B) - \mathbb{E}[h_n(Z_0, Z_B) \mid Z_B]$

Asymptotic Normality of CV

CV Central Limit Theorem [Bayle, Bayle, Janson, and Mackey, 2020]

Suppose $Z_0, Z_1, \cdots, Z_n$ are i.i.d., and define the expected loss function

$$\bar{h}_n(Z_0) = \mathbb{E}[h_n(Z_0, Z_{1:n(1-1/k)}) \mid Z_0] \quad \text{with} \quad \sigma_n^2 = \mathrm{Var}(\bar{h}_n(Z_0)).$$

If loss stability $= o(\sigma_n^2/n)$ and $(\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)])^2/\sigma_n^2$ is uniformly integrable then

$$\frac{\sqrt{n}}{\sigma_n}(\hat{R}_n - R_n) \xrightarrow{d} \mathcal{N}(0, 1).$$

Sufficient condition: $\sup_n \mathbb{E}[|\bar{h}_n(Z_0) - \mathbb{E}[\bar{h}_n(Z_0)]|^\alpha/\sigma_n^\alpha] < \infty$ for some $\alpha > 2$

Many learning algorithms enjoy decaying loss stability

- Stochastic gradient descent on convex and non-convex objectives [Hardt, Recht, and Singer, 2016]
- Empirical risk minimization of strongly convex, Lipschitz objective [Bousquet and Elisseeff, 2002]
  - Note: training objective need not match the validation loss h_n!
- k-nearest neighbor methods [Devroye and Wagner, 1979], even when overfit with 0 training error
- Decision trees [Arsov, Pavlovski, and Kocarev, 2019] and ensemble methods [Elisseeff, Evgeniou, and Pontil, 2005]

Asymptotic Normality of CV: Related Work

Theorem 3 of Dudoit and van der Laan [2005]

- Requires a bounded loss function
- Excludes leave-one-out CV
- Requires prediction rule to be loss-consistent for a risk-minimizing prediction rule

Theorem 4.1 of LeDell, Petersen, and van der Laan [2015]

- Applies only to AUC loss
- Requires bounded number of folds k
- Requires prediction rule to be loss-consistent for a risk-minimizing prediction rule

Theorem 1 of Austern and Zhou [2020]

- Assumes variance parameter $\tilde{\sigma}_n \geq \sigma_n$ converging to a non-zero limit
- Requires $o(1/n)$ mean-square stability and $o(1/n^2)$ 2nd-order mean-square stability
- Assumes learning algorithm is symmetric in the training points

Application: Confidence Intervals for Test Error

Problem

Construct an asymptotically-exact $(1-\alpha)$-confidence interval for k-fold test error R_n

Solution: CV Confidence Interval for Test Error

Under the assumptions of the CV CLT, if a variance estimator $\hat{\sigma}_n^2$ satisfies relative error consistency $(\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1)$, then the interval

$$C_\alpha \triangleq \hat{R}_n \pm q_{1-\alpha/2}\,\hat{\sigma}_n/\sqrt{n}$$

satisfies

$$\lim_{n\to\infty} \mathbb{P}(R_n \in C_\alpha) = 1-\alpha$$

where $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a standard normal distribution

Application: Tests for Algorithm Improvement

Problem

Construct an asymptotically-exact level α test of whether $\mathcal{A}_1$ has smaller k-fold test error than $\mathcal{A}_2$

Solution: CV Test for Improved Test Error

For a target loss function ℓ, define the $\mathcal{A}_1$-$\mathcal{A}_2$ loss difference

$$h_n(Z_0, Z_B) = \ell(Y_0, \hat{f}_1(X_0; Z_B)) - \ell(Y_0, \hat{f}_2(X_0; Z_B)),$$

and consider testing $H_0 : R_n \geq 0$ ($\mathcal{A}_1$ not better) against $H_1 : R_n < 0$ ($\mathcal{A}_1$ is better). Under the assumptions of the CV CLT, if a variance estimator $\hat{\sigma}_n^2$ satisfies relative error consistency ($\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$), then the test

$$\text{REJECT } H_0 \Leftrightarrow \hat{R}_n < q_\alpha \hat{\sigma}_n/\sqrt{n}$$

has asymptotic level α for q_α the α-quantile of a standard normal distribution

Consistent Variance Estimation

Goal: Find a practical estimator $\hat{\sigma}_n^2$ satisfying $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{p} 1$ under weak conditions.

Within-fold variance estimator $\hat{\sigma}^2_{n,in}$

Computes the variance of $h_n(Z_i, Z_{B_j})$ in each fold and takes the average across folds

All-pairs variance estimator $\hat{\sigma}^2_{n,out}$

$$\hat{\sigma}^2_{n,out} \triangleq \frac{1}{n} \sum_{j=1}^{k} \sum_{i \in B_j'} (h_n(Z_i, Z_{B_j}) - \hat{R}_n)^2$$

- Computes the empirical variance of $h_n(Z_i, Z_{B_j})$ across all folds
- Advantage: can also be used for leave-one-out cross-validation

Low computational cost

$\hat{\sigma}^2_{n,in}$ and $\hat{\sigma}^2_{n,out}$ can be computed in $O(n)$ time and in $O(k)$ time if loss is binary

Consistent Variance Estimation

Theorem (Consistent Estimation of CV Variance [Bayle, Bayle, Janson, and Mackey, 2020])

Under exactly the same conditions given for the CV central limit theorem (loss stability $= o(\sigma_n^2/n)$ and uniform integrability), we have

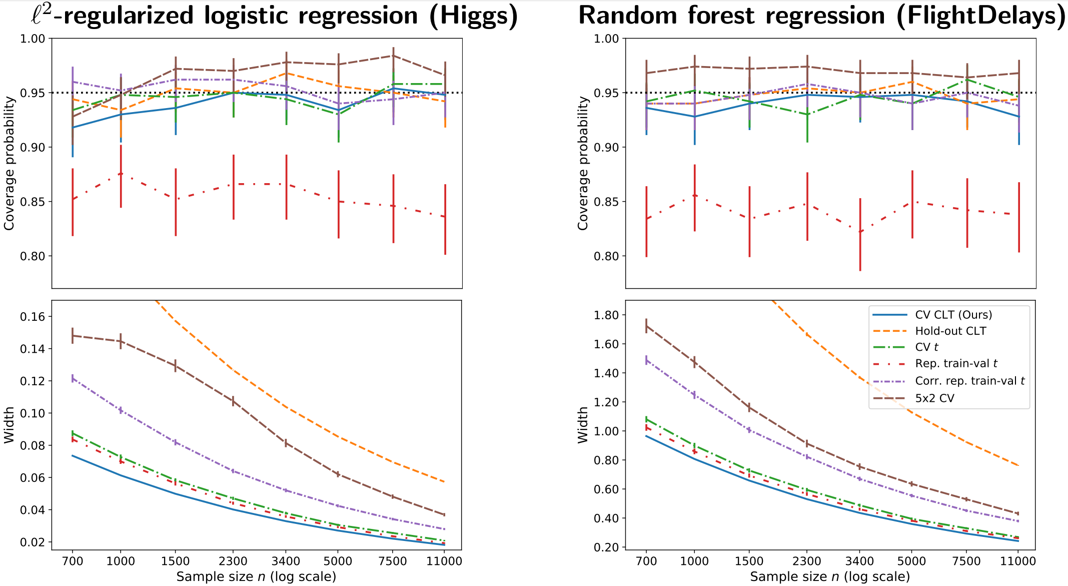
$$\hat{\sigma}_{n,in}^2 / \sigma_n^2 \xrightarrow{L^1} 1.$$

If, additionally, mean-square stability $= o(k\sigma_n^2/n)$, then

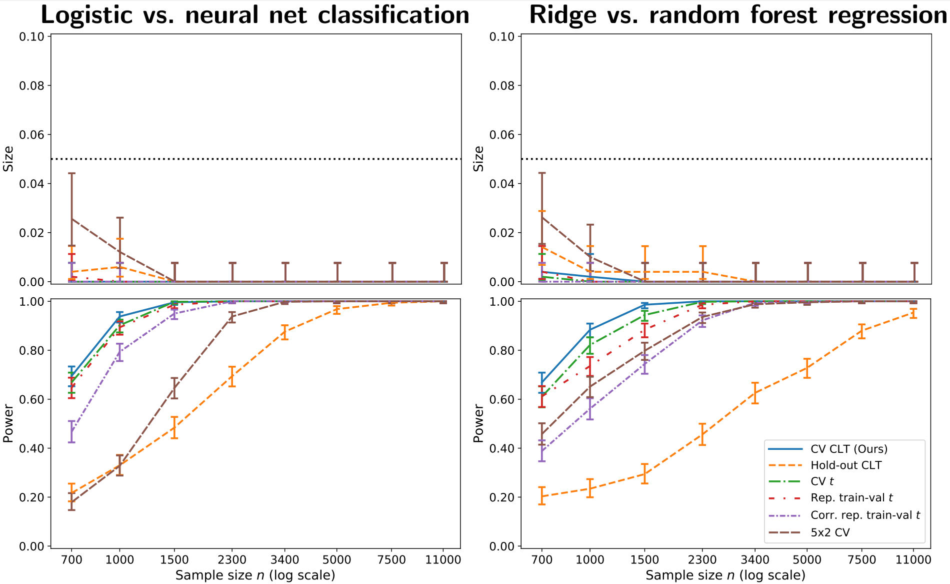
$$\hat{\sigma}_{n,out}^2 / \sigma_n^2 \xrightarrow{L^1} 1.$$

- Mean-square stability condition particularly mild for leave-one-out CV ($k = n$)

Confidence Intervals for Test Error, $1 - \alpha = 0.95$, $k = 10$



Testing for Algorithm Improvement, $\alpha = 0.05$, $k = 10$



Leave-one-out CV Confidence Intervals, $1 - \alpha = 0.95$

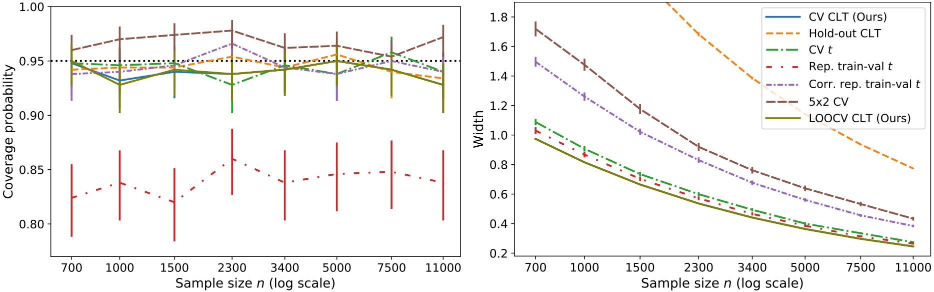
Misconception: Leave-one-out CV (LOOCV, $k = n$) only relevant for small n

Reality

- Ridge regression LOOCV only slightly slower than a single regression
- For many models, LOOCV can be efficiently approximated with only $O(1/n^2)$ error

[Beirami, Razaviyayn, Shahrampour, and Tarokh, 2017, Giordano, Stephenson, Liu, Jordan, and Broderick, 2019, Koh, Ang, Teo, and Liang, 2019, Wilson, Kasy, and Mackey, 2020]

Ridge regression



Conclusions

Summary

- New CV central limit theorem under algorithmic stability
- Consistent estimators of CV variance
- Asymptotically exact confidence intervals and tests for k-fold test error

Opportunities for future work

- Practical valid tests and confidence intervals in the absence of stability
- Analogous tools for *expected* test error $\mathbb{E}[R_n]$ [see, e.g., Austern and Zhou, 2020]

Cross-validation Confidence Intervals for Test Error

Paper: `https://arxiv.org/abs/2007.12671`

Code: `https://github.com/alexandre-bayle/cvci`

References I

N. Arsov, M. Pavlovski, and L. Kocarev. Stability of decision trees and logistic regression. *arXiv preprint arXiv:1903.00816v1*, 2019.

M. Austern and W. Zhou. Asymptotics of Cross-Validation. *arXiv preprint arXiv:2001.11111v2*, 2020.

P. Bayle, A. Bayle, L. Janson, and L. Mackey. Cross-validation confidence intervals for test error. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16339–16350. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/bce9abf229ffd7e570818476ee5d7dde-Paper.pdf.

A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh. On optimal generalizability in parametric learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 3455–3465, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

A. Blum, A. Kalai, and J. Langford. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proc. COLT*, pages 203–208, 1999.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

L. Devroye and T. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, Oct. 1998. ISSN 0899-7667. doi: 10.1162/089976698300017197. URL https://doi.org/10.1162/089976698300017197.

S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2 (2):131–154, 2005.

A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79, Dec. 2005. ISSN 1532–4435.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975.

R. Giordano, W. Stephenson, R. Liu, M. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1139–1147. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/giordano19a.html.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning - Volume 48*, ICML'16, pages 1225–1234. JMLR.org, 2016.

S. Kale, R. Kumar, and S. Vassilvitskii. Cross-validation and mean-square stability. In *Proceedings of the Second Symposium on Innovations in Computer Science (ICS2011)*. Citeseer, 2011.

References II

P. W. Koh, K.-S. Ang, H. H. K. Teo, and P. Liang. On the Accuracy of Influence Functions for Measuring Group Effects. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'19, pages 5254–5264, 2019.

R. Kumar, D. Lokshtanov, S. Vassilvitskii, and A. Vattani. Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pages 27–35, 2013.

E. LeDell, M. Petersen, and M. van der Laan. Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. *Electronic Journal of Statistics*, 9(1):1583–1607, 2015.

C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52(3):239–281, 2003.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

A. Wilson, M. Kasy, and L. Mackey. Approximate cross-validation: Guarantees for model assessment and selection. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4530–4540, Online, 26–28 Aug 2020. PMLR. URL http://proceedings.mlr.press/v108/wilson20a.html.