CBIO meeting

The trouble with cross-validation

Chloé-Agathe Azencott

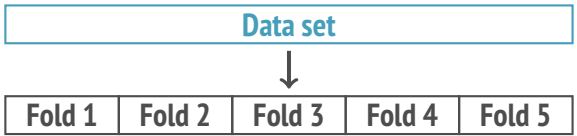
Center for Computational Biology (CBIO)
Mines Paris PSL – Institut Curie – INSERM U900
PSL Research University & PR[AI]RIE, Paris, France

May 27, 2024

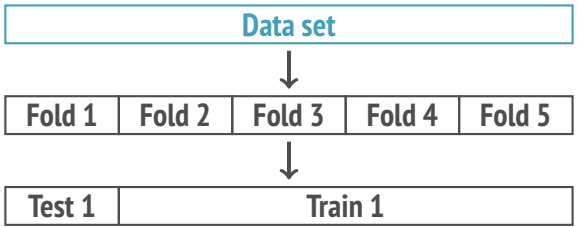
http://cazencott.info chloe-agathe.azencott@mines-paristech.fr @cazencott@lipn.info

Model evaluation by cross-validation

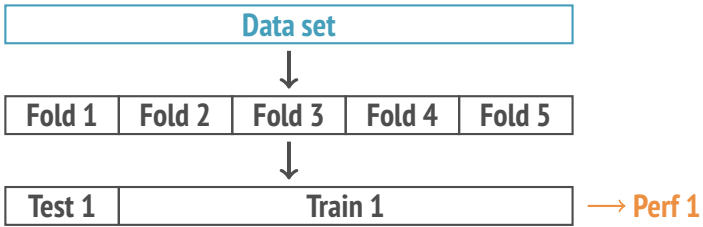
Model evaluation by cross-validation



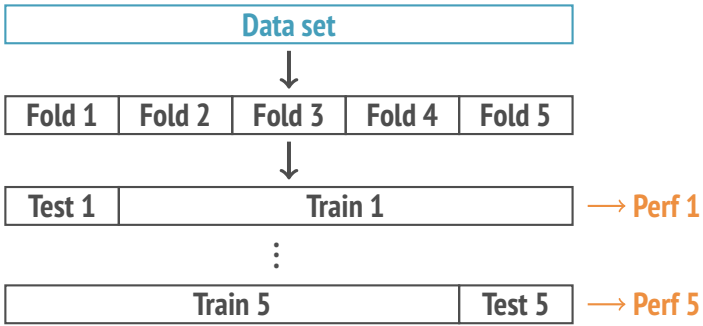
Model evaluation by cross-validation



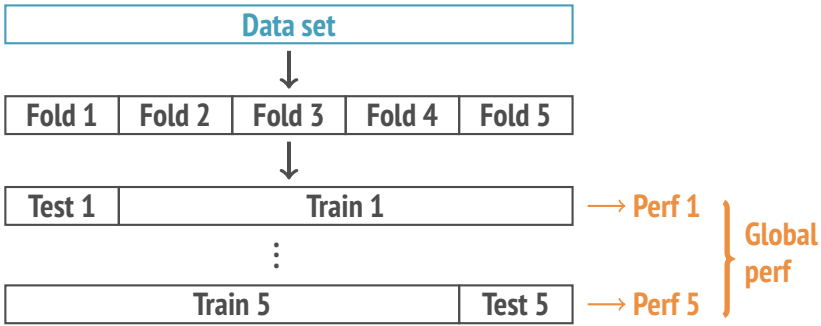
Model evaluation by cross-validation



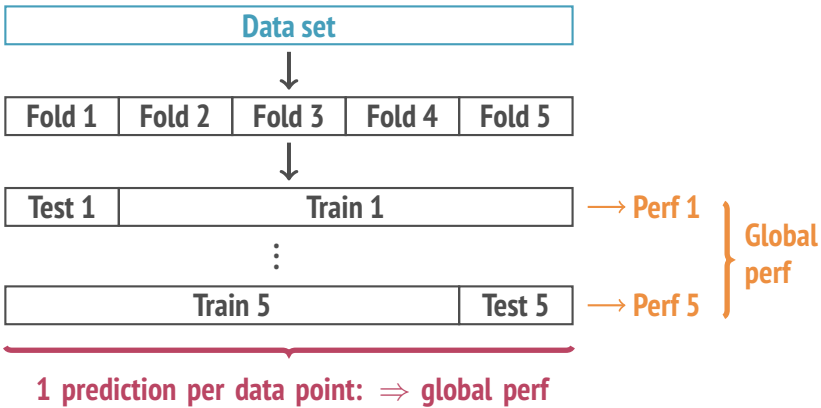
Model evaluation by cross-validation



Model evaluation by cross-validation



Model evaluation by cross-validation



Historical perspective

Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society: Series B, 1974 (Stone 1974)

Historical perspective

Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society: Series B, 1974 (Stone 1974)

– Footnote: “The term **assessment** is preferred to **validation** which has a ring of excessive confidence about it.”

Historical perspective

Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society: Series B, 1974 (Stone 1974)

- Footnote: "The term **assessment** is preferred to **validation** which has a ring of excessive confidence about it."
- The idea of assessing prediction quality on a separate sample dates back to at least 1931 (Wherry 1931)
- "Symposium: The need and means of cross-validation" in 1951

Historical perspective

Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions, Journal of the Royal Statistical Society: Series B, 1974 (Stone 1974)

- Footnote: "The term **assessment** is preferred to **validation** which has a ring of excessive confidence about it."
- The idea of assessing prediction quality on a separate sample dates back to at least 1931 (Wherry 1931)
- "Symposium: The need and means of cross-validation" in 1951
- Followed by a discussion of the paper by Barnard, Atkinson, Chan, Dawid, Baker, Cox, Geisser, Hinkley, Hocking, and Young.

But still a topic of much discussion

1. **Error bar sizes for small sample sizes**

   Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage, 2018 (Varoquaux 2018)

2. **What does cross-validation estimate?**

   Bates, S., Hastie, T., & Tibshirani, R. Cross-validation: what does it estimate and how well does it do it? Journal of the American Statistical Association, 2023 (Bates, Hastie, and Tibshirani 2023)

Issue number 1: error bar size

Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage, 2018 (Varoquaux 2018)

Confidence intervals from cross-validation

- n predictions $\Rightarrow$ n errors $e_1, e_2, \ldots, e_n$

  Assuming they are the realization of n iid square integrable random variables $E_1, E_2, \ldots, E_n$ of expectation μ and variance σ^2,

  by the **central limit theorem:** $\frac{\bar{E}-\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ with $\bar{E} = \frac{1}{n}\sum_{i=1}^{n} E_i$

- hence $\mathbb{P}\left(\bar{E} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{E} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

Confidence intervals from cross-validation

- n predictions $\Rightarrow$ n errors $e_1, e_2, \ldots, e_n$

  Assuming they are the realization of n iid square integrable random variables $E_1, E_2, \ldots, E_n$ of expectation μ and variance σ^2,

  by the **central limit theorem:** $\frac{\bar{E}-\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ with $\bar{E} = \frac{1}{n}\sum_{i=1}^{n} E_i$

- hence $\mathbb{P}\left(\bar{E} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{E} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

$$\boxed{\text{CI}_{95\%} = \bar{e} \pm 1.96\frac{\hat{\sigma}}{\sqrt{n}}}$$

$\bar{n} = \frac{1}{n}\sum_{i=1}^{n} e_i$ $\quad \hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2}$ $\quad \frac{\hat{\sigma}}{\sqrt{n}}$ = standard error

Confidence intervals from cross-validation

- n predictions $\Rightarrow$ n errors $e_1, e_2, \ldots, e_n$

  Assuming they are the realization of n iid square integrable random variables $E_1, E_2, \ldots, E_n$ of expectation μ and variance σ^2,

  by the **central limit theorem:** $\frac{\bar{E}-\mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ with $\bar{E} = \frac{1}{n}\sum_{i=1}^{n} E_i$

- hence $\mathbb{P}\left(\bar{E} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{E} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

$$\boxed{\text{CI}_{95\%} = \bar{e} \pm 1.96\frac{\hat{\sigma}}{\sqrt{n}}}$$

$\bar{n} = \frac{1}{n}\sum_{i=1}^{n} e_i$ $\quad \hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2}$ $\quad \frac{\hat{\sigma}}{\sqrt{n}}$ = standard error

CI width as training size increases

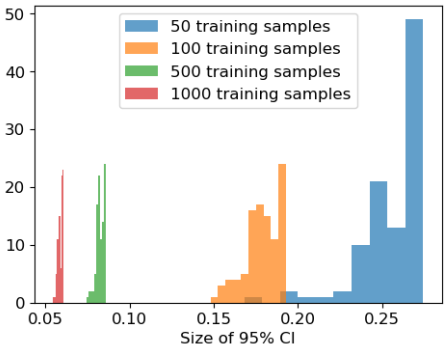
Experiment:

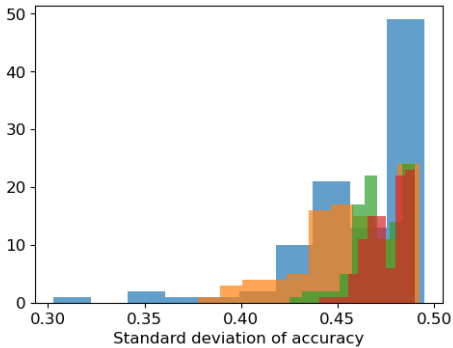
- Model: $g(x) = \frac{1}{1+\exp(-x^\top \beta)}, \quad f(x) = 1_{g(x)>0.5} \quad \beta = (c, c, c, c, c, 0, \cdots, 0)$
- number of features = 10 $\times$ number of samples; 200 repeats
- For each repeat: draw X_{tr}, generate y_{tr}, 10-fold cross-validate an l1-regularized logistic regression (with fixed regularization parameter), compute 95% confidence interval

CI width as training size increases

Experiment:

- Model: $g(x) = \frac{1}{1+\exp(-x^\top \beta)}, \quad f(x) = 1_{g(x)>0.5} \quad \beta = (c, c, c, c, c, 0, \cdots, 0)$
- number of features = 10 $\times$ number of samples; 200 repeats
- For each repeat: draw X_{tr}, generate y_{tr}, 10-fold cross-validate an l1-regularized logistic regression (with fixed regularization parameter), compute 95% confidence interval





What can we do?

- **Beware** of this effect, do not overinterpret results on small scale studies
- Endeavour to **increase sample sizes**

What can we do?

- **Beware** of this effect, do not overinterpret results on small scale studies
- Endeavour to **increase sample sizes**
- Braga-Neto, U.M. and Dougherty, E.R. Is cross-validation valid for small-sample microarray classification? Bioinformatics, 2004 (Braga-Neto and Dougherty 2004)
  - **cross-validation error estimation:** excessive **variance**
  - **bootstrap** has **lower variance** but **higher bias**
  - small improvement seen when **averaging multiple repeats** of cross-validation

Issue number 2: what does cross-validation estimate?

Bates, S., Hastie, T., & Tibshirani, R. Cross-validation: what does it estimate and how well does it do it? Journal of the American Statistical Association, 2023 (Bates, Hastie, and Tibshirani 2023)

Generalization error vs. expected test error

- Cross-validation is often used to estimate the **prediction error of a model**
  - Assumption: $Y = f(X) + \epsilon$
  - **Generalization error:** for a fixed train set $(X_{\text{tr}}, y_{\text{tr}})$, from which we learn model $\hat{f}$

$$\text{Err}_{(X_{\text{tr}}, y_{\text{tr}})} = \mathbb{E}\left[L(Y, \hat{f}(X))|(X_{\text{tr}}, y_{\text{tr}})\right]$$

Generalization error vs. expected test error

- Cross-validation is often used to estimate the **prediction error of a model**
  - Assumption: $Y = f(X) + \epsilon$
  - **Generalization error:** for a fixed train set $(X_{\text{tr}}, y_{\text{tr}})$, from which we learn model $\hat{f}$

$$\text{Err}_{(X_{\text{tr}}, y_{\text{tr}})} = \mathbb{E}\left[L(Y, \hat{f}(X))|(X_{\text{tr}}, y_{\text{tr}})\right]$$

- But what it actually estimates is the **average prediction error across training sets**
  - **Expected test error:**

$$\text{Err} = \mathbb{E}\left[\text{Err}_{(X_{\text{tr}}, y_{\text{tr}})}\right]$$

- This has been known at least since (Zhang 1995) and appears explicitely in The Elements of Statistical Learning

Experiment with the cross-validation estimator

- **Model:**
  - $X \sim \mathcal{N}(0, \text{diag}(1))$ of dimension p
  - $g(X) = \frac{1}{1+\exp(-X^\top \beta)}, \quad \beta = (c, c, c, c, c, 0, \cdots, 0)$
  - $U \sim \mathcal{U}([0,1]), \quad f(x) = 1_{g(x)>u}$
  - c calibrated so that the error of the model using $u = 0.5$ is about 15%

Experiment with the cross-validation estimator

- **Model:**
  - $X \sim \mathcal{N}(0, \text{diag}(1))$ of dimension p
  - $g(X) = \frac{1}{1+\exp(-X^\top \beta)}, \quad \beta = (c, c, c, c, c, 0, \cdots, 0)$
  - $U \sim \mathcal{U}([0,1]), \quad f(x) = 1_{g(x)>u}$
  - c calibrated so that the error of the model using $u = 0.5$ is about 15%
- **Learning algorithm:** l1-regularized logistic regression (with fixed regularization parameter)
- n=100 samples, p=1 000 features, 100 simulations

Experiment with the cross-validation estimator

- **Model:**
  - $X \sim \mathcal{N}(0, \text{diag}(1))$ of dimension p
  - $g(X) = \frac{1}{1+\exp(-X^\top \beta)}, \quad \beta = (c, c, c, c, c, 0, \cdots, 0)$
  - $U \sim \mathcal{U}([0,1]), \quad f(x) = 1_{g(x)>u}$
  - c calibrated so that the error of the model using $u = 0.5$ is about 15%
- **Learning algorithm:** l1-regularized logistic regression (with fixed regularization parameter)
- n=100 samples, p=1 000 features, 100 simulations
- **Experiment:**
  1. Compute the **expected test accuracy**
     For each simulation, generate a training set and compute
     2. The **generalization accuracy**
     3. Multiple estimations and CI95%, using the **cross-validation estimator**

Experiment with the cross-validation estimator

1. Compute the **expected test accuracy** Acc:
   - repeatedly
     - draw $X_{\text{tr-inner}}$ and $X_{\text{te-inner}}$,
     - generate $y_{\text{tr-inner}}$ and $y_{\text{te-inner}}$,
     - train model on $(X_{\text{tr-inner}}, y_{\text{tr-inner}})$,
     - compute accuracy acc-inner on $(X_{\text{te-inner}}, y_{\text{te-inner}})$
   - expected test accuracy = average all values of acc-inner

Experiment with the cross-validation estimator

– draw X_{tr}, generate y_{tr}

2. Compute the **generalization accuracy** $\text{Acc}_{(X_{\text{tr}}, y_{\text{tr}})}$:
   – train model on $(X_{\text{tr}}, y_{\text{tr}})$
   – draw X_{te}, y_{te},
   – $\text{Acc}_{(X_{\text{tr}}, y_{\text{tr}})}$ = model accuracy on $(X_{\text{te}}, y_{\text{te}})$

3. Multiple estimations and CI95%, using the **cross-validation estimator** $\widehat{\text{Acc}}^{\text{CV}}$

   For each estimation:

   – $\widehat{\text{Acc}}^{\text{CV}}$ = 5-fold cross-validated accuracy on $(X_{\text{tr}}, y_{\text{tr}})$
   – CI95 = $\widehat{\text{Acc}}^{\text{CV}} \pm 1.96 \frac{\hat{\sigma}}{\sqrt{n}}$

Side note

How to train an l1-regularized logistic regressions with a **given regularization parameter?**

- What we want:

$$\arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i) + \boldsymbol{\lambda} \, ||\beta||_1$$

Side note

How to train an l1-regularized logistic regressions with a **given regularization parameter?**

- What we want:

$$\arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i) + \boldsymbol{\lambda} \, ||\beta||_1$$

- This is what `glmnet` solves when you pass it $(x_i, y_i)_{i=1,\ldots,n}$ and λ

Side note

How to train an l1-regularized logistic regressions with a **given regularization parameter?**

- What we want:

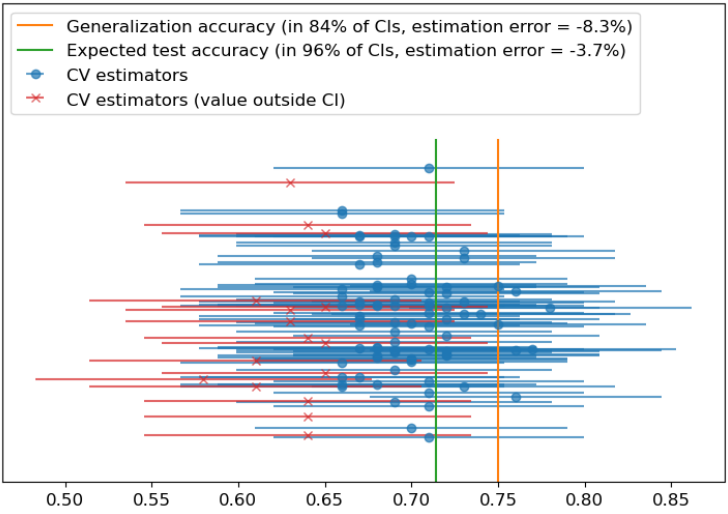
$$\arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i) + \boldsymbol{\lambda} ||\beta||_1$$

- This is what `glmnet` solves when you pass it $(x_i, y_i)_{i=1,\ldots,n}$ and λ
- But `scikit-learn` solves

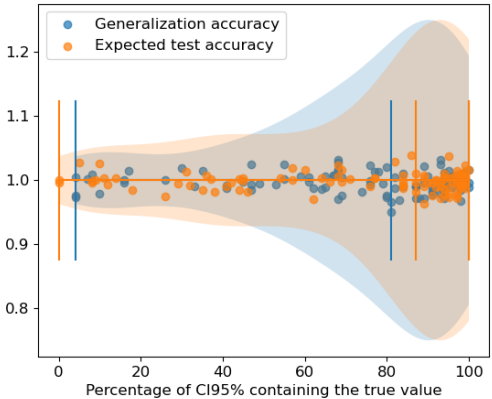
$$\arg\min_{\beta} \frac{1}{2} ||\beta||_1 + C \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i)$$

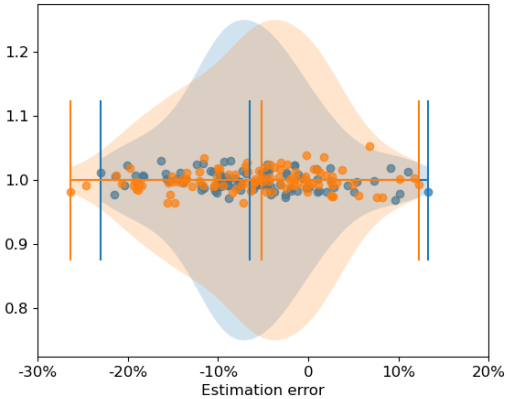
- $C = \frac{1}{2\boldsymbol{n}\lambda}$
- Working with **training sets of different sizes**, you have to make sure to use $C = C^*/n_{\text{train}}$

One simulation



100 simulations





Contributions of Bates et al. (2023)

1. Explain why the cross-validation estimator is a better estimate of expected test error than of generalization error

   (Also true for data splitting, bootstrap, Mallow's Cp.)

2. **Propose a new unbiased estimator of generalization error**

Nested CV estimator of generalization error: CIs

- There is **no unbiased estimator** of the variance of an estimator based on **a single instance of CV** (Bengio and Grandvalet 2004)

Nested CV estimator of generalization error: CIs

- There is **no unbiased estimator** of the variance of an estimator based on **a single instance of CV** (Bengio and Grandvalet 2004)

Nested CV estimator of generalization error: CIs

- There is **no unbiased estimator** of the variance of an estimator based on **a single instance of CV** (Bengio and Grandvalet 2004)
- Bates et al. use the following lemma:
  - Split the training data $(X_{\text{tr}}, y_{\text{tr}})$ into $(X_{\text{in}}, y_{\text{in}})$ and $(X_{\text{out}}, y_{\text{out}})$
    (typically "in" will be the union of $(K-1)$ folds of CV and "out" the remaining fold)
  - Train model $\hat{f}$ on $(X_{\text{in}}, y_{\text{in}})$ and obtain average predictions error e_{out} on $(X_{\text{out}}, y_{\text{out}})$
  - then the **mean squared error** of an estimator of prediction error on $(X_{\text{in}}, y_{\text{in}})$ is:

$$\mathbb{E}\left[\left(\widehat{\text{Err}}_{\text{in}} - \text{Err}_{\text{in}}\right)^2\right] = \mathbb{E}\left[\left(\widehat{\text{Err}}_{\text{in}} - e_{\text{out}}\right)^2\right] - \mathbb{E}\left[\left(e_{\text{out}} - \text{Err}_{\text{in}}\right)^2\right]$$

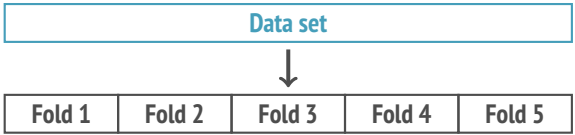
  - for confidence intervals: $\pm z_{1-\alpha/2}\sqrt{\widehat{\text{MSE}}}$

Nested CV estimator of generalization error: CIs

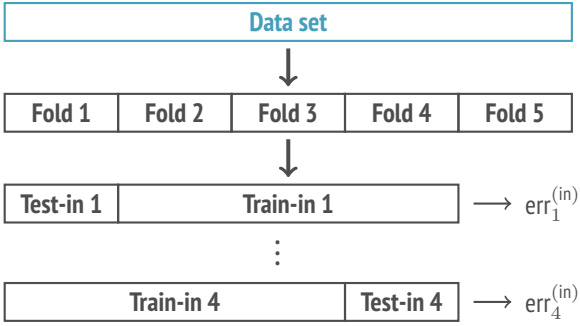
$$\mathbb{E}\left[\left(\widehat{\mathrm{Err}}_{\mathrm{in}} - \mathrm{Err}_{\mathrm{in}}\right)^2\right] = \mathbb{E}\left[\left(\widehat{\mathrm{Err}}_{\mathrm{in}} - e_{\mathrm{out}}\right)^2\right] - \mathbb{E}\left[\left(e_{\mathrm{out}} - \mathrm{Err}_{\mathrm{in}}\right)^2\right]$$

- The **first term** can be estimated with
  - a CV estimator on $(X_{\mathrm{in}}, y_{\mathrm{in}})$
  - the error on $(X_{\mathrm{out}}, y_{\mathrm{out}})$ of a model trained on $(X_{\mathrm{in}}, y_{\mathrm{in}})$
- The **second term** can be estimated as the mean empirical variance of the errors on $(X_{\mathrm{out}}, y_{\mathrm{out}})$

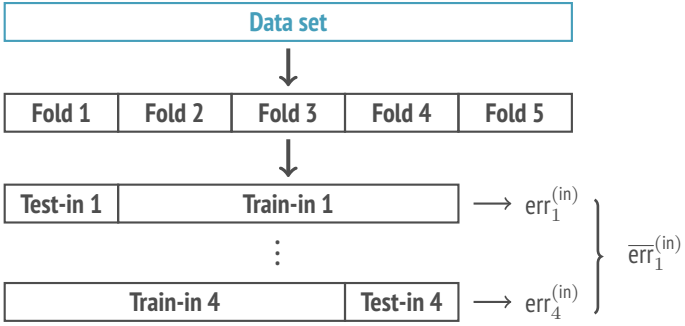
Nested CV estimator of generalization error: algorithm



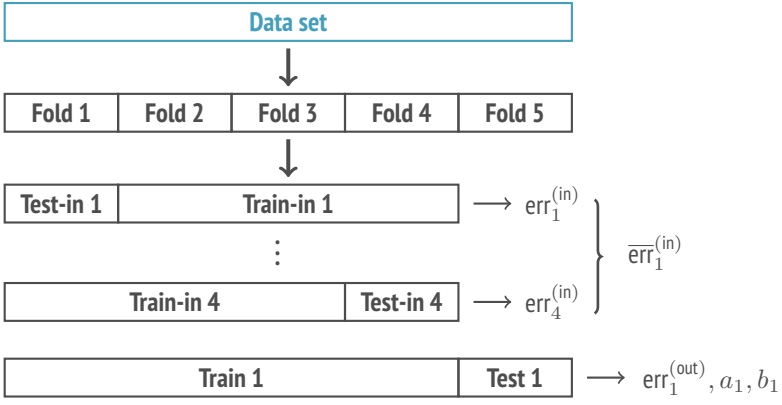
Nested CV estimator of generalization error: algorithm



Nested CV estimator of generalization error: algorithm

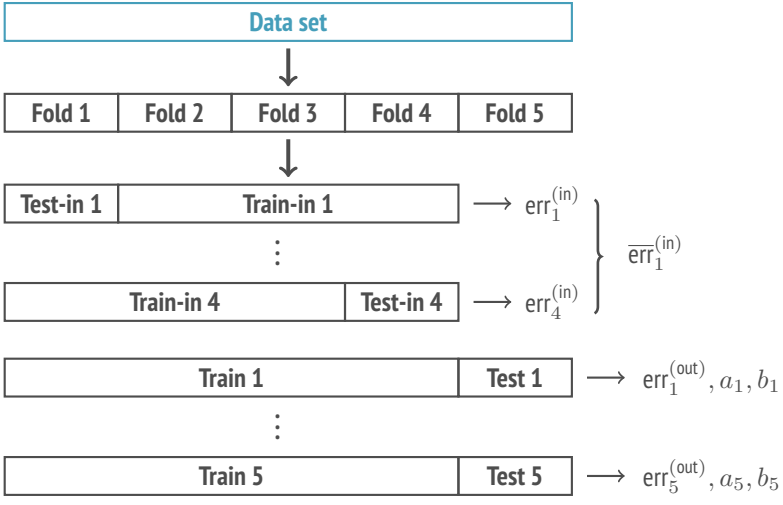


Nested CV estimator of generalization error: algorithm



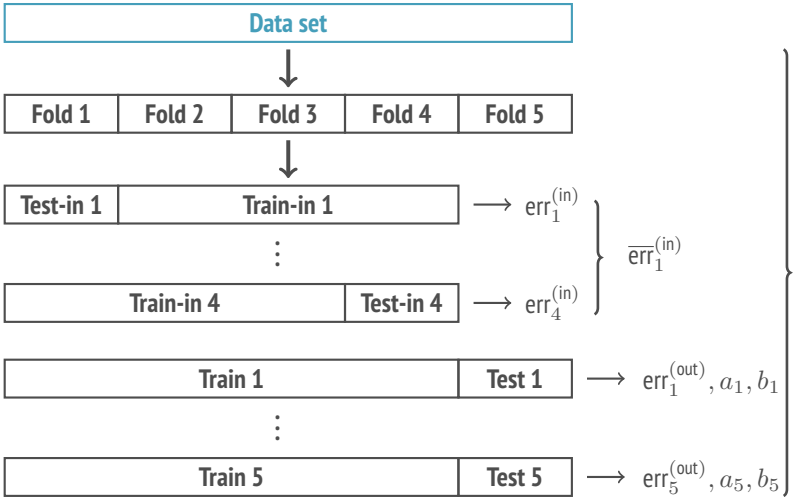
$$a_k = \left(\overline{\mathrm{err}}_k^{(\mathrm{in})} - \mathrm{mean}\left(\mathrm{err}_k^{(\mathrm{out})}\right)\right)^2 \qquad b_k = \mathrm{var}\left(\mathrm{err}_k^{(\mathrm{out})}\right) / |\mathrm{Test}_k|$$

Nested CV estimator of generalization error: algorithm



$$a_k = \left(\overline{\mathrm{err}}_k^{(\mathrm{in})} - \mathrm{mean}\left(\mathrm{err}_k^{(\mathrm{out})}\right)\right)^2 \qquad b_k = \mathrm{var}\left(\mathrm{err}_k^{(\mathrm{out})}\right) / |\mathrm{Test}_k|$$

Nested CV estimator of generalization error: algorithm



$$a_k = \left(\overline{\text{err}}_k^{(\text{in})} - \text{mean}\left(\text{err}_k^{(\text{out})}\right)\right)^2 \qquad b_k = \text{var}\left(\text{err}_k^{(\text{out})}\right) / |\text{Test}_k|$$

Nested CV estimator of generalization error: corrections

- The **MSE estimator** estimates the MSE of $(K-1)$-fold cross-validation on a sample of size $n(K-1)/K$

$\Rightarrow$ multiply it by $(K-1)/K$

Nested CV estimator of generalization error: corrections

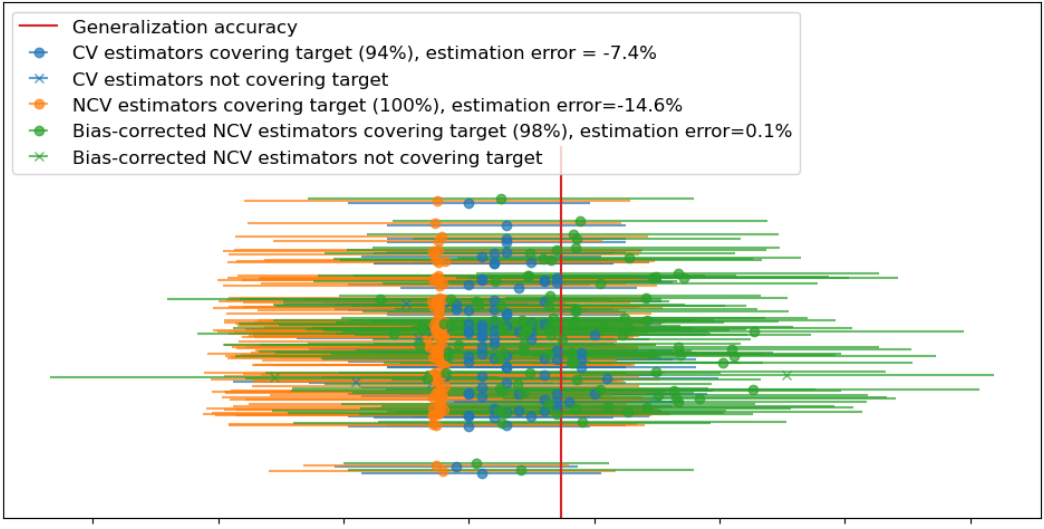
- The **MSE estimator** estimates the MSE of $(K-1)$-fold cross-validation on a sample of size $n(K-1)/K$

$\Rightarrow$ multiply it by $(K-1)/K$

- The nested CV estimator is **biased**
  - The nested CV algorithm uses models fit with $n(K-2)/K$ samples
  - The CV algorithm uses models fit with $n(K-1)/K$ samples

$$\widehat{\text{bias}} = \left(1 + \frac{K-2}{K}\right)^{1.5} \left(\widehat{\text{Err}}^{(\text{NCV})} - \widehat{\text{Err}}^{(\text{CV})}\right).$$

Experiment (one simulation)



Conclusions

- Cross-validation is more complex than you may have thought
- Estimating the **generalization error** of a model is hard
- Estimating the **expected test error** of a model is not much easier

Conclusions

- Cross-validation is more complex than you may have thought
- Estimating the **generalization error** of a model is hard
- Estimating the **expected test error** of a model is not much easier
- This matters more for **error estimation** than for **model selection**: cross-validation is asymptotically consistant for model selection (Wager 2020)

Conclusions

- Cross-validation is more complex than you may have thought
- Estimating the **generalization error** of a model is hard
- Estimating the **expected test error** of a model is not much easier
- This matters more for **error estimation** than for **model selection**: cross-validation is asymptotically consistant for model selection (Wager 2020)
- `scikit-learn`'s formulation of the regularized logistic regression does not scale the regularization parameter with training set size.

References I

Bates, Stephen, Trevor Hastie, and Robert Tibshirani (2023). "Cross-Validation: What Does It Estimate and How Well Does It Do It?" In: *Journal of the American Statistical Association*.

Bengio, Yoshua and Yves Grandvalet (2004). "No Unbiased Estimator of the Variance of K-Fold Cross-Validation". In: *The Journal of Machine Learning Research* 5, pp. 1089–1105.

Braga-Neto, Ulisses M and Edward R Dougherty (2004). "Is cross-validation valid for small-sample microarray classification". In: *Bioinformatics (Oxford, England)* 20.3, pp. 374–380.

Stone, M. (1974). "Cross-Validatory Choice and Assessment of Statistical Predictions". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 111–133.

Varoquaux, Gaël (2018). "Cross-validation failure: Small sample sizes lead to large error bars". In: *NeuroImage* 180.Pt A, pp. 68–77.

Wager, Stefan (2020). "Cross-Validation, Risk Estimation, and Model Selection: Comment on a Paper by Rosset and Tibshirani". In: *Journal of the American Statistical Association* 115.529, pp. 157–160.

Wherry, R. J. (1931). "A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation". In: *The Annals of Mathematical Statistics* 2.4, pp. 440–457. DOI: 10.1214/aoms/1177732951.

Zhang, Ping (1995). "Assessing Prediction Error in Non-Parametric Regression". In: *Scandinavian Journal of Statistics* 22.1, pp. 83–94.