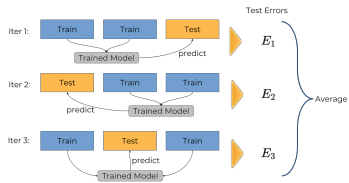


Introduction to Machine Learning

Evaluation Resampling 1

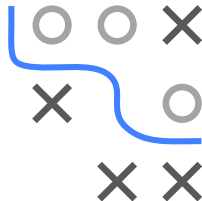
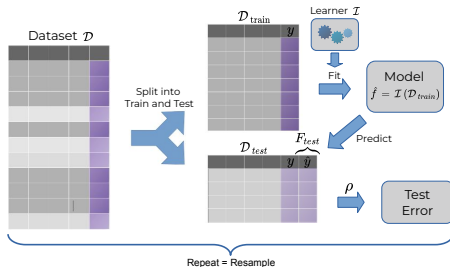


Learning goals

- Understand how resampling techniques extend the idea of simple train-test splits
- Understand the ideas of cross-validation, bootstrap and subsampling

RESAMPLING I

- **Goal:** estimate $\text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n, \rho_L) = \mathbf{E} [L(y, \mathcal{I}(\mathcal{D}_{\text{train}}, \boldsymbol{\lambda})(\mathbf{x}))]$.
- Holdout: Small trainset = high pessimistic bias; small testset = high var.
- Resampling: Repeatedly split in train and test, then average results.
- Allows to have large trainsets large (low pessimistic bias) since we use $\text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\text{train}}, \rho)$ as a proxy for $\text{GE}(\mathcal{I}, \boldsymbol{\lambda}, n, \rho)$
- And reduce var from small testsets via averaging over repetitions.



RESAMPLING STRATEGIES I

- Represent train and test sets by index vectors:

$$\mathbf{J}_{\text{train}} \in \{1, \dots, n\}^{n_{\text{train}}} \text{ and } \mathbf{J}_{\text{test}} \in \{1, \dots, n\}^{n_{\text{test}}}$$

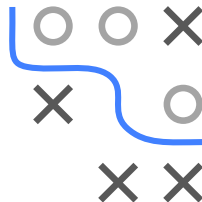
- Resampling strategy = collection of splits:

$$\mathcal{J} = ((\mathbf{J}_{\text{train},1}, \mathbf{J}_{\text{test},1}), \dots, (\mathbf{J}_{\text{train},B}, \mathbf{J}_{\text{test},B})).$$

- Resampling estimator:

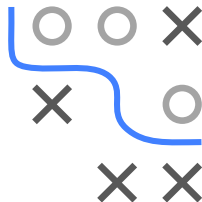
$$\begin{aligned} \widehat{\text{GE}}(\mathcal{I}, \mathcal{J}, \rho, \boldsymbol{\lambda}) = & \text{agr} \left(\rho \left(\mathbf{y}_{\mathbf{J}_{\text{test},1}}, \mathbf{F}_{\mathbf{J}_{\text{test},1}}, \mathcal{I}(\mathcal{D}_{\text{train},1}, \boldsymbol{\lambda}) \right), \right. \\ & \vdots \\ & \left. \rho \left(\mathbf{y}_{\mathbf{J}_{\text{test},B}}, \mathbf{F}_{\mathbf{J}_{\text{test},B}}, \mathcal{I}(\mathcal{D}_{\text{train},B}, \boldsymbol{\lambda}) \right) \right), \end{aligned}$$

- Aggregation agr is typically "mean" and $n_{\text{train}} \approx n_{\text{train},1} \approx \dots \approx n_{\text{train},B}$.

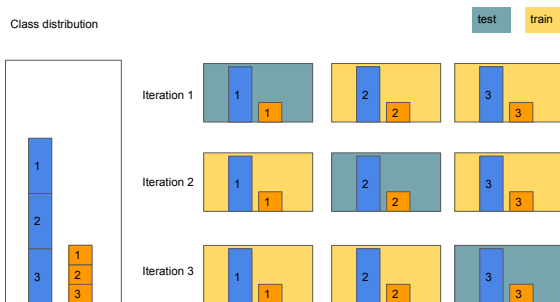


CROSS-VALIDATION - STRATIFICATION I

- Used when target classes are very imbalanced
- Then small classes can randomly get very small in samples
- Preserve distrib of target (or any feature) in each fold
- For classes: simply CV-split the class data, then join

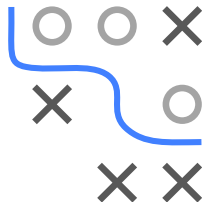


Example: stratified 3-fold cross-validation



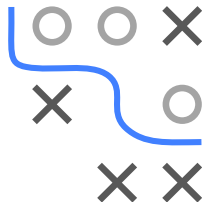
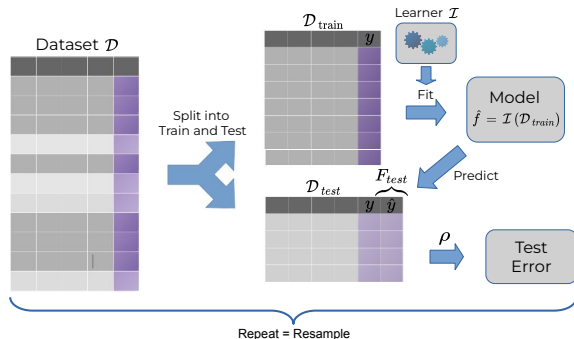
CROSS-VALIDATION I

- 5 or 10 folds are common.
- $k = n$ is known as "leave-one-out" CV (LOO-CV)
- Bias of \widehat{GE} : The more folds, the smaller. LOO nearly unbiased.
- LOO has high var, better many folds for small data but not LOO
- Repeated CV (avg over high-fold CVs) good for for small data.



SUBSAMPLING I

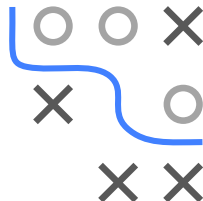
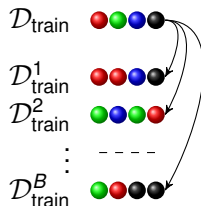
- Repeated hold-out with averaging, a.k.a. Monte Carlo CV.
- Typical choices for splitting: $\frac{4}{5}$ or $\frac{9}{10}$ for training.



- Smaller subsampling rate = larger pessimistic bias
- More reps = smaller var

BOOTSTRAP I

- Draw B trainsets of size n with replacement from orig \mathcal{D}
- Testsets = Out-Of-Bag points: $\mathcal{D}_{\text{test}}^b = \mathcal{D} \setminus \mathcal{D}_{\text{train}}^b$



- Similar analysis as for subsampling
- Trainsets contain about 2/3 unique points:

$$1 - \mathbb{P}((\mathbf{x}, y) \notin \mathcal{D}_{\text{train}}) = 1 - \left(1 - \frac{1}{n}\right)^n \xrightarrow{n \rightarrow \infty} 1 - \frac{1}{e} \approx 63.2\%$$
- Replicated train points can lead to problems and artifacts
- Extensions B632 and B632+ also use trainerr for better estimate when data very small

