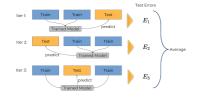
## **Introduction to Machine Learning**

# Evaluation Resampling 1



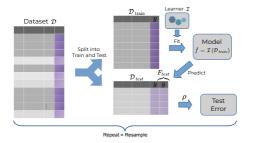


#### Learning goals

- 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  $GE(\mathcal{I}, \lambda, n, \rho_L) = E[L(y, \mathcal{I}(\mathcal{D}_{train}, \lambda)(\mathbf{x}))].$
- 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  $GE(\mathcal{I}, \lambda, n_{\text{train}}, \rho)$  as a proxy for  $GE(\mathcal{I}, \lambda, n, \rho)$ )
- And reduce var from small testsets via averaging over repetitions.





#### **RESAMPLING STRATEGIES I**

- Represent train and test sets by index vectors:  $J_{\text{train}} \in \{1, \dots, n\}^{n_{\text{train}}}$  and  $J_{\text{test}} \in \{1, \dots, n\}^{n_{\text{test}}}$
- Resampling strategy = collection of splits:

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

Resampling estimator:

$$\begin{split} \widehat{\mathrm{GE}}(\mathcal{I}, \mathcal{J}, \rho, \pmb{\lambda}) &= \mathrm{agr}\Big(\rho\Big(\mathbf{y}_{J_{\mathrm{test},1}}, \pmb{F}_{J_{\mathrm{test},1}, \mathcal{I}(\mathcal{D}_{\mathrm{train},1}, \pmb{\lambda})}\Big), \\ &\vdots \\ &\rho\Big(\mathbf{y}_{J_{\mathrm{test},B}}, \pmb{F}_{J_{\mathrm{test},B}, \mathcal{I}(\mathcal{D}_{\mathrm{train},B}, \pmb{\lambda})}\Big)\Big), \end{split}$$

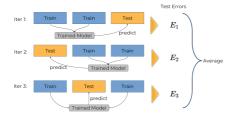
ullet Aggregation  $\operatorname{agr}$  is typically "mean" and  $n_{\operatorname{train}} pprox n_{\operatorname{train},1} pprox \cdots pprox n_{\operatorname{train},B}$ .



#### **CROSS-VALIDATION I**

- Split the data into *k* roughly equally-sized partitions.
- Each part is test set once, join k-1 parts for training.
- Obtain *k* test errors and average.
- Fraction (k-1)/k is used for training, so 90% for 10CV
- Each observation is tested exactly once.

#### Example: 3-fold CV

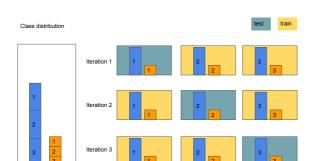




### **CROSS-VALIDATION - STRATIFICATION I**

- 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







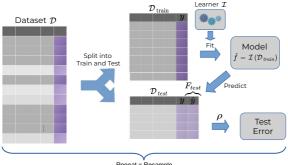
#### **CROSS-VALIDATION I**

- 5 or 10 folds are common.
- k = n is known as "leave-one-out" CV (LOO-CV)
- $\bullet\,$  Bias of 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.

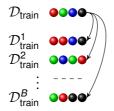


- Repeat = Resample
- Smaller subsampling rate = larger pessimistic bias
- More reps = smaller var



#### **BOOTSTRAP I**

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





- Similar analysis as for subsampling
- Trainsets contain about 2/3 unique points:  $1 \mathbb{P}((\mathbf{x}, \mathbf{y}) \notin \mathcal{D}_{\text{train}}) = 1 \left(1 \frac{1}{n}\right)^n \stackrel{n \to \infty}{\longrightarrow} 1 \frac{1}{e} \approx 63.2\%$
- Replicated train points can lead to problems and artifacts
- Extensions B632 and B632+ also use trainerr for better estimate when data very small

#### LEAVE-ONE-OBJECT-OUT I

- Used when we have multiple obs from same objects, e.g., persons or hospitals or base images
- Data not i.i.d. any more
- Data from same object should either be in train or testset
- ullet Otherwise we likely bias  $\widehat{GE}$
- CV on objects, or leave-one-object-out

