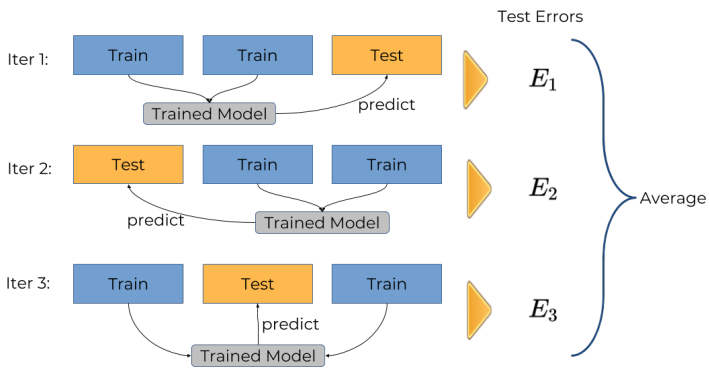
Introduction to Machine Learning

Evaluation: Resampling 2



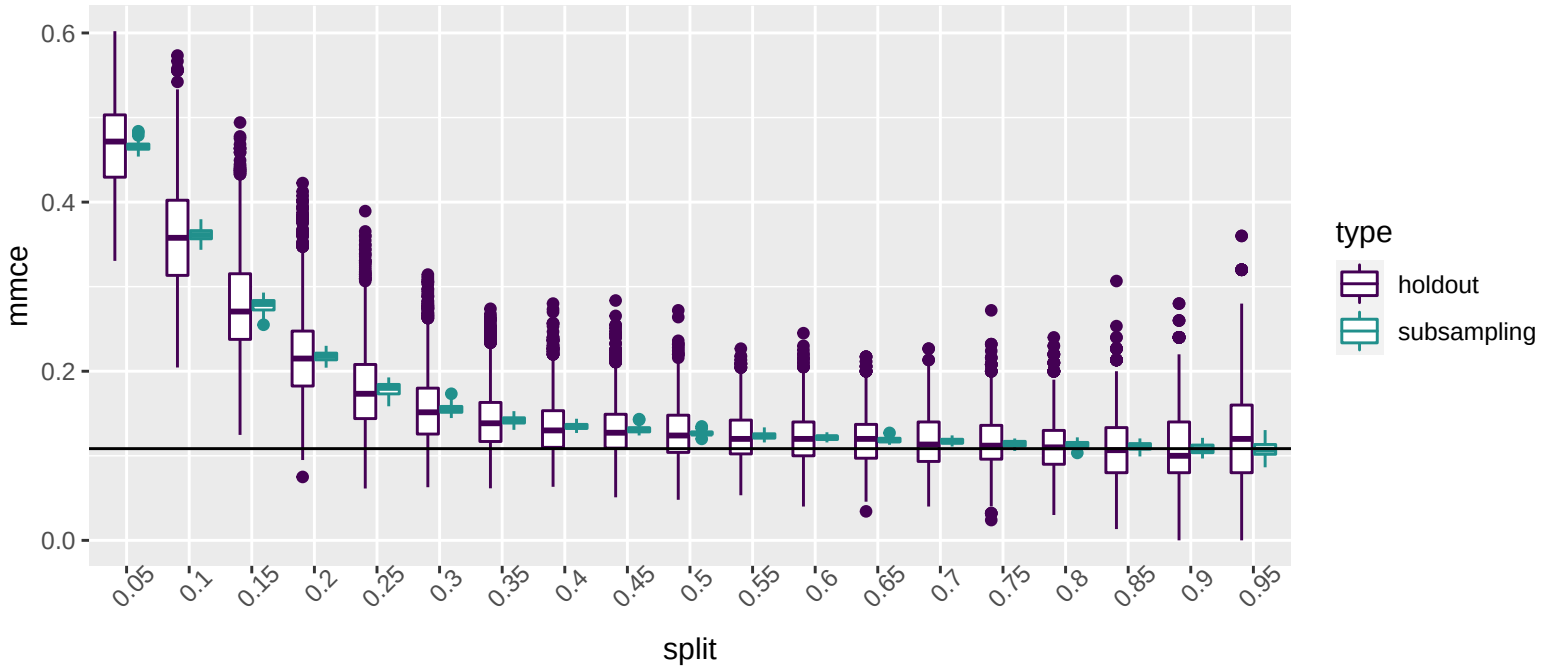
Learning goals

- Understand advantages of subsampling over single hold-out split
- Understand challenges when comparing learners based on CV results
- Understand what pessimistic bias means
- Be able to compare different resampling strategies

BIAS-VARIANCE ANALYSIS FOR SUBSAMPLING

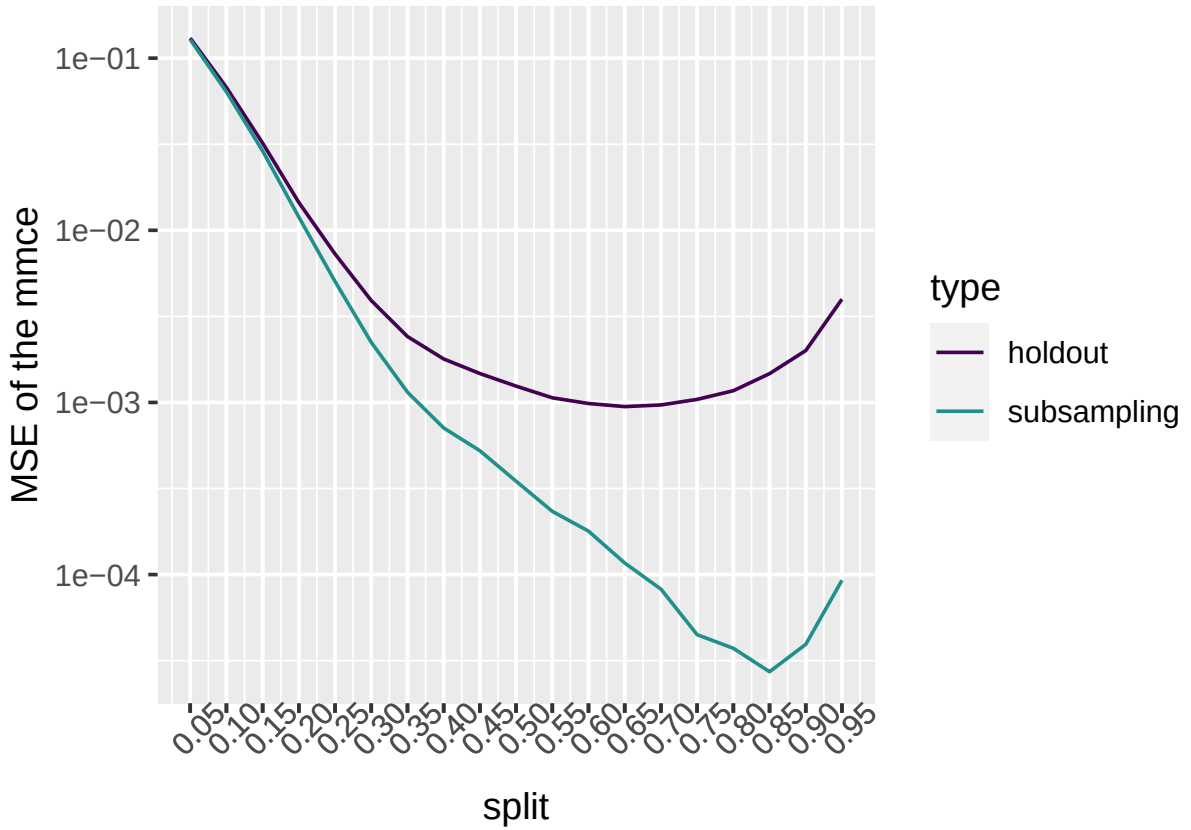
- Reconsider our hold-out experiment on the `spirals` data from the train-test unit (maybe re-read it again)
- Again, we use split rates $s \in \{0.05, 0.1, ..., 0.95\}$ for training with $|\mathcal{D}_{\text{train}}| = s \cdot 500$.
- But: now we compare 50 subsampling experiments with $50 \cdot 50$ hold-out experiments per split.
- Every subsampling experiment is the result of averaging 50 hold-out experiments, so each performance estimate is much more reliable (but also more expensive) than one computed by a single hold-out experiment.

BIAS-VARIANCE ANALYSIS FOR SUBSAMPLING



- Both experiments are compared to the "real" MMCE (black line).
- Subsampling has the same pessimistic bias for small split rates but much less variance overall.
- This allows to use much smaller test sets with good results.

BIAS-VARIANCE ANALYSIS FOR SUBSAMPLING



- The MSE is strictly better for subsampling compared to hold-out.
- The optimal split rate now is a higher $s \approx 0.8$.
- We see the variance picking up at the end because training sets increasingly overlap.

COMPARING LEARNERS BASED ON CV RESULTS

- Since the k training sets and their respective independent test sets stem from the same distribution we get k unbiased estimators of $\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\mathrm{train}}, \rho)$.
- However, in order to compare two different learners we also need to assess the uncertainty of our overall estimator.
- This becomes challenging since the k splits are not independent of each other:
- $\mathbb{V}[\widehat{\mathrm{GE}}]$ of CV is a linear combination of
  - the average variance we get from estimating from finite training sets,
  - the covariance arising from the dependence of test errors the learners made since they were trained on overlapping training sets,
  - the covariance due to the dependence of training sets and that test sets appear also in training sets. [Bengio, 2004]

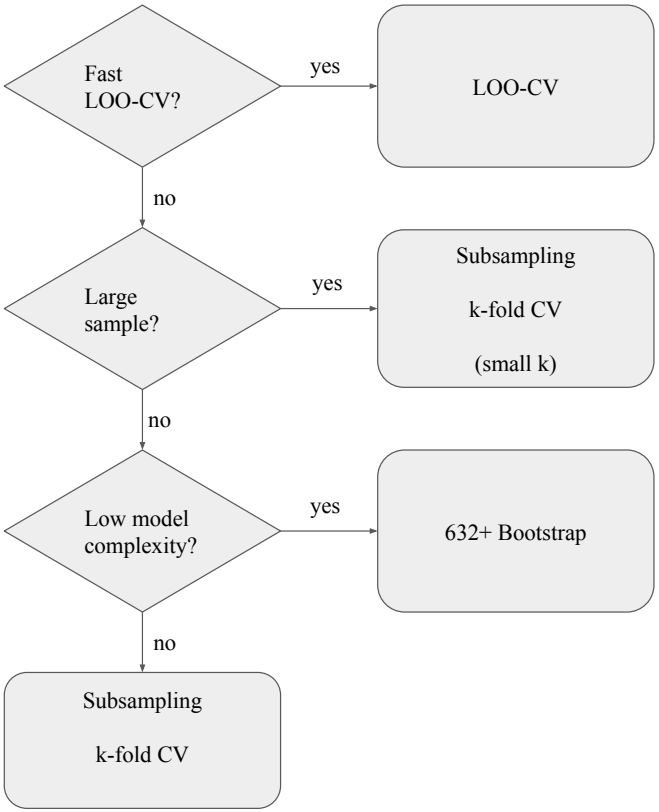
$\Rightarrow$ Taking the empirical variance of the k $\widehat{\mathrm{GE}}$s yields a biased estimator of $\mathbb{V}[\widehat{\mathrm{GE}}]$.

$\Rightarrow$ These dependences should be taken into account when testing if a learner is significantly better than another.

RESAMPLING DISCUSSION

- In ML we fit, at the end, a model on all our given data.
- **Problem:** We need to know how well this model will perform in the future, but no data is left to reliably quantify this.
  $\Rightarrow$ Approximate using hold-out / CV / bootstrap / subsampling estimate
- **But:** pessimistic bias because we don't use all data points.
- The final model is (usually) computed on all data points.
- Strictly speaking, resampling only produces one number, the performance estimator. It does NOT produce models, parameters, etc. These are intermediate results and discarded.
- The model and parameters are only obtained when we finally fit the learner on the complete data.

RESAMPLING DISCUSSION



- 5-CV or 10-CV have become standard.
- Do not use hold-out, CV with few iterations, or subsampling with a low subsampling rate for small samples, since this can cause the estimator to be extremely biased, with large variance.
- For small-data situations with less than 500 or 200 observations, use LOO or, probably better, repeated CV.
- For some models, computationally fast calculations or approximations for LOO exist.
- A data set $\mathcal{D}$ with $|\mathcal{D}| = 100.000$ can have small-sample properties if one class has few observations
- Research indicates that subsampling has better properties than bootstrapping. The repeated observations can cause problems in training, especially in nested setups where the “training” set is split up again.