Chapter 5 Overfitting and its Avoidance

10/29/19 9:02 PM

Generalization:

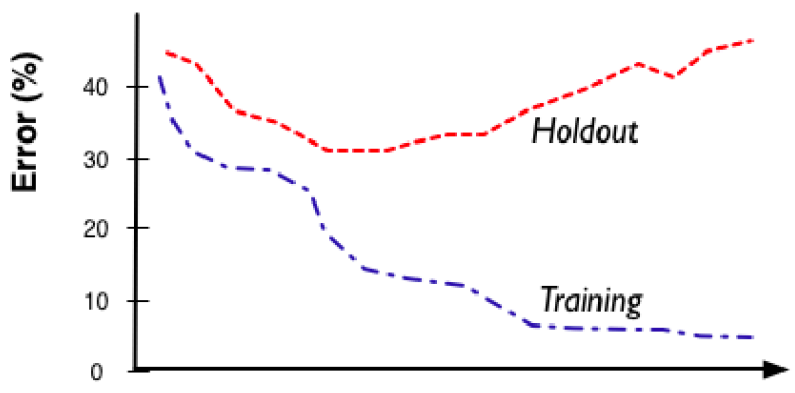
A character of the model whereby the model applies to data beyond training data

Overfitting:

A model is tailored to training data at the expense of generalization

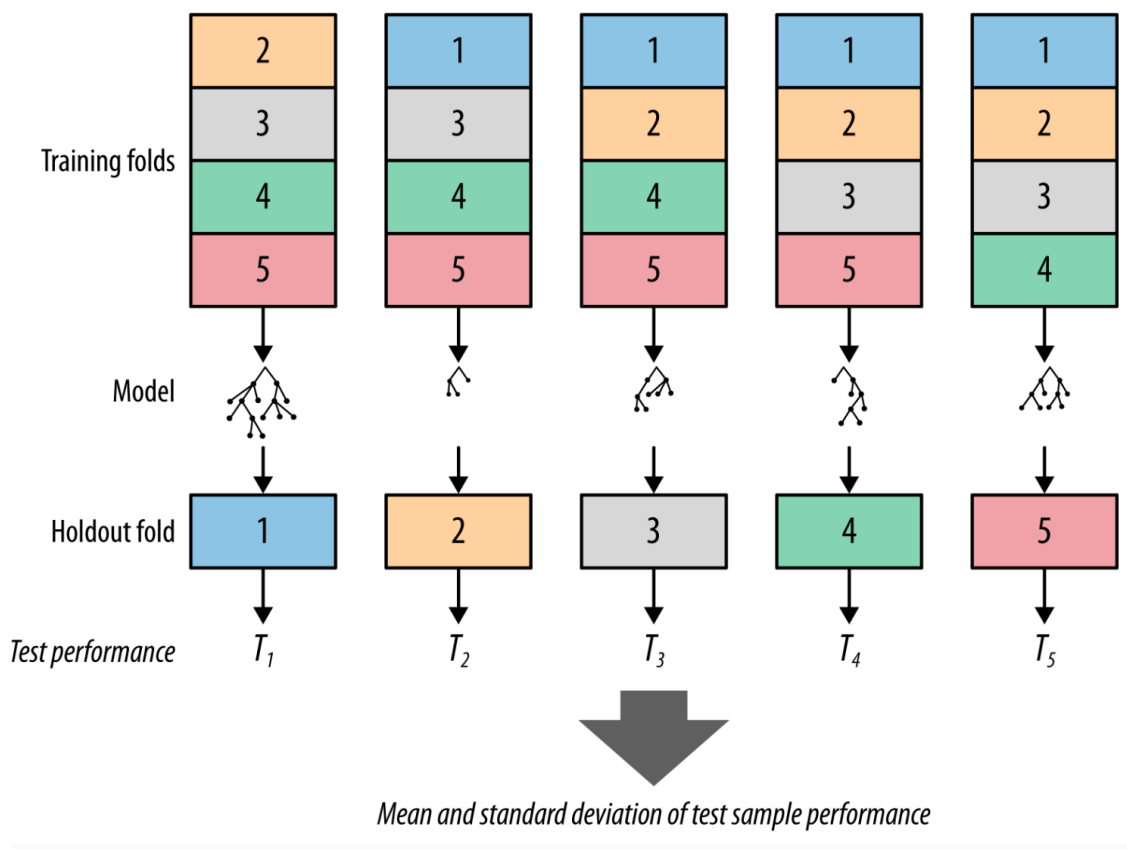
Fitting Graph (to examine overfitting):

- Fitting graph: model complexity and error for predicting
- Holdout data: test data



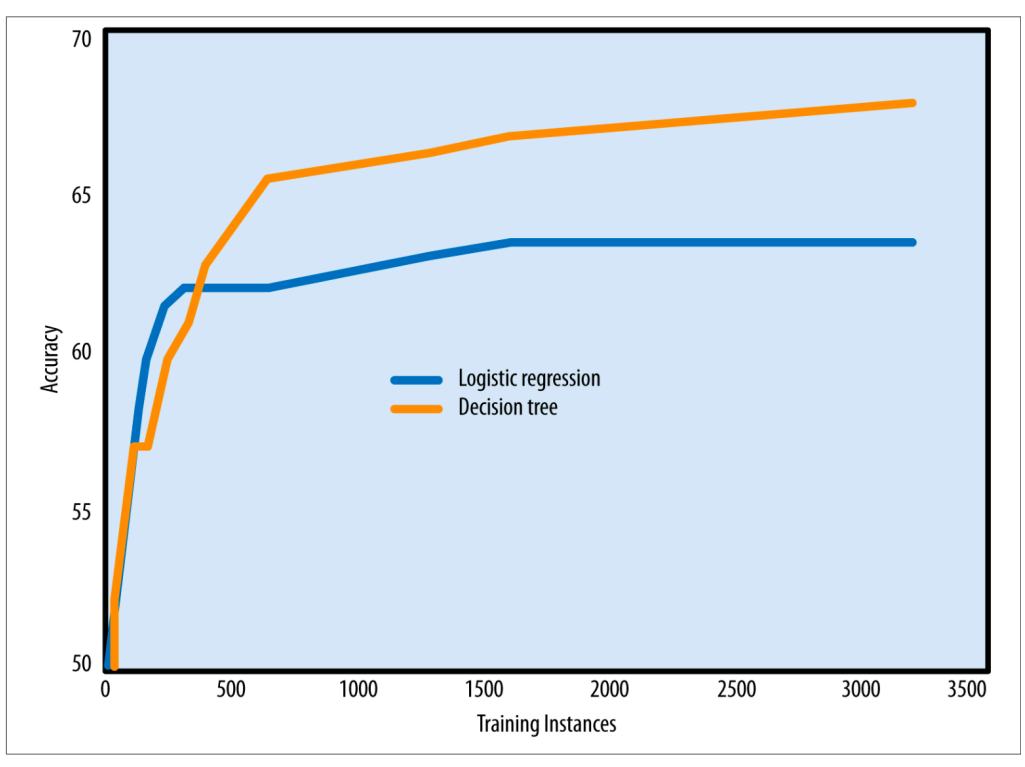
Cross-validation

- Generate a set of performance, allow some statistics (mean, variance) on performance
- Makes better use of limited sample size



Learning Curves

- Generalization performance against training sample size
  - X axis: amount of training data used
  - Y axis: performance on **test** data
- Learning curve also would suggest whether investing in more data would be worthy. When sample size is very small, increasing training data size would significant improve model performance. However, if the learning curve is already almost flat, it's not valuable to have larger sample size at cost.



Avoid overfitting in tree induction

- 1st method: Control smallest size of each node, only split nodes with many instances
  - Statistically, can do hypothesis test on each leaf to observe whether the reduced IG is due to chance (if by chance only, keep it)
- 2nd method: get the whole tree, then prune, based on whether the pruning will reduce accuracy
- In practice, can try a list of min size of leaf, compute accuracy, and choose the "best"

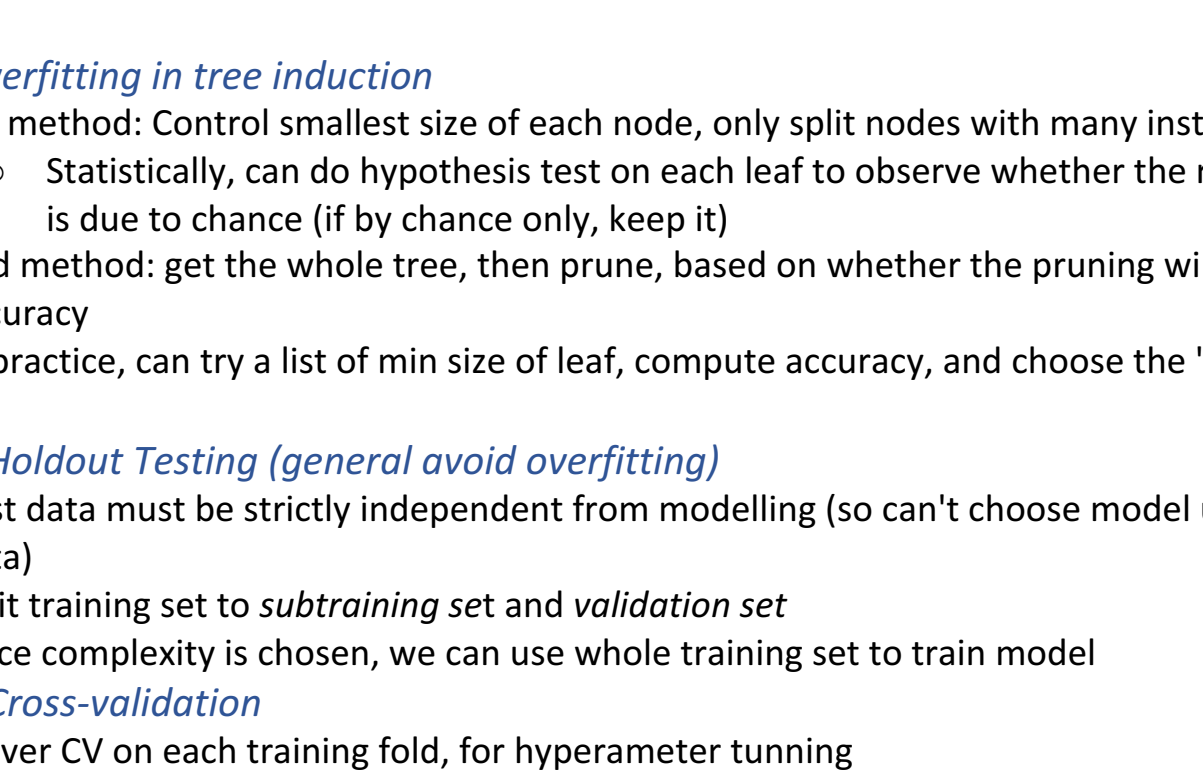
Nested Holdout Testing (general avoid overfitting)

- Test data must be strictly independent from modelling (so can't choose model using test data)
- Split training set to *subtraining se*t and *validation set*
- Once complexity is chosen, we can use whole training set to train model

Nested Cross-validation

- Innver CV on each training fold, for hyperameter tunning
- Performace evaluation on each test fold, using hyperameter from inner CV
- Outer CV across all folds for steps above
- Final performance score is average of each outer CV.

-



Avoid Overfitting for Parameter Optimization

- Model Regularization
  - Complexity control via regularization works by adding to objective function a penalty for complexity
  - Ridge Regression: L-2 Norm as penalty; Lasso Regression: L-1 norm as penalty