Introduction to Machine Learning Applications

Spring 2023

Model fitting

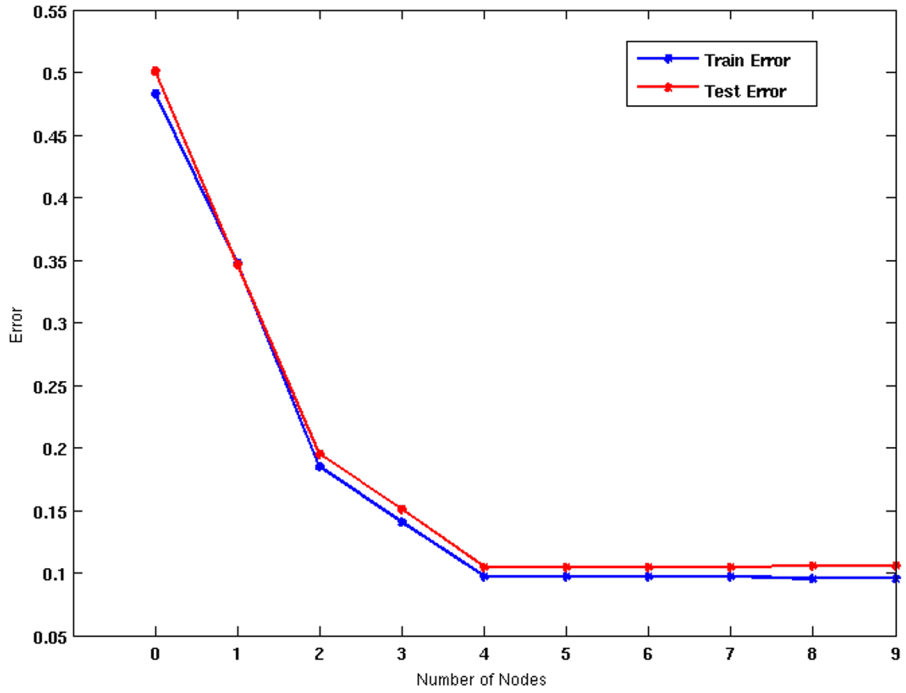
Minor Gordon

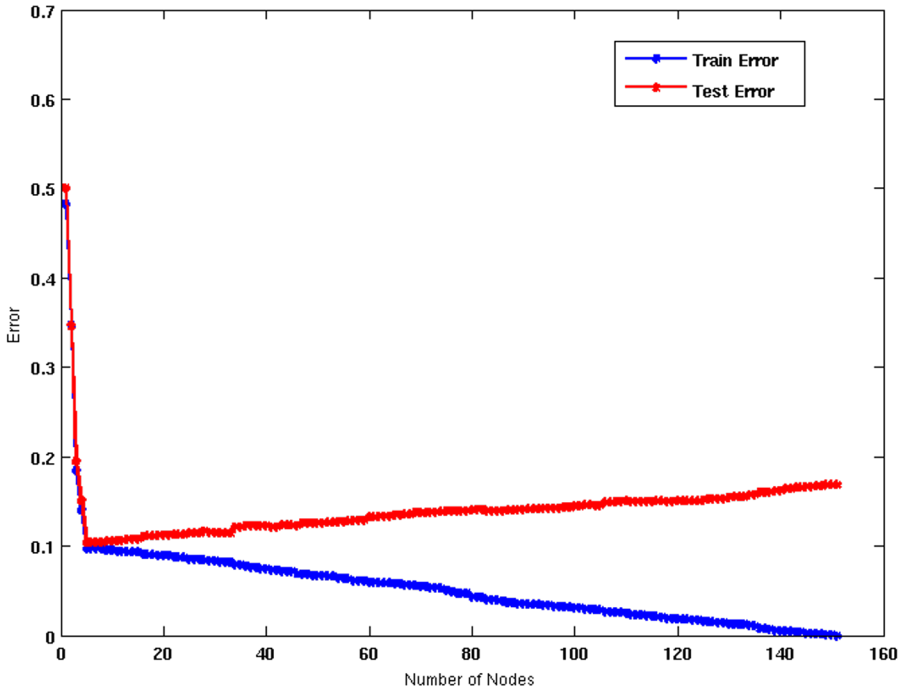
gordom6@rpi.edu

Rensselaer

Model training should result in a model that can adequately generalize to new data.

Model Overfitting & Underfitting



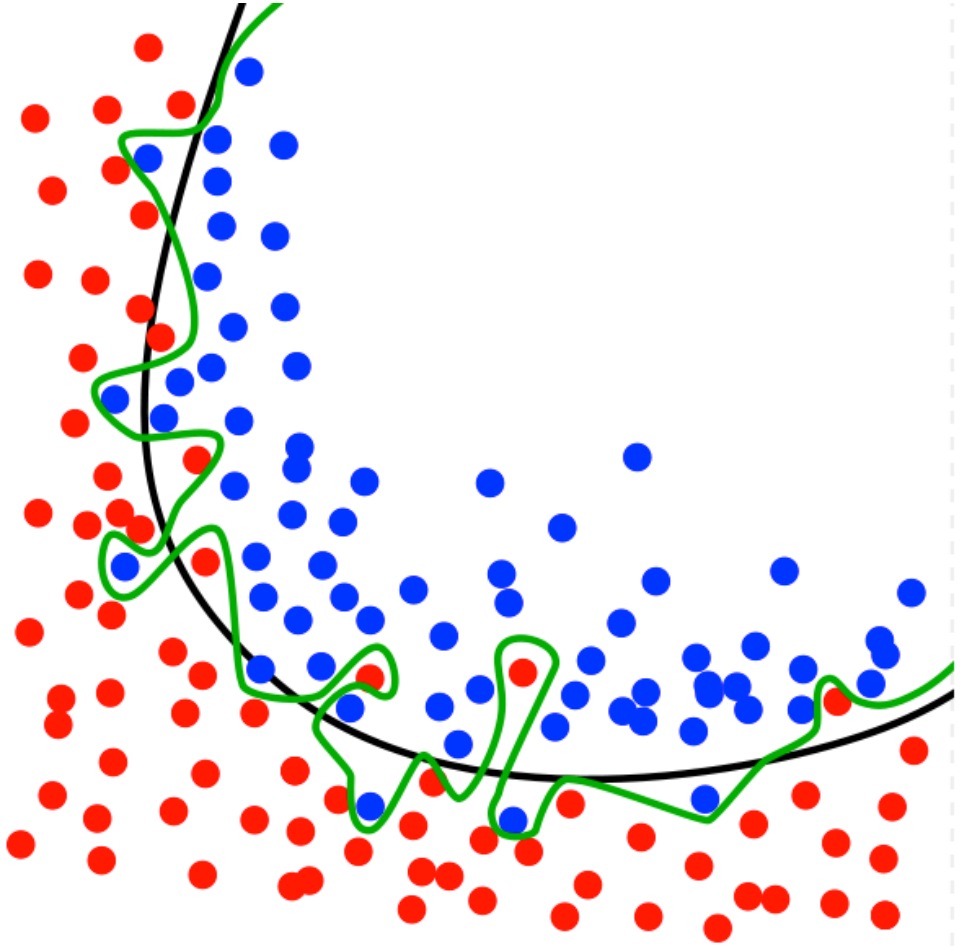


Underfitting: when model is too simple, both training and test errors are large

Overfitting: when model is too complex, training error is small but test error is large

Overfitting Training Set

- This model with have a high level of accuracy but will not generalize to other data

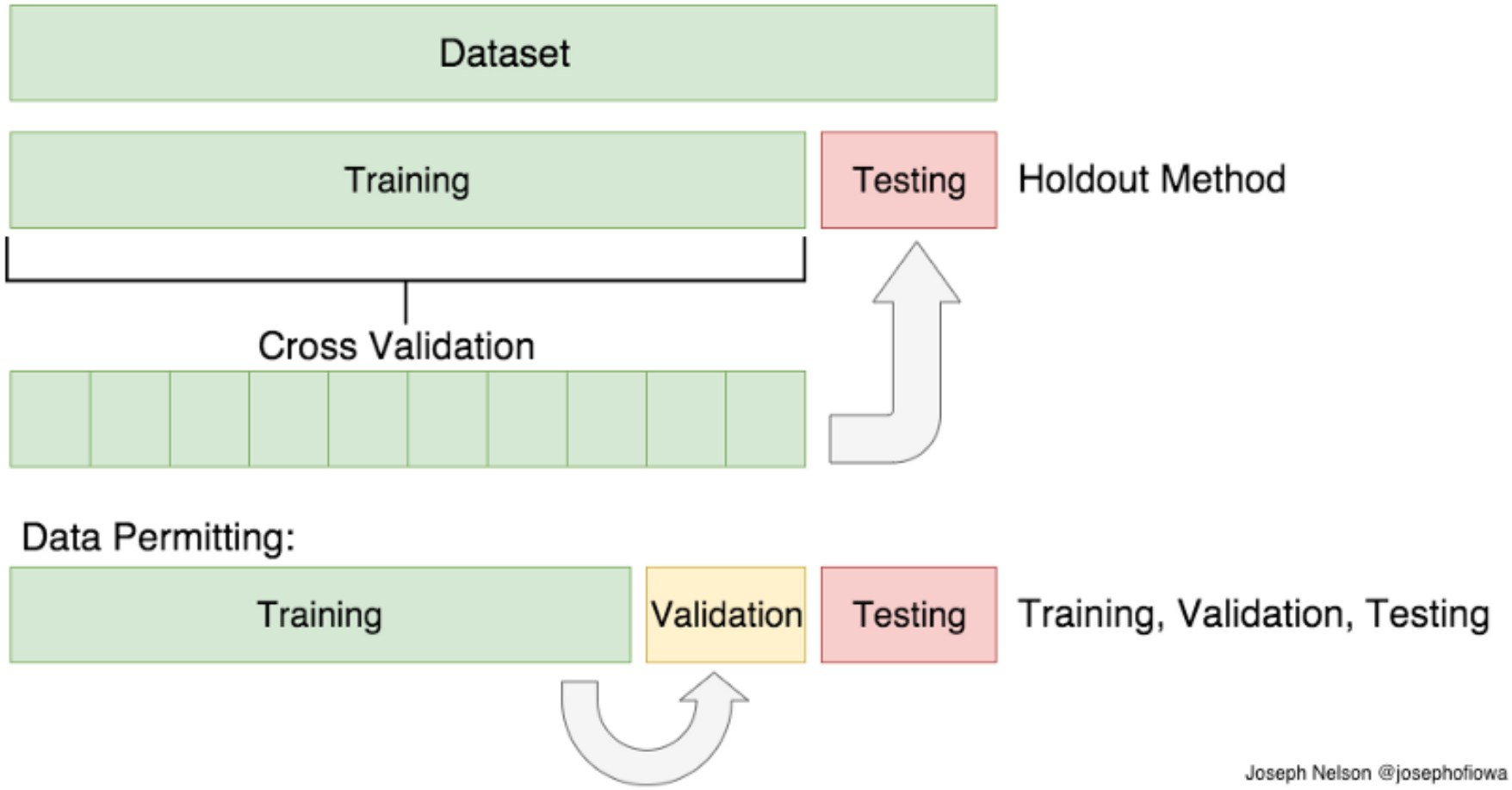


By Chabacano (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 4.0-3.0-2.5-2.0-1.0 (https://creativecommons.org/licenses/by-sa/4.0-3.0-2.5-

Overfitting is likely occurring when there is a large difference in the performance of the model between the training and test set.

How do we prepare data to train a model and prevent over fitting?

Separation Into Train/Test



https://medium.com/towards-data-science/train-test-split-and-cross-validation-in-python-80b61beca4b6

Cross Validation

- Used to prevent overfitting of model and/or improving fit
- Many Different Types
  - Holdout Method
  - K-fold Cross Validation
  - Repeated random sub-sampling validation (small n)
  - Leave-one-out cross-validation (small n)

2 Fold Cross Validation/Holdout Method

- For each fold, we randomly assign data points to two sets d_0 and d_1, so that both sets are equal size (this is usually implemented by shuffling the data array and then splitting it in two). We then train on d_0 and test on d_1, followed by training on d_1 and testing on d_0.
- This has the advantage that our training and test sets are both large, and each data point is used for both training and validation on each fold.

Example: 2 Fold Cross Validation/Holdout Method

Titanic Dataset (Goal is to predict survival)

- Split sample randomly [DF_1, DF_2]
- Using DF_1 train survival model use the model to predict survival in the DF_1 sample
- Using DF_2 train survival model use the model to predict survival in the DF_1 sample

k-fold Cross-validation

- Resampling procedure to evaluate machine learning models on a given data sample.
- The parameter *k* refers to the number of groups that a given data sample is to be split into.
- If k=10, it is 10-fold cross-validation where the sample data is divided into 10 groups.

k-fold Cross-validation

> Shuffle the dataset (better)

> Split the dataset into k disjoint groups

> For each unique group:

  > Take the group as a hold out or test (validation) data set

  > Take the remaining groups as a training data set

  > Fit a model on the training set and evaluate it on the test set

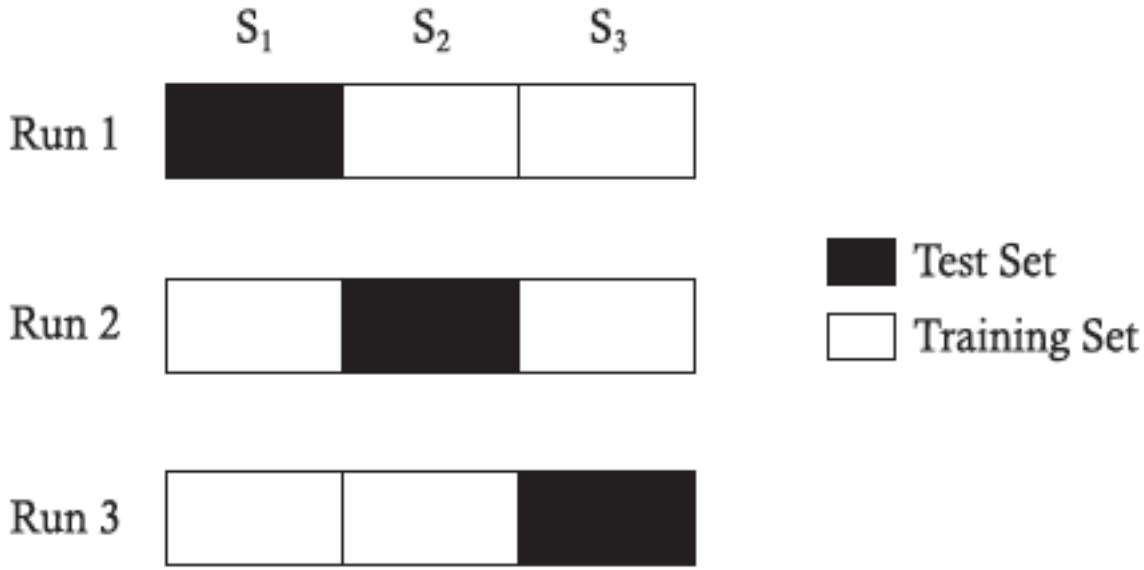
  > Record the evaluation score

> Find the mean of all the sample of model evaluation scores

k-fold Cross-validation

[1, 2, 3, 4, 5, 6]

Fold1: [5, 3]
Fold2: [1, 6]
Fold3: [2,4]

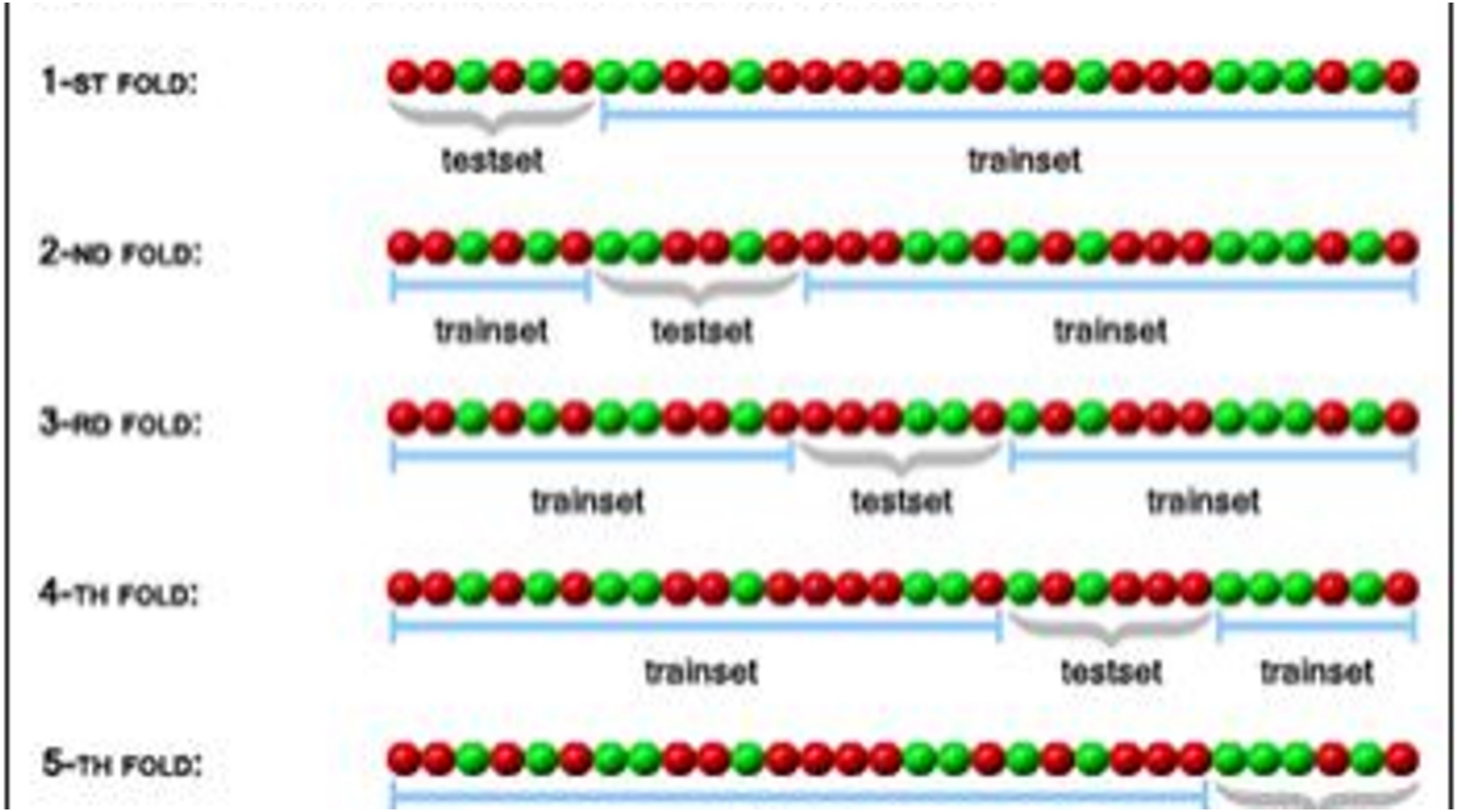


Model1: Trained on Fold2 + Fold3, Tested on Fold1
Model2: Trained on Fold1 + Fold3, Tested on Fold2
Model3: Trained on Fold1 + Fold2, Tested on Fold3

Example

- Given a set of data points – {a, b, c, d, e, f, g, h}
  - Perform 4-fold cross validation
  - Explain it in your own terms – what are the folds and how do you use them?

5 Fold Cross Validation



Train-test split notebook

Hyperparameter Tuning

“**Hyperparameters** are parameters whose values are set prior to the commencement of the learning process. By contrast, the values of other parameters are derived via training. Hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm.” - Wikipedia

Hyperparameter Optimization

- Some types of models have different parameters, which impact how the model behaves.
- For example, for a tree, how many layers of the tree do we want? Do we want lots of branches or just the key branches?
- Tuning of these parameters, using the resulting performance, is call hyperparameter optimization

Models and Hyperparameters

- Can be optimized through a *grid search model* where hyperparameters are adjusted systematically
  - Advantages: Exhaustive
  - Disadvantages: Time consuming

- Can lead to overfitting. If you try a bunch of models on your test set, your model may be “tuned” specifically to the characteristics.

KNN notebook