WELCOME TO

TECH I.S.

Cross Validation

Cross Validation

Almost all the algorithm that we have seen have one or more parameters that can be chosen:

- **The number of neighbors k in a kNN Classification Rule.**
- **The number of features n to preserve in a Subset Selection problem.**

Two issues arise at this point:

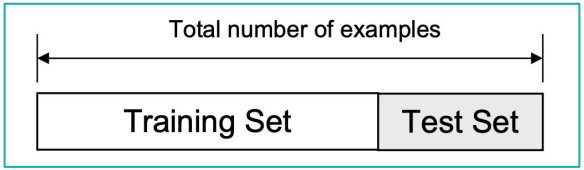
- **Model Selection: How do we select the "optimal" parameter(s) for a given classification problem?**
- **Validation: Once we have chosen a model, how do we estimate its true error rate?**

If we had access to an unlimited data, these questions would have a straightforward answer, however in real applications only a finite set of examples is available as Data collection is a very expensive process.

The Holdout Method

Split dataset into two groups:

- **Training set: used to train the classifier**
- **Test set: used to estimate the error rate of the trained classifier**



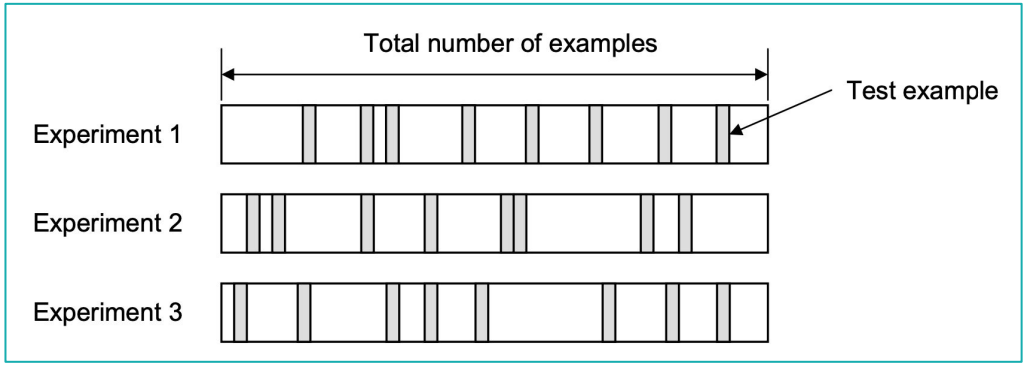
The holdout method has two basic drawbacks:

- **In problems where we have a sparse(small) dataset we may not be able to afford the "luxury" of setting aside a portion of the dataset for testing.**
- **Since the splitting is random, If we happen to get an "unfortunate" split, the the model will not fit properly.**

Random Subsampling

Random Subsampling performs K data splits of the entire dataset:

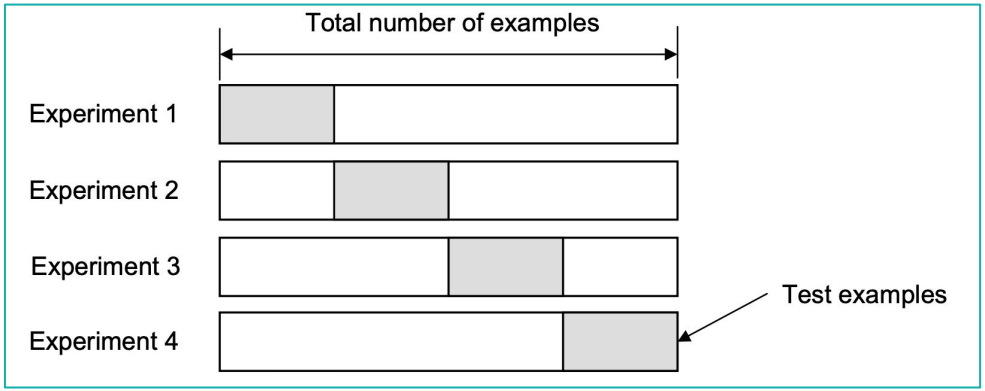
- Each data split randomly selects a (fixed) number of examples without replacement.
- For each data split we retrain the classifier from scratch with the training examples and then estimate E_i with the test examples which is averaged out later.



K-Fold Cross-validation

Creates a K-fold partition of the the dataset:

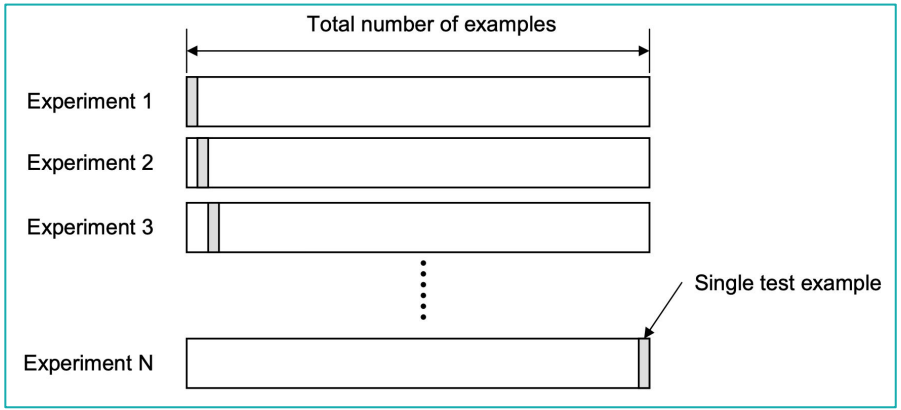
- For each of K experiments, use K-1 folds for training and a different fold for testing
- The advantage of K-Fold Cross validation is that all the examples in the dataset are eventually used for both training and testing.
- This procedure is illustrated in the following figure for K=4



Leave-one-out Cross Validation

Leave-one-out is the degenerate case of K-Fold Cross Validation, where K is chosen as the total number of examples.

- For a dataset with N examples, perform N experiments.
- For each experiment use N-1 examples for training and the remaining example for testing.



How many folds (k) are needed?

With a large number of folds:

- The bias of the true error rate estimator will be small (the estimator will be very accurate).
- The variance of the true error rate estimator will be large.
- The computational time will be very large as well (many experiments).

With a small number of folds

- The number of experiments and, therefore, computation time are reduced.
- The variance of the estimator will be small.
- The bias of the estimator will be large (conservative or smaller than the true error rate)

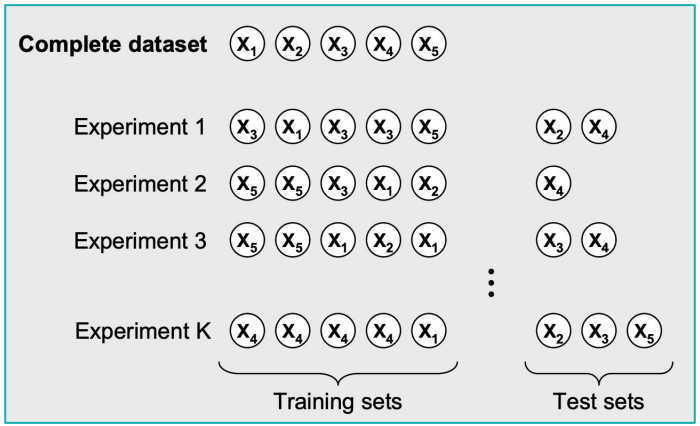
In practice, the choice of the number of folds depends on the size of the dataset:

- For large datasets, even 3-Fold Cross Validation will be quite accurate
- For very sparse datasets, we may have to use higher number of k.

The Bootstrapping

The bootstrap is a resampling technique with replacement:

- From a dataset with N examples
  - Randomly select (with replacement) N examples and use this set for training
  - The remaining examples that were not selected for training are used for testing
- Repeat this process for a specified number of folds (K)



Advantages of Bootstrapping

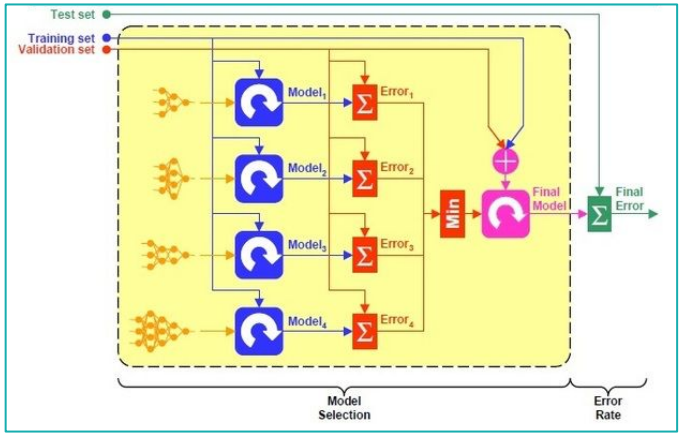
Advantages of Bootstrapping as compared to basic cross-validation:

- Compared to basic cross-validation, the bootstrap increases the variance that can occur in each fold.
- An additional benefit of the bootstrap is its ability to obtain accurate measures of BOTH the bias and variance of the true error estimate

Three-way Data Splits

If Cross Validation and Performance are to be computed simultaneously, the data needs to be divided into three disjoint sets:

- Training set: a set of examples used for learning: to fit the parameters of the classifier.
- Validation set: a set of examples used to tune the parameters of a classifier.
- Test set: a set of examples used only to assess the performance of a fully-trained classifier.



Why separate test and validation sets?

The error estimate of the final model on validation data will be smaller than the true error rate.

Much obliged.

TECH I.S.