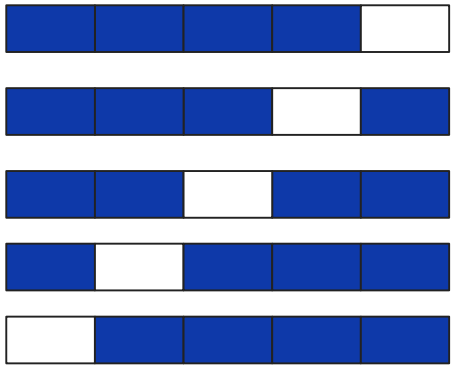
Cross validation - concept and procedure

Need for cross-validation

- We wish to know how well a Machine Learning model is likely to perform in production
- To estimate the model production score, we divide the data into 3 parts.
- Usually the available data is not sufficient to split into training, test and validation sets and expect them to represent the universe.
- As a result, the performance on training data may not be a good estimate of the performance in the universe.
- Using cross validation techniques can help us achieve a generalised model performance.
- This technique is also very useful in the absence of large data sets.

Cross-validation

- Cross-validation divides data into k folds
- k-1 folds are used for training and the kth fold is used for testing purposes
- The process is repeated k times
- k different scores will be obtained
- The final model performance is calculated by taking the average of k scores
- We can also calculate the standard deviation of those scores and using that we can say that with 95% confidence our model's performance will belong to



(avg - 2*std. dev., avg + 2*std dev)

How to choose the value of ‘k’

- k is an integer
- Minimum value of k has to be 2, there will be two iterations in this case
- Max value of k can be the number of data points, this is also known as Leave One Out Cross Validation or LOOCV
- Whatever the value of k chosen, the resulting training and test data should be representative of the unseen data as much as possible
- There is no formula to decide the k but k = 10 is usually considered good
- Too large a k, means less variance across the training sets thus limit the model differences across iterations
- For a sample size of n, And k = p, Number Of Records (r) per fold = n/p