Consider how we can get the most out of our data using k-fold cross validation. Then consider a new training data sampling technique called Bagging.

**Keywords** Cross Validation · k-Fold Cross Validation · Bagging

**Preface** This is the format for ArXiv articles. I wanted to spice things up.

1 Introduction

Cardinal sin is using your test data while choosing your model, so we’re going to introduce a new portion of data. Train, validation and test. We use the validation set (about same size as test set) to choose our model without ever looking at the test data. Good slides visualize this.

Idea: use your data to the maximum capacity by reusing it strategically.

**k-Fold Cross Validation**

Con 1: Reduces training data (especially with very low k). It divides up your training data too much. Generally, with less training data, you can’t do as much cross validation. If you have a very low K (like 2 results in half the training data). K = 3, results in 2/3rds of training data).

Con 2: Larger K ⇒ higher computational cost.

1.1 Choosing k

Larger k better for smaller data sets – like k=n (below).

Three common tactics for choosing a value for k are as follows[1]

**Representative**: The value for k is chosen such that each train/test group of data samples is large enough to be statistically representative of the broader dataset.

**k=10**: The value for k is fixed to 10, a value that has been found through experimentation to generally result in a model skill estimate with low bias a modest variance.

**k=n**: The value for k is fixed to n, where n is the size of the dataset to give each test sample an opportunity to be used in the hold out dataset. This approach is called leave-one-out cross-validation. Extremely expensive, create n different models. But good for small datasets.

To summarize, there is a bias-variance trade-off associated with the choice of k in k-fold cross-validation.

---

[1] Much of the Choosing K info is from here: [https://machinelearningmastery.com/k-fold-cross-validation/](https://machinelearningmastery.com/k-fold-cross-validation/)
Typically, given these considerations, one performs k-fold cross-validation using $k = 5$ or $k = 10$, as these values have been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.

1.2 In Practice

Often don’t separate a test set, they use all their data in the K-fold process.
But normally, we would want to separate out test data first, then only do k-fold on the train data.

2 Bagging

Bagging = Bootstrap Aggregation
Bootstrap - randomly sample with replacement. Replacement is important: never changing/reducing the subset of data you’re drawing from.
Run the learning algorithm on each data set independently.