Unit 12 - Training and test set

Overfitting training data

Imagine a model that would just repeat the labels of the samples that it has just seen.

It would have a perfect score, but fail to predict anything on new data.

This is called **overfitting**.

What is overfitting?

Overfitting happens when your model detects apparent relationships in data that, well, aren't really there.

This means that we see relationships that **don't generalize**.

xkcd explains it.

Is overfitting a common thing in pratice?

Yes.

It's very easy to over fit by including too many things in our models, a somewhat equivalent method to *knowing something by heart*.

But how do I know if I'm overfitting?

A common practice when performing a supervised machine learning model is to **hold out part of the available data as a test set**.

The testing dataset must contain only unknown or first seen data.

This way, we assess how our results will generalize to a new data set and estimate the accuracy of our model in the real world, limiting overfitting.

Train and test split

From the Elements of Statistical Learning book:

- The training set is used to fit the models
- The test set is used for assessment of the generalization error
- Ideally, the test set should be kept in a vault, and be brought out only at the end of the analysis (why?)

How do I test different models?

Keep a validation set (if you have enough data)

Sometimes, we divide the raw data set in three parts:

- A training set
- A *validation* set
- Plus the test set

The validation set is used to **estimate the prediction error for model selection**.

Cross-validation

Cross-validation is a way of measuring predictive performance.

Performing **k-fold cross-validation** requires that the original sample is randomly partitioned into k subsamples and one is left out at each step, or iteration.

One round of cross-validation involves performing the analysis using k-1 subsets, and using the remaining one for validation. The process is repeated k times, and then the k results can then be averaged.

Cross-validation

The advantage is that all observations are used for both training and validation, and each observations is used for validation exactly once.

Cross-validation can be used when there's not enough data to keep partitions without losing modelling or testing capability.