Andrew Gard - equitable.equations@gmail.com

Overfitting

The purpose of machine learning (and statistical inference more generally) is to use sample data to draw conclusions about the population from which its drawn.

The purpose of machine learning (and statistical inference
more generally) is to use sample data to draw conclusions about the population from which its drawn. Typically we have complete information about the former but only really care about the latter.

The purpose of machine learning (and statistical inference more generally) is to use sample data to draw conclusions about the population from which its drawn. Typically we have complete information about the former but only really care about the latter.

We build a model using the sample data with the goal of applying it to new data from the population to make predictions about response variables of interest.

Statistical and machine learning models are built to fit the data on which they're trained as closely as possible within their particular structures.

Statistical and machine learning models are built to fit the data on which they're trained as closely as possible within their particular structures.

For instance, linear regression picks out the line that minimizes the squared residuals (RSE) in the set while logistic regression minimizes residual deviance. A classification models might be optimized to identify as high a proportion of the sample correctly as possible.

Unfortunately, models that performs better on their training data are often *worse* at making predictions from new data.

Unfortunately, models that performs better on their training data are often *worse* at making predictions from new data. **Overfitting** refers to when a model goes too far out of its way to accommodate the peculiarities of the set used to build it, integrating random noise as if it were meaningful information.

Unfortunately, models that performs better on their training data are often *worse* at making predictions from new data. **Overfitting** refers to when a model goes too far out of its way to accommodate the peculiarities of the set used to build it, integrating random noise as if it were meaningful information.

More flexible modeling techniques are more susceptible to overfitting.

Since measuring a model's performance, whether by residual standard error or some other metric, on the data used to train it is guaranteed to be unreliable, another approach is needed.

Since measuring a model's performance, whether by residual standard error or some other metric, on the data used to train it is guaranteed to be unreliable, another approach is needed. The simplest and most computationally efficient is the **testing set** approach.

Since measuring a model's performance, whether by residual standard error or some other metric, on the data used to train it is guaranteed to be unreliable, another approach is needed. The simplest and most computationally efficient is the **testing set** approach.

The idea is simple: the sample data is split into two sets, a **training set** used to fit the model and a **testing set** used to evaluate the model.

The testing set approach has drawbacks.

The testing set approach has drawbacks.

- Setting aside part of the sample for evaluation leaves less for fitting the model. This increases the variability of the model, in the sense that repeated samples are more likely to produce different fits.

The testing set approach has drawbacks.

- Setting aside part of the sample for evaluation leaves less for fitting the model. This increases the variability of the model, in the sense that repeated samples are more likely to produce different fits.
- Under this approach, both the model and the evaluation depend on the specific split. For instance, if the split is done randomly (as is typically the case), repeated builds will lead to different results.

The testing set approach has drawbacks.

- Setting aside part of the sample for evaluation leaves less for fitting the model. This increases the variability of the model, in the sense that repeated samples are more likely to produce different fits.
- Under this approach, both the model and the evaluation depend on the specific split. For instance, if the split is done randomly (as is typically the case), repeated builds will lead to different results.

On the other hand, the approach is easy to understand, and to code. It's also very computationally effective.