✓ 100 XP ▶

Nuances of test sets

6 minutes

Test sets are considered best practice for most aspects of machine learning, though the field is still relatively young, and so exactly how and when is often debated. Let's go through some things to consider.

Test sets can be misleading

Although test sets are helpful to identify overtraining, they can provide us with false confidence. Specifically, test sets are only useful if they reflect data that we expect to see in the real world. For example, our test set is very small, it will not be representative of the variety of data that are likely to be seen in the real world. Test datasets are also only as good as their source. If our test dataset come from a biased source, our metrics will not reflect how things will behave in the real world.

For example, let's say we're trying to find the relationship between number of rescues and the age a dog started training. If our test set was only three dogs, it's possible that these dogs aren't a good representation of the wide variety of working dogs in the real world. Also, imagine that we obtained our test set from a single breeder, who doesn't know how to work with puppies. Our model might predict that older dogs are best to train, and our test dataset would confirm this, when in fact other trainers might have enormous success with younger animals.

Test sets aren't free

We've already seen how the more training data we have, the less likely it's that our model will overfit. Similarly, the larger the test sets, the more we feel we can trust our test results. However, we usually work with finite amounts of data and a datapoint cannot be in both the training and the test set. This means that as we get larger test sets, we get smaller training datasets and vice versa. Exactly how much data should be sacrificed to appear in the test dataset depends on individual circumstances, with anything between 10 - 50% being relatively common, depending on the volume of data available.

Train and test isn't the only approach

It’s worthwhile keeping in mind that train-and-test is common but not the only widely used approach. Two of the more common alternatives are the hold-out approach, and statistical methods.

The hold-out approach

The hold-out approach is like train-and-test, but instead of splitting a dataset into two, it's split into three: *training, test—also known as validation—and hold-out.* The training and test datasets are as we’ve described above. The hold-out dataset is a kind of test set that is used only once, when we're ready to deploy our model for real-world use. In other words, it's not used until we have finish experimenting with different kinds of training regimens, different kinds of models, and so on.

This approach tackles the fact that we usually experiment with different models and training regimens. For example, we fit a model, find it doesn't work well with the test dataset, change some aspects of the model being trained, and try again until we can get a good result. This means we're purposefully altering our model to work for a particular set of data, just like normal training does with the training dataset. By doing this, we can end up with a model that is essentially too overtrained to work on our test dataset!

The idea of a third dataset is we can test for this too. This approach means splitting the data three ways, which means we start with even less training data. If we don't have a lot of data to work with, this approach can reduce our ability to obtain a good model.

Statistical approaches

Simpler models that have originated in statistics often don't need test datasets. Instead, what degree the model is overfit can be calculated directly as statistical significance: a ‘p-value’.

These statistical methods are powerful, well established, and form the foundation of modern science. The advantage is that the training set doesn't ever need to be split and we get a much more precise understanding of how confident we can be about a model. For example, a p-value of 0.01 mean there's a very small chance that our model has found a relationship that doesn't actually exist in the real world. By contrast, a p-value of 0.5 means that while our model might look good with our training data, it will be no better than flipping a coin in the real-world.

The downside to these approaches is that they're only easily applied to certain model types, such as the linear regression models we have been practicing with. For all but the simplest models, these calculations can be extremely complex to perform properly, and so are out of scope for the current course. They also suffer the same limitation regarding data selection – if our training data are biased, our p-values will be misleading.

Next unit: Exercise – Test set nuances

Continue >