How accurate is my accuracy?

Can your Data Scientist tell you how confident they are in their predictions?

Data science and predictive analytics are an integral part of data driven decision making. Now suppose you have commissioned a demand forecasting model from your data science team, and you can now receive detailed predictions for the volume of sales for the next month or the next year. What is the level of confidence you should have in these predictions? Would you require a particular confidence level before introducing the predictive model to your decision-making process?

Fundamental to statistics is the problem of inferring properties of an entire population from the properties of just a representative sample. For example, if we measure the height of ten thousand Australians, what can we infer about the average height of the total population of Australians and what is the confidence interval around our prediction?

When it comes to machine learning and data science, a major complication is that our data is no longer normally distributed. Can we produce some measure of confidence for our data science models?

In general this is a research level question but there is one technique which all data science teams should be familiar with. This is "bootstrapping". I will first describe how bootstrapping would be used when estimating heights of Australians.

We would like to know the mean and variance of the heights of Australians. We have taken just a single sample of ten thousand measurements, so we can compute the mean and standard deviation of this sample. But more than that, we would like to use this sample to estimate the variance of the entire population.

Bootstrapping (or jack-knifing) is an ingenious trick which consists of the following steps:

1. By sampling with replacement, we produce multiple samples, all of size ten thousand from this single sample. It is known that roughly two-thirds of each new sample will be unique, the rest will have multiplicity greater than one.

2. We then compute the average for each sample, followed by computing the variance of all these computations.

3. We relate this variance of the sample-mean to the variance of the entire population.

This is a well-established method, but how can it be used for our new machine learning model? Well, suppose you have a training set of size ten thousand, you can then follow these steps:

1. Resample thirty times with replacement from the training set, to produce thirty training sets, all of size ten thousand.

2. Train your model once on each of these thirty training sets, producing thirty models.

3. Evaluate your metric using the test set for each of the thirty models.

4. Compute the variance of these thirty metric evaluations.

You can find more details in the reference below but following these techniques, your data science team can provide you with a measure of confidence for the predictions of their great new data science models.

If your business is experiencing a lack of confidence in their data science models, or you would like to find about more about data science, please contact a Fujitsu Data & AI specialist now.

Reference: The Elements of Statistical Learning, T. Hastie, R. Tibshirani and J. Friedman, Chapter 7.

Contact

Fujitsu Data & AI
+61 3 9924 3000