Representative-Data

September 28, 2017

1 When Is the Dataset Representative or Balanced?

Representativeness means one thing in machine learning and something different in statistics.

Let's understand what it means in both areas.

Let's start with machine learning.

1.1 Representativeness in Machine Learning

In machine learning, *representativeness of a feature* is a measure of the frequency with which the values of the feature appear. For categorical features it's simply the relative frequency of each possible value of that feature.

If you think back to the telco churn dataset we began the course with, you would have noticed that the feature of being a senior citizen (or not) had values that were distributed as follows.

If we look at the dataset as a whole we find that very few rows have the senior citizen feature value of 1 compared to those rows that have value 0 for this feature. Hence this feature is not representative -- it has many more occurrences of one of the possible values over the other.

Representativeness is highest when the distribution over the range of possible values is *uniform*.

So having 50% senior citizens and 50% not senior citizens in the dataset would make that feature representative. As the percentages shift away from 50% (one feature value going lower than 50% and the other going higher than 50%), the feature becomes more and more unrepresentative.

You can think of it in terms of how *balanced* the feature's values are between 0 (not a senior citizen) and 1 (senior citizen).

There are no rules on when these values go from representative to non-representative. That's a matter of judgement -- but 90% and 10% is clearly not representative and 52% and 48% is clearly representative.

On the other hand, the feature of gender is balanced in the telco churn dataset.

Here is a made-up example to illustrate the point of balance and why it matters to machine learning.

Because the learning algorithm sees very few 2-year contracts, it can't learn as much from this feature value.

On the contrary, here is a representative or balanced feature from a machine learning standpoint.

In this dataset the machine learning algorithm is able to learn (roughly) equally from both values of the feature.

1.2 Balanced Features in Machine Learning

A feature is balanced in a dataset if its values are (roughly) uniformly distributed in that dataset. In particular, machine learning models are sensitive to how balanced the *target* feature in a dataset is.

We'll see this come to the fore when we look at measuring the performance of logistic regression models.

1.3 Representativeness in Statistics

In statistics, representativeness is a characteristic or property of the sample taken from a population.

Let's see what this means...

Here only 10% of the population are Ph.D.s; but 50% of the sample consists of Ph.D.s.

For the sample to be representative, it must have the same proportion of non-Ph.D.s to Ph.D.s that the population has.

In this case, both the population and the sample have the same proportion of non-Ph.D.s to Ph.D.s. The sample is representative of the population.

1.4 Summary

Non-representative samples can lead to terrible mistakes in statistical inference.

One way to make sure you have a representative sample is to use stratified sampling techniques.

Perhaps the most famous error in statistical inference due to non-representative sampling is the 1948 US presidential election.