

stats.terms Populations, samples, and machine learning

An experiment's population is a complete collection of objects that we would like to study. These objects can be people, machines, processes, or anything else we would like to understand experimentally.

Of course, we typically can't measure all of the population. Instead, we take a subset of the population—called a sample—and infer the characteristics of the entire population from this sample.

However, this inference that the sample is somehow representative of the population assumes the sample size is sufficiently large and that the sampling is random. This means selection of the sample should be such that no one group within a population are systematically over- or under-represented in the sample.

Machine learning is a field that makes extensive use of measurements and statistical inference. In it, an algorithm is trained by exposure to sample data, which is called a training set. The variables measured are called features.

Typically, a predictive model is developed that can be used to extrapolate from the data to a new situation. The methods of statistical analysis we introduce in this chapter are the foundation of most machine learning methods.

population

sample

random

machine learning

training

training set

features

predictive model

Example stats.terms– 1

Consider a robot, Pierre, with a particular gravitas and sense of style. He seeks just the right-looking pair of combat boots for wearing in the autumn rains. Pierre is to purchase the boots online via image recognition, and decides to gather data by visiting a hipster hangout one evening to train his style. For contrast, he

re: combat boots

• also watches footage of a White Nationalist rally, focusing special attention on the boots of wearers of khakis and polos. Comment on Pierre's methods.