

Data valuation with Leave-One-Out (LOO) test and Shapley methods

Nathan Martin
Science & Engineering
Manalapan High School
Englishtown, NJ
425nmartin@frhsd.com

Abstract

During my fall internship with WIT, I explored methods to value data from progressive profiling questions that I designed. These profiling questions were split into demographic (username, email, country,) behavioral (actions), and psychographic (values, interests, lifestyle) questions. I was to then design a database that recorded estimated values to serve as a base for future progress. After this, I researched data valuation with two tests: Leave-One-Out (LOO) and Shapley. One valued data by essentially finding the least common answer by eliminating all datasets except one to then record the effect the removal had on the distribution. This was efficient in valuing multiple choice answers but was slow and did not accurately affect each removal on the distribution. After making the program, I worked on valuing with Shapley, the marginal contribution each dataset's removal has with respect to another's removal. This yielded more accurate results and eliminated less valuable data at a more successful rate.

Index Terms

data valuation, leave-one-out test, Shapley methods, internship, data science