The Value of Data

Data valuation, debugging, and understanding is a great challenge today to improve a machine learning model. In this project, we will be focusing on one specific algorithm that uses Shapley value over K-nearest neighbor classifiers for the purpose of data debugging [1].

References: [1] https://arxiv.org/abs/1908.08619

Algorithm and Workpackages

Input:

- A training set, which is already in the format of a N*(d+1) matrix.
- A validation set, which is already in the format of a M*d matrix.

Work Packages

WP1. Implement Algorithm 1 in [1], first with a simple C++/C implementation, and then optimize its performance using what you have learned in this course.

WP2. Implement Algorithm 2 in [1], first with a simple C++/C implementation, and then optimize its performance using what you have learned in this course. Algorithm 2 is an approximation of Algorithm 1, but we expect them to rely on different optimization techniques.

WP3. If you have time after finishing WP1 and WP2, you can play with the LSH-based algorithm in Section 3.2 in [1]. This algorithm replace line 2 in Algorithm 1 with a probing of LSH instead of sorting all numbers. Better writeup will be provided if we reach this point.

We provide a reference implementation in Python for all three parts.