Constraint-Based Learning from Low-Cost Supervision

Bert Huang, Tufts University Dept. of Computer Science and Data Intensive Studies Center
bert@cs.tufts.edu http://berthuang.com

Modern demonstrations of machine learning (ML) have excited practitioners across disciplines about its potential. Unfortunately, most successful ML applications depend on costly annotation. Supervised learning requires expensive, large datasets of individually labeled examples.

My work focuses on methods that learn from low-cost noisy and biased signals as weak supervision. The approaches I have developed model the weak signals as constraints. A learning algorithm is given access to a large unlabeled dataset and a set of weak signals that approximately label the data. Given information such as estimated error rates of the weak signals, each signal defines a restriction on the possible label values for all examples. Using constraints to model the signals avoids the need for statistical assumptions that may cause cascading errors if false.

I have introduced various algorithms that learn by reasoning about these constraints. Adversarial label learning (ALL) trains a model to minimize the worst loss obtainable by any feasible labeling (Arachie & Huang, AAAI '19, JMLR '21). This objective is a form of upper bound on the unknown true error. Constrained label learning (Arachie & Huang, UAI '21) is less pessimistic. It instead trains using a random labeling within the constrained space. Our most recent approach is to add constraints based on data consistency (Arachie & Huang, in submission). We further constrain the space of labels by also requiring that the labels can be output by a function of the data features.

The modeling of weak signals as constraints makes these methods robust against redundant errors in weak signals. We have demonstrated this robustness in various experiments, where our methods outperform other recent approaches to weak supervision.

My future research directions will incorporate other low-cost supervision signals together into a constraint-based framework. For example, we will define constrained optimizations that choose which examples would most benefit an active learning algorithm. And we will use learning-theoretic bounds to define constraints on semi-supervised methods as weak labelers. Our work will continue to characterize the power of constraint-based learning to handle the bias and noise of low-cost labeling.