Programmatic Labeling of Dark Data for Artificial Intelligence in Spatial Informatics

Jason Meil
SAIC, Corporation, United States of America

Data preparation process generally consumes up to 80% of the Data Scientists time, with 60% of that being attributed to cleaning and labeling data.[1] Our solution is to use automated pipelines to prepare, annotate, and catalog data. The first step upon ingestion, especially in the case of real world—unstructured and unlabeled datasets—is to leverage Snorkel, a tool specifically designed around a paradigm to rapidly create, manage, and model training data. Configured properly, Snorkel can be leveraged to temper this labeling bottle-neck through a process called weak supervision. Weak supervision uses programmatic labeling functions—heuristics, distant supervision, SME or knowledge base—scripted in python to generate “noisy labels”. The function traverses the entirety of the dataset and feeds the labeled data into a generative—conditionally probabilistic—model. The function of this model is to output the distribution of each response variable and predict the conditional probability based on a joint probability distribution algorithm. This is done by comparing the various labeling functions and the degree to which their outputs are congruent to each other. A single labeling function that has a high degree of congruence with other labeling functions will have a high degree of learned accuracy, that is, the fraction of predictions that the model got right. Conversely, single labeling functions that have a low degree of congruence with other functions will have low learned accuracy. Each prediction is then combined by the estimated weighted accuracy, whereby the predictions of the higher learned functions are counted multiple times. The result yields a transformation from a binary classification of 0 or 1 to a fuzzy label between 0 and 1—there is “x” probability that based on heuristic “n”, the response variable is “y”. The addition of data to this generative model multi-class inference will be made on the response variables positive, negative, or abstain, assigning probabilistic labels to potentially millions of data points. Thus, we have generated a discriminative ground truth for all further labeling efforts and have improved the scalability of our models. Labeling functions can be applied to unlabeled data to further machine learning efforts.

Once our datasets are labeled and a ground truth is established, we need to persist the data into our delta lake since it combines the most performant aspects of a warehouse with the low-cost storage for data lakes. In addition, the lake can accept unstructured, semi structured, or structured data sources, and those sources can be further aggregated into raw ingestion, cleaned, and feature engineered data layers. By sectioning off the data sources into these “layers”, the data engineering portion is abstracted away from the data scientist, who can access model ready data at any time. Data can be ingested via batch or stream.
The design of the entire ecosystem is to eliminate as much technical debt in machine learning paradigms as possible in terms of configuration, data collection, verification, governance, extraction, analytics, process management, resource management, infrastructure, monitoring, and post verification.