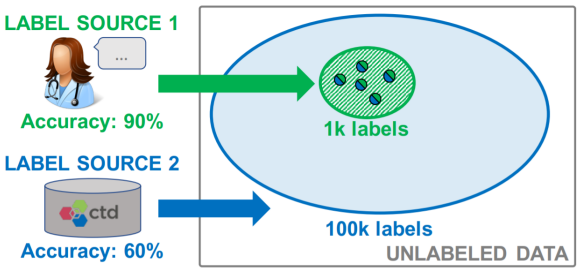
Multi-fidelity supervised learning with Snorkel

Conor Perreault

November 29 2021

Weak Supervision

- Modern ML problems require huge amounts of data. It is expensive to hand label all this data.
- Weak supervision uses noisy labels for training data learn a model that can (hopefully) improve on the performance of the noisy labels.
- This makes more problems feasible because noisy labels can be generated automatically.

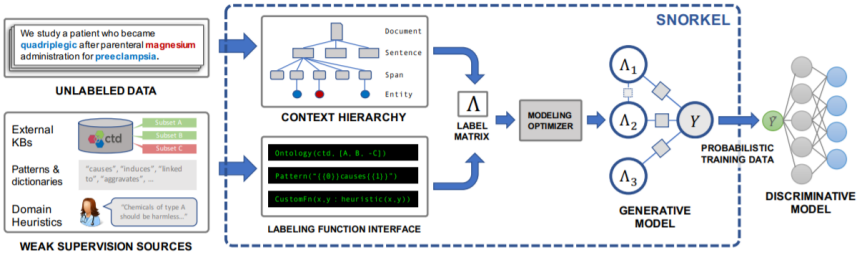


Snorkel

- Snorkel is an open source project that makes weak supervision projects tractable.
- Provides an interface to easily create noisy training labels easily.
- End-to-end implementation of 'data programming' which makes it easy to train ML models for problems where there is not enough labeled data.

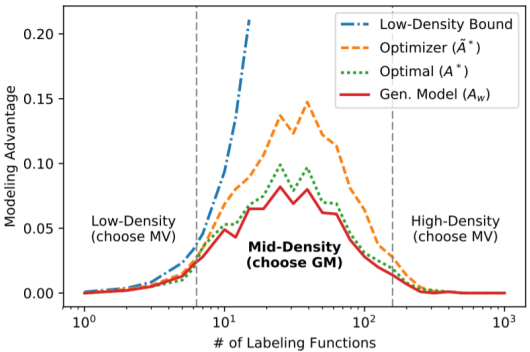
Snorkel Architecture

- Labeling functions: noisy automatic label generators. Should be programmed by subject area experts
- Generative model: model that learns to generate probabilistic labels for training data based on labeling functions.
- Discriminative model: train a predictive model based on probabilistic labels.



Results and Extensions

- Within 3.6% of hand-labeled accuracy on 'average' over a couple datasets
- Creating labeling functions is $2-3\times$ faster than hand labeling data in the examples tested.



Task	Text	Image	Cross-Modal	Cross-Over
CT 1	1.12	1.43	1.52	60k examples
CT 2	1.49	2.32	2.43	50k examples
CT 3	0.88	0.95	1.14	5k examples
CT 4	1.74	2.00	2.45	4k examples
CT 5	1.67	2.03	2.42	750k examples

Generative Model

- For each training data point, the generative model creates a vector describing the 'votes' of each labeling function.
- Snorkel wants to learn coefficients for to account for correlations between the labeling functions.
- Snorkel alternates Gibbs sampling and SGD steps to maximize the likelihood of the aggregated votes by changing correlation parameters.
- From here, Snorkel can create probabilistic training labels.

Discriminative Model

- Any type of discriminative model can be trained on the probabilistic labels created by the generative model.
- Loss function should be noise-aware:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{m} \mathbb{E}_{y \sim \overline{Y}}[l(h_\theta(x_i), y)].$$

- By combining the 'knowledge' of labeling functions, the discriminative model can sometimes generalize beyond the predictions of any labeling function.

Project Overview

- First goal: understand how the accuracy of the labeling functions affects end model performance.
- Can Snorkel discriminate between good and bad labeling functions? How accurate should labeling functions be to avoid 'damaging' a model?
- Extension goal: Design a 'hook' in Snorkel to input true data labels for a small hand-labeled dataset.
- The generalization of this would be to allow prior knowledge of labeling function accuracy.

Project Plan

- Recreate a Snorkel model based on the original paper and tutorials.
- Split true dataset into train/test groups and use 'training data' as an extra labeling function.
- Create a bad labeling function that gives random votes for each input.
- Measure the accuracy and coverage of each labeling function created by domain experts in the paper and tutorial.
- Test different combinations of these labeling functions to learn how well Snorkel actually can generalize from noisy labels.

Project Extensions

- Use Snorkel for multi-fidelity modeling: give it access to a small amount of real data to learn labeling function correlations and errors more accurately.
- Other papers have showed success with multifidelity modeling by training multiple neural networks: one to make predictions based on noisy data, and the rest to learn correlations between noisy and true labels.
- Another idea: train two GPs to estimate the loss functions of 2 different discriminative models: high and low fidelity. Then, pick a label based on these estimates.
- Emukit is another open source project that supports multi-fidelity emulation, could be modified for this architecture.

References

A. Ratner, S. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré: Snorkel: Rapid Training Data Creation with Weak Supervision, VLDB, 2017. A. Ratner, B. Hancock, J. Dunnmon, R. Goldman, and C. Ré:

Snorkel MeTaL: Weak Supervision for Multi-Task Learning, DEEM, 2018.

Andrei Paleyes, Mark Pullin, Maren Mahsereci, Cliff McCollum, Neil D. Lawrence, Javier Gonzalez. Emulation of Physical Processes with Emukit. arXiv:2110.13293. 2021.

Xuhui Meng, George Em Karniadakis. A composite neural network that learns multi-fidelity data: Application to function approximation and inverse PDE problems. Journal of Computational Physics. 2019.

Xuhui Meng, Hessam Babaee, George Em Karniadakis. Multi-Fidelity Bayesian Neural Networks: Algorithms and Applications. CoRR. 2020. Sahaana Suri, Raghuveer Chanda, Neslihan Bulut,

Pradyumna Narayana, Yemao Zeng, Peter Bailis, Sugato Basu, Girija Narlikar, Christopher Re, Abishek Sethi. Leveraging Organizational Resources to Adapt Models to New Data Modalities. VLDB Endowment. 2020.