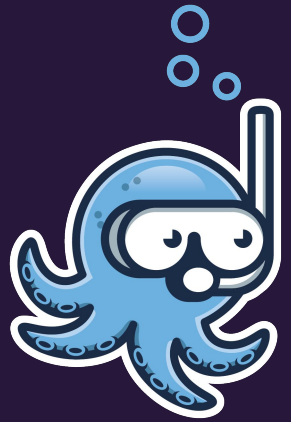
David Gomes

Snorkel



“Programmatically Build and Manage Training Data”

Problems

- The need for a huge amount of labeled data to train deep neural networks
- Cost of SMEs (subject matter experts) to hand-label data
- Poor performance of distant supervision baseline

Objective

- Combine multiple weak supervision sources
- Decrease the cost of labeled data
- Facilitate the job of SMEs to label data

Weak supervision sources

- Patterns
- Heuristics
- Distant supervision
- Crowdsource labels
- Weak and noisy classifiers

Snorkel Architecture

1. Writing labeling functions
2. Modeling accuracies and correlations using a generative model
3. Train a discriminative model using the probabilistic labels

Labeling functions

- Can be written as a Python function or using Snorkel declarative interface
- The function receives the data and must return a label or abstain
- Each labeling function is used as an independent noisy voter
- One may model dependency between labeling functions that express similar heuristics in order to avoid double counting

Labeling functions

```python
from snorkel.labeling import labeling_function


@labeling_function()
def lf_contains_link(x):
    # Return a label of SPAM if "http" in comment text, otherwise ABSTAIN
    return SPAM if "http" in x.text.lower() else ABSTAIN
```

Generative Model

- The generative model is constructed by applying all labeling functions to the unlabeled data points
- It results in a label matrix
- The model is encoded using the labeling propensity, accuracy and pairwise correlations of labeling functions

Discriminative Model

- A discriminative model can be trained with the probabilistic labels
- Predictions can be used by transforming the probabilistic labels
- Then, a discriminative model can be used, such as RandomForest and SVM, trained with the predictions of the generative model

Results

- Exceeds models trained using distant supervision baseline by an average of 132%
- By writing a few dozen of labeling functions, the Snorkel performance is able to approach results using hand-labeled data
- Facilitates its use by offering an easy interaction paradigm of writing labeling functions

Thanks!