Training Data: The New New Oil in Today’s ML

Motivation:
- For many of today’s models (e.g. deep learning), we no longer need to do manual feature engineering!
- However, these models require massive training sets...
- Training set creation & management is the key bottleneck in real-world applications! The current way of hand-labeling data is prohibitively expensive, slow, and brittle

Example ML pipeline today (information extraction problem):

<table>
<thead>
<tr>
<th>Unstructured Input</th>
<th>Large-scale hand-labeled training set</th>
<th>Structured Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td></td>
<td>No feature engineering...</td>
</tr>
</tbody>
</table>

Generating Training Data Programmatically

Creating Noisy Training Sets with Labeling Functions
In data programming, users write labeling functions (LFs), which are just scripts that noisily label subsets of the data. An example where we label relations in text based on an existing knowledgebase:

```
def if(v): cid = (x, gene, x, pheno) if cid in KB: return 1 else: return 0
```

A Unifying Framework for Weak Supervision

We can also include dependencies between the LFs!

The Data Programming Pipeline: Modeling the Training Set Creation Process

DOMAINEXPERTS

\[
\begin{align*}
\lambda_1 : X &\mapsto Y \cup \{0\} \\
\lambda_2 &
\end{align*}
\]

Labeling Functions (LFs)
In data programming, users focus on writing labeling functions (LFs). Labeling functions encode various heuristics or weak supervision signals that programatically label training data.

Generative Model
The LFs define a generative model of the labeling process. By learning this model, based on the overlaps between LFs, we learn the accuracies of the LFs and are able to denoise the training set they generate.

Noise-Aware Discriminative Model
We then train any discriminative model using a modified noise-aware loss function, which simply minimizes the expected loss with respect to the predictions of the generative model.

Theorem: Scaling with Unlabeled Data
Given a constant-order number of LFs, we get the same asymptotic scaling as in supervised methods—but with respect to unlabeled data!

Theorem: Independent Case*

If:
1. \(\pi\) can be represented by our model family
2. Our noise-aware risk minimizer has bounded gen. risk
3. \(y \perp f(x) | \lambda(x)\)
4. We have a sufficient number of LFs with enough coverage & accuracy

Then:
\[
\hat{\delta}(e^{-c}) \text{ unlabeled} \text{ training points allow the algorithm to achieve } O(e^{-c}) \text{ generalization risk (using SGD + Gibbs sampling)}
\]

*For full theorem statement, more general form, and corresponding theorem for the case when LF dependencies are included, see paper

Experimental Results: Information Extraction from Text
We test data programming (DP) on text information extraction problems, comparing to a distant supervision approach where rules to create training data were encoded as a simple if-then-return (ITR) script

We get significant improvements across a range of applications and LFs—notably, even more so with deep learning approaches!