Motivation: The Training Data Bottleneck

Multi-task learning models require large labeled training sets... for each task!

Snorkel MeTaL: An Open-Source System for Multi-Task Weak Supervision

1. Users write labeling functions for multiple related tasks
2. We model the labeling functions’ behavior to denoise them
3. We use the probabilistic labels to train a multi-task model

Modeling Multi-Task Supervision as Matrix Completion

Representation & Model

- Y
  - X1
  - X2
  - X3
  - X4
  - X5

Probabilistic graphical model defined by source dependency graph G, with latent label Y

\[ Y = \text{Labeling Functions} \]

\[ X_1, X_2, X_3, X_4, X_5 \]

Algorithm: Modeling Weak Supervision as Matrix Completion

1. Split into observed (O) and separator set (S) cliques; consider the covariance & inverse covariance matrices
   \[ \text{Cov}(Y \cup S) = \Sigma = \begin{bmatrix} \Sigma_O & \Sigma_{OS} \\ \Sigma_{SO} & \Sigma_S \end{bmatrix} \]
   \[ K = \Sigma^{-1} = \begin{bmatrix} K_O & K_{OS} \\ K_{SO} & K_S \end{bmatrix} \]
   We use the recent result that K has graph-structured sparsity (Loh & Wainwright 2013); i.e. has zeros for indices \( \Omega \) where no edge in our graphical model!

2. Use block matrix inversion lemma to rearrange into an equation w/ observed matrix low-rank parameters, and a graph-structured sparse matrix
   \[ \hat{K}_O = \Sigma_{O}^{-1} + zz^T \]
   \[ \hat{K}_S = \Sigma_{S}^{-1} + zz^T \]
   Sparse
   Low-rank

3. Result: A matrix completion problem w.r.t. the inverse observed covariance!
   \[ \hat{y} = \text{arg} \min_x \| y - x \|_2^2 \]
   This is a standard problem to solve!
   We then show that a simple, deterministic check of the dependency graph \( G \) can determine identifiability of this solution

Theory

I. Result: Given \( n \) unlabeled data points, the source accuracy & correlation estimation error decreases as \( n \to \infty \)

\[ \hat{E}_{\text{corr}} = \frac{\mu_{\text{corr}}}{\sqrt{n}} \]

II. Result: Given standard assumptions on relationship between end model features and our labels (see paper), end model generalization controlled by the above as well:

\[ E_{\text{generalization}} = \frac{\sigma}{\sqrt{n}} \]

Synthetic Experiments: Accuracies, Correlations, and Runtime Scaling

- Our estimates of the label sources’ accuracies improve with more unlabeled data
- Our approach successfully handles the effect of a given source dependency structure (G)
- Our approach is orders-of-magnitude faster than previous Gibbs sampling-based techniques

Real-World Experiments: Fine Grain Text Tagging & Classification

- NER
  - Gold Dev: 63.7 ± 2.1
  - MV: 76.9 ± 2.6
  - DP [28]: 78.4 ± 1.2
- RE
  - Doc: 41.6 ± 0.9
- Average: 51.6, 65.0, 67.7

显著性能增长在三类测试应用中显著：(i) 一个小型手标数据集，对应时间相似于完善监督（Gold Dev）；(ii) 大多数来自的（MV）；(iii) 一个单任务弱监督训练方法(DP)

Scaling with Unlabeled Data

As predicted by theory, our end model accuracy scales with more unlabeled data

Open-source code and tutorials: github.com/HazyResearch/metal