More data means less inference: A pseudo-max approach to structured learning

W10

David Sontag
Microsoft Research

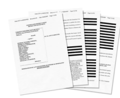
Ofer Meshi
Hebrew University

Tommi Jaakkola
MIT

Amir Globerson
Hebrew University

Structured prediction

- Multi-label prediction:

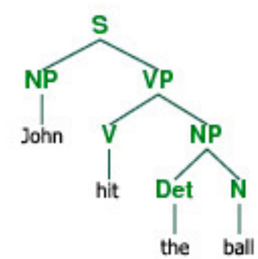
x : 

Politics Sports Finance

$\mathbf{y}$: ✓ ✗ ✓

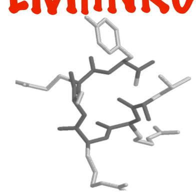
- Parsing of natural language:

x : John hit the ball

$\mathbf{y}$: 

- Protein side-chain placement:

x : KDLMHNKCYHFFM

$\mathbf{y}$: 

More data means less inference: A pseudo-max approach to structured learning

W10

David Sontag
Microsoft Research

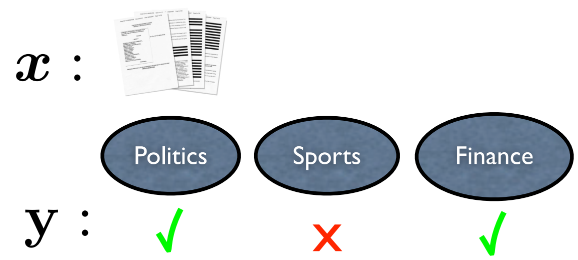
Ofer Meshi
Hebrew University

Tommi Jaakkola
MIT

Amir Globerson
Hebrew University

Structured prediction

- Multi-label prediction:



- Each prediction task is specified by a feature function $f(\boldsymbol{x}, \mathbf{y})$ and a weight vector $\mathbf{w}$.
- Prediction is given by $\mathbf{y} \leftarrow \underset{\hat{\mathbf{y}} \in Z}{\operatorname{argmax}} \; \mathbf{w} \cdot f(\boldsymbol{x}, \hat{\mathbf{y}})$
- Typically decomposes as $f(\boldsymbol{x}, \mathbf{y}) = \sum_c f_c(\boldsymbol{x}, \mathbf{y}_c)$, where c is a small set of variables

More data means less inference: A pseudo-max approach to structured learning

W10

David Sontag
Microsoft Research

Ofer Meshi
Hebrew University

Tommi Jaakkola
MIT

Amir Globerson
Hebrew University

Learning problem (separable setting)

- Given training data $\{\boldsymbol{x}_m, \mathbf{y}_m\}_{m=1}^{M}$
- Assume that there exist "true" parameters $\mathbf{w}$ such that

$$\mathbf{y}_m \leftarrow \underset{\hat{\mathbf{y}} \in Z}{\operatorname{argmax}}\ \mathbf{w} \cdot f(\boldsymbol{x}_m, \hat{\mathbf{y}}) \qquad \text{for all } m$$

- Structured perceptron, stochastic subgradient, cutting-plane, ...

  All repeatedly do prediction during learning - very slow!

- Is there some way to circumvent prediction during learning?
- We give an efficient learning algorithm which, when distribution of training examples is sufficiently "nice", is asymptotically consistent

More data means less inference: A pseudo-max approach to structured learning

W10

David Sontag, Microsoft Research
Ofer Meshi, Hebrew University
Tommi Jaakkola, MIT
Amir Globerson, Hebrew University

The Pseudo-max Method

- Exact: $\{\mathbf{w} \cdot f(\boldsymbol{x}^m, \mathbf{y}^m) > \mathbf{w} \cdot f(\boldsymbol{x}^m, \mathbf{y}), \ \forall m \text{ and } \mathbf{y} \neq \mathbf{y}^m\}$
- Pseudo-max: $\{\mathbf{w} \cdot f(\boldsymbol{x}^m, \mathbf{y}^m) > \mathbf{w} \cdot f(\boldsymbol{x}^m, \mathbf{y}^m_{-i}, y_i), \ \forall m \text{ and } i, y_i \neq y_i^m\}$
  - Very small number of constraints: M × #Vars × #Values
- Does this ever work?
  - Yes, under some conditions on $p(\boldsymbol{x})$.
  - When f corresponds to a pairwise Markov random field, these constraints suffice to identify $\mathbf{w}^*$.
- We also show how to apply to non-separable setting
- **Very fast**, and gives good results for multi-label prediction and protein side-chain placement