Partially Observed Maximum Entropy Discrimination Markov Networks

Jun Zhu
Dept. of Comp. Sci. & Tech., Tsinghua University
Machine Learning Dept., Carnegie Mellon University

Joint work with Eric P. Xing and Bo Zhang

Outline

- Introduction
  - Structured Prediction
  - Max-margin Markov Networks
- Max-Entropy Discrimination Markov Networks (MaxEnDNet)
  - Basic Theorems
  - Partially Observed MaxEnDNet
  - Experimental Results
- Summary

Classification

- Inputs:
  - a set of training samples $\mathcal{D} = \{(\mathrm{x}^i, y^i)\}_{i=1}^N$, where $\mathrm{x}^i = [x_1^i, x_2^i, \cdots, x_d^i]^\top$ and $y^i \in C \triangleq \{c_1, c_2, \cdots, c_L\}$
- Outputs:
  - a predictive function $h(\mathrm{x})$: $y^\star = h(\mathrm{x}) \triangleq \arg\max_y F(\mathrm{x}, y; \mathbf{w})$
- Examples ($F(\mathrm{x}, y; \mathbf{w}) = \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)$):
  - Logistic Regression
    - Max-likelihood estimation

$$\max_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \sum_{i=1}^{N} \log p(y^i|\mathrm{x}^i)$$

$$p(y|\mathrm{x}) = \frac{\exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)\}}{\sum_{y'} \exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y')\}}$$

  - Support Vector Machine (SVM)
    - Max-margin learning

$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^{N} \xi_i;$$

$$\text{s.t.} \quad \mathbf{w}^\top \Delta\mathbf{f}_i(y) \geq 1 - \xi_i, \ \forall i, \forall y \neq y^i.$$

Structured Prediction

- Inputs:
  - a set of training samples: $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$, where

$$\mathbf{x}^i = (\mathrm{x}_1^i, \mathrm{x}_2^i, \cdots, \mathrm{x}_{\ell_i}^i) \text{ and } \mathrm{x}_j^i = [x_{j1}^i, \cdots, x_{jd}^i]^\top$$
$$\mathbf{y}^i = (y_1^i, y_2^i, \cdots, y_{\ell_i}^i) \text{ and } y_j^i \in C \triangleq \{c_1, c_2, \cdots, c_L\}$$

  - Examples:
    - Part-of-speech (POS) Tagging:

      $\mathbf{x}$ = "Do you want fries with that?" -> $\mathbf{y}$ = <verb pron verb noun prep pron>
    - Image segmentation

$$\mathbf{x} = \begin{pmatrix} \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots \\ \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_{11} & y_{12} & \cdots \\ y_{21} & y_{22} & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix}$$

- Outputs:
  - a predictive function $h(\mathbf{x}) : \mathbf{y}^\star = h(\mathbf{x}) \triangleq \arg\max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}; \mathbf{w})$

Structured Prediction Models

- Conditional Random Fields (CRFs) (Lafferty et al., 2001)
  - Based on Logistic Regression
  - Max-likelihood estimation (point-estimate)

$$\max_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \sum_{i=1}^{N} \log p(\mathbf{y}^i | \mathbf{x}^i)$$

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp\{\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})\}}{\sum_{\mathbf{y}'} \exp\{\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}')\}}$$

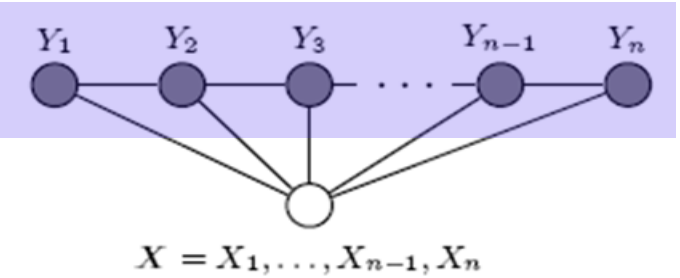
- Max-margin Markov Networks (M^3Ns) (Taskar et al., 2003)
  - Based on SVM
  - Max-margin learning (point-estimate)

$$\text{P0 } (\text{M}^3\text{N}): \min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{s.t. } \forall i, \forall \mathbf{y} \neq \mathbf{y}^i : \mathbf{w}^\top \Delta \mathbf{f}_i(\mathbf{y}) \geq \Delta \ell_i(\mathbf{y}) - \xi_i,\ \xi_i \geq 0$$

where $\mathbf{w}^\top \Delta \mathbf{f}_i(\mathbf{y})$ denotes the margin and $\Delta \ell_i(\mathbf{y})$ is a loss function.

- Markov properties are encoded in the feature functions $\mathbf{f}(\mathbf{x}, \mathbf{y})$



Between MLE and max-margin learning

- Likelihood-based estimation
  - Probabilistic (joint/conditional likelihood model)
  - Easy to perform Bayesian learning, and consider prior knowledge, missing data

- Max-margin learning
  - Non-probabilistic (concentrate on input-output mapping)
  - Not obvious how to perform Bayesian learning or consider prior, and missing data
  - Sound theoretical guarantee with limited samples

- Maximum Entropy Discrimination (MED) (Jaakkola, et al., 1999)
  - A Bayesian learning approach $\hat{y} = \text{sign} \int p(\mathbf{w}) F(x; \mathbf{w}) \, d\mathbf{w}$ $(y \in \{+1, -1\})$
  - The optimization problem (binary classification)

$$\min \; KL(p(\Theta)||p_0(\Theta))$$

$$\text{s.t.} \int p(\Theta)[y_i F(x_i; \mathbf{w}) - \xi_i] \, d\Theta \geq 0, \forall i,$$

MED subsumes SVM.

where Θ is the parameter $\mathbf{w}$ when ξ are kept fixed or the pair $(\mathbf{w}, \xi)$ when we want to optimize over ξ

MaxEnt Discrimination Markov networks

- MaxEnt Discrimination Markov Networks (MaxEnDNet):

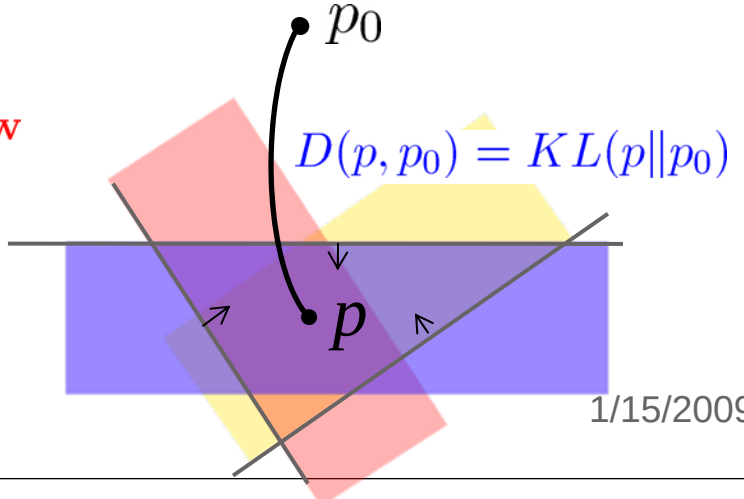
$$\text{P1(MaxEnDNet)}: \quad \min_{p(\mathbf{w}),\xi} \; KL(p(\mathbf{w})||p_0(\mathbf{w})) + U(\xi)$$
$$\text{s.t.} \;\; p(\mathbf{w}) \in \mathcal{F}_1, \;\; \xi_i \geq 0, \forall i.$$

  - *Generalized* maximum entropy or regularized KL-divergence
  - Subspace of distributions defined with *expected* margin constraints

$$\mathcal{F}_1 = \{p(\mathbf{w}) : \int p(\mathbf{w})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta \ell_i(\mathbf{y})]\,\mathrm{d}\mathbf{w} \geq -\xi_i, \;\; \forall i, \forall \mathbf{y} \neq \mathbf{y}^i\},$$

- Bayesian-style Prediction

$$h_1(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \int p(\mathbf{w}) F(\mathbf{x},\mathbf{y};\mathbf{w})\,\mathrm{d}\mathbf{w}$$



Solution to MaxEnDNet

- Theorem 1 (Solution to MaxEnDNet):
  - Posterior Distribution:

$$p(\mathbf{w}) = \frac{1}{Z(\alpha)} p_0(\mathbf{w}) \exp\left\{ \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) [\Delta F_i(\mathbf{y}; \mathbf{w}) - \Delta \ell_i(\mathbf{y})] \right\}$$

  - Dual Optimization Problem:

$$\text{D1}: \quad \max_{\alpha} \; -\log Z(\alpha) - U^\star(\alpha)$$
$$\text{s.t.} \;\; \alpha_i(\mathbf{y}) \geq 0, \;\; \forall i, \;\; \forall \mathbf{y},$$

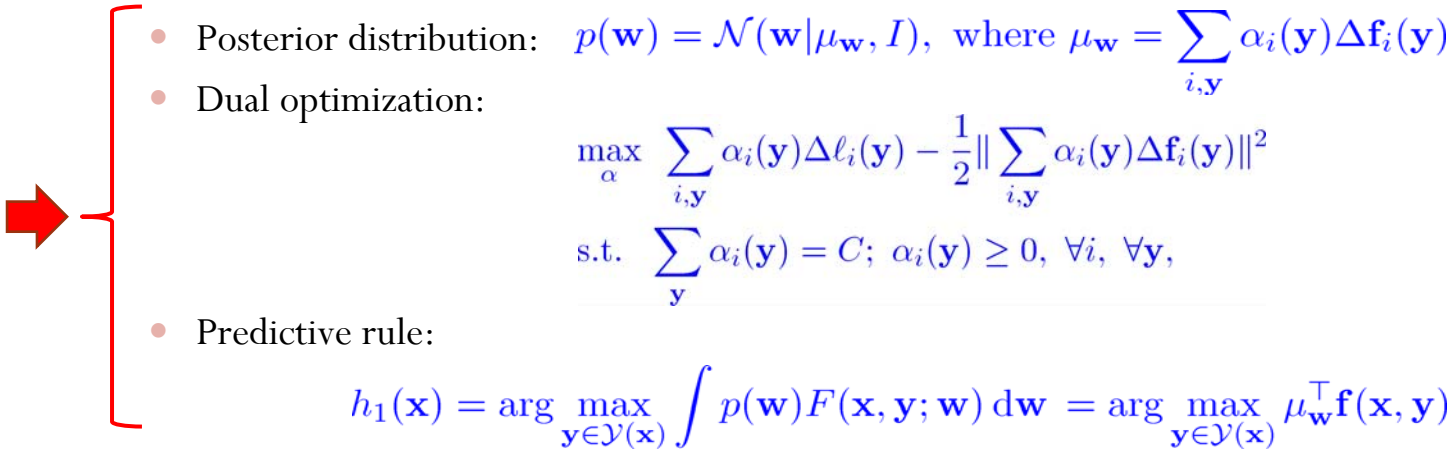
  - $U^\star(\cdot)$ is the conjugate of the $U(\cdot)$, i.e., $U^\star(\alpha) = \sup_\xi \left(\sum_{i,\mathrm{y}} \alpha_i(\mathrm{y}) \xi_i - U(\xi) \right)$
- Convex conjugate (closed proper convex $\phi(\mu)$)
  - Def: $\phi^\star(\nu) = \sup_\mu [\nu^\top \mu - \phi(\mu)]$
  - Ex:

$$\phi(p(\mu)) = KL(p(\mu)||p_0(\mu)) \qquad \phi^\star(\nu) = \log Z(\nu)$$

$$\phi(\xi) = C \sum_i |\xi_i| \qquad \phi^\star(\nu) = \mathbb{I}_\infty(|\nu_i| \leq C, \;\; \forall i)$$

Reduction to M^{3}Ns

- Theorem 2 (Reduction of MaxEnDNet to M^{3}Ns):
  - Assume $F(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}), U(\xi) = C \sum_i \xi_i$, *and* $p_0(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, I)$



  - Posterior distribution: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}, I)$, where $\mu_{\mathbf{w}} = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) \Delta \mathbf{f}_i(\mathbf{y})$
  - Dual optimization:

$$\max_{\alpha} \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) \Delta \ell_i(\mathbf{y}) - \frac{1}{2} \| \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) \Delta \mathbf{f}_i(\mathbf{y}) \|^2$$

$$\text{s.t.} \quad \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C; \ \alpha_i(\mathbf{y}) \geq 0, \ \forall i, \ \forall \mathbf{y},$$

  - Predictive rule:

$$h_1(\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \int p(\mathbf{w}) F(\mathbf{x}, \mathbf{y}; \mathbf{w}) \, d\mathbf{w} = \arg \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \mu_{\mathbf{w}}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})$$

- Thus, MaxEnDNet subsumes M^{3}Ns and admits all the merits of max-margin learning
- Furthermore, MaxEnDNet has at least three advantages …

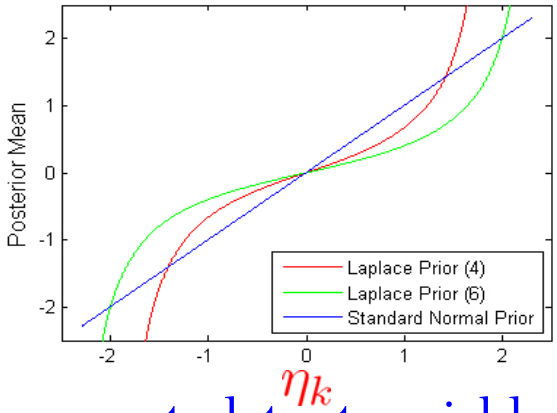
Three Advantages

- PAC-Bayesian prediction error guarantee

$$\Pr_Q(M(h,\mathbf{x},\mathbf{y}) \le 0) \le \Pr_{\mathcal{D}}(M(h,\mathbf{x},\mathbf{y}) \le \gamma) + O\left(\sqrt{\frac{\gamma^{-2}KL(p||p_0)\ln(N|\mathcal{Y}|) + \ln N + \ln \delta^{-1}}{N}}\right).$$

- Introduce regularization effects, such as sparsity bias
  - Laplace prior => Posterior shrinkage effects (Laplace M^3N, ICML'08)

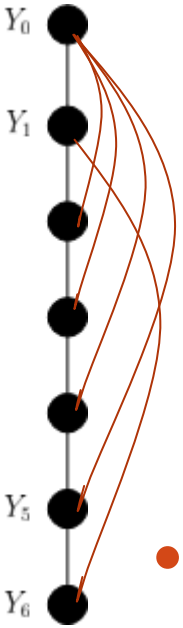
$$\forall k,\ \langle w_k \rangle_p = \frac{2\eta_k}{\lambda - \eta_k^2}$$



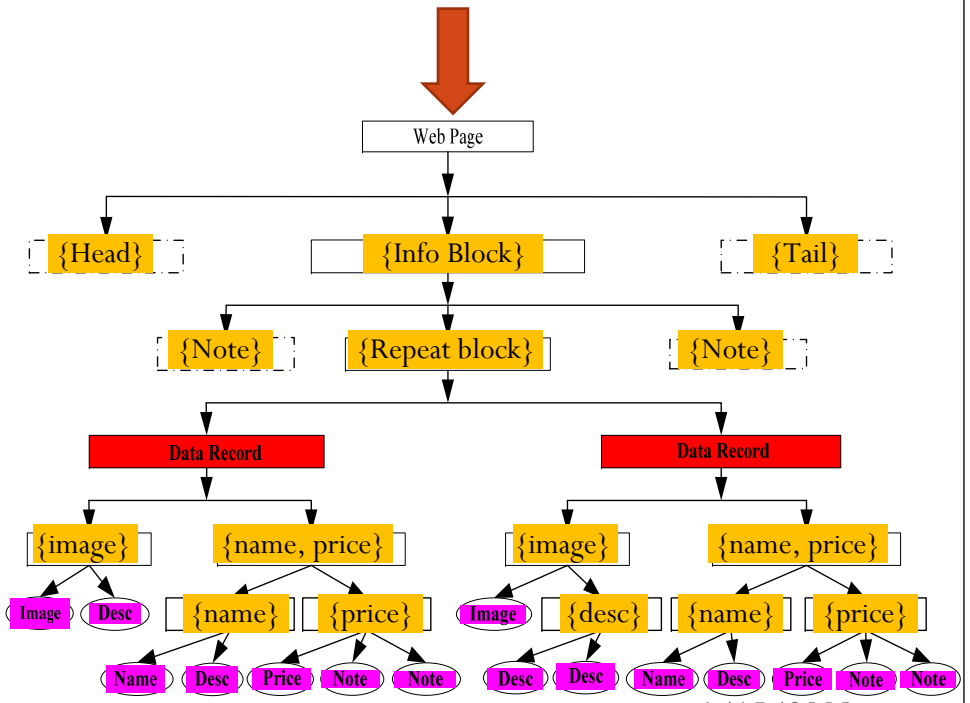
- An elegant approach to incorporate latent variables and structures
  - Partially observed MaxEnDNet

Motivating Example

- Web data extraction
  - Goal: *Name*, *Image*, *Price*, *Description*, etc.



Given Data Record



- Hierarchical model
- Advantages:
  - Computational efficiency
  - Long-range dependency
  - Joint extraction

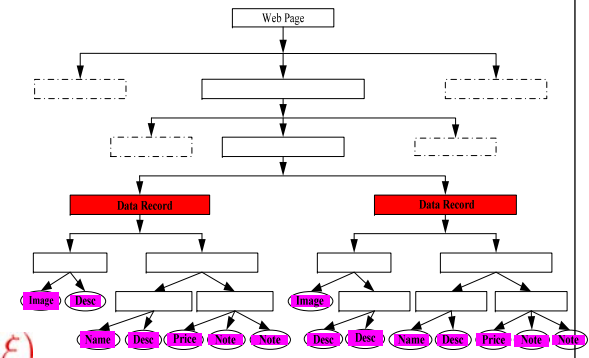
Learn hierarchical model with latent variables

- Can we learn a hierarchical model from partially labeled data?
  - Yes!
  - Partially observed CRFs for object recognition (NIPS'04)
  - Dynamic Hierarchical MRFs for web data extraction (ICML'07)
- How about max-margin learning?
  - Yes!
  - Easy with MaxEnDNet

Partially observed MaxEnDNet

- $\mathcal{D} = \{< \mathbf{x}^i, \mathbf{y}^i >\}_{i=1}^N$ augmented with hidden variables $\{\mathbf{z}\} = (\mathbf{z}^1, \cdots, \mathbf{z}^N)$

$$\text{P1(MaxEnDNet)}: \quad \min_{p(\mathbf{w}),\xi} \; KL(p(\mathbf{w})||p_0(\mathbf{w})) + U(\xi)$$
$$\text{s.t.} \;\; p(\mathbf{w}) \in \mathcal{F}_1, \;\; \xi_i \geq 0, \forall i.$$



- Def of PoMEN:

$$\text{P2(PoMEN)}: \quad \min_{p(\mathbf{w},\{\mathbf{z}\}),\xi} \; KL(p(\mathbf{w},\{\mathbf{z}\})||p_0(\mathbf{w},\{\mathbf{z}\})) + U(\xi)$$
$$\text{s.t.} \;\; p(\mathbf{w},\{\mathbf{z}\}) \in \mathcal{F}_2, \;\; \xi_i \geq 0, \forall i.$$

$$\mathcal{F}_2 = \{p(\mathbf{w},\{\mathbf{z}\}) : \sum_{\mathbf{z}} \int p(\mathbf{w},\mathbf{z})[\Delta F_i(\mathbf{y},\mathbf{z};\mathbf{w}) - \Delta \ell_i(\mathbf{y})]\, d\mathbf{w} \geq -\xi_i, \;\; \forall i, \forall \mathbf{y} \neq \mathbf{y}^i\},$$

- Prediction:

$$h_2(\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \sum_{\mathbf{z}} \int p(\mathbf{w},\mathbf{z}) F(\mathbf{x},\mathbf{y},\mathbf{z};\mathbf{w})\, d\mathbf{w}$$

Alternating Minimization Alg.

- Factorization assumption:

$$p_0(\mathbf{w}, \{\mathbf{z}\}) = p_0(\mathbf{w}) \prod_{i=1}^{N} p_0(\mathbf{z}_i) \qquad p(\mathbf{w}, \{\mathbf{z}\}) = p(\mathbf{w}) \prod_{i=1}^{N} p(\mathbf{z}_i)$$

- Alternating minimization:
  - Step 1: keep $p(\mathbf{z})$ fixed, optimize over $p(\mathbf{w})$

$$\min_{p(\mathbf{w}),\xi} \; KL(p(\mathbf{w})||p_0(\mathbf{w})) + C\sum_i \xi_i$$
$$\text{s.t.} \;\; p(\mathbf{w}) \in \mathcal{F}_1', \;\; \xi_i \geq 0, \forall i.$$

$$\mathcal{F}_1' = \{p(\mathbf{w}) : \int p(\mathbf{w}) E_{p(\mathbf{z})}[\Delta F_i(\mathbf{y}, \mathbf{z}; \mathbf{w}) - \Delta \ell_i(\mathbf{y})] \, d\mathbf{w} \geq -\xi_i, \; \forall i, \; \forall \mathbf{y}\}$$

- Normal prior
  - M^3N problem (QP)
- Laplace prior
  - Laplace M^3N problem (VB)

  - Step 2: keep $p(\mathbf{w})$ fixed, optimize over $p(\mathbf{z})$

$$\min_{p(\mathbf{w}),\xi} \; KL(p(\mathbf{z})||p_0(\mathbf{z})) + C\xi_i$$
$$\text{s.t.} \;\; p(\mathbf{z}) \in \mathcal{F}_1^\star, \;\; \xi_i \geq 0.$$

$$\mathcal{F}_1^\star = \{p(\mathbf{z}) : \sum_{\mathbf{z}} p(\mathbf{z}) \int p(\mathbf{w})[\Delta F_i(\mathbf{y}, \mathbf{z}; \mathbf{w}) - \Delta \ell_i(\mathbf{y})] \, d\mathbf{w} \geq -\xi_i, \; \forall i, \; \forall \mathbf{y}\}$$

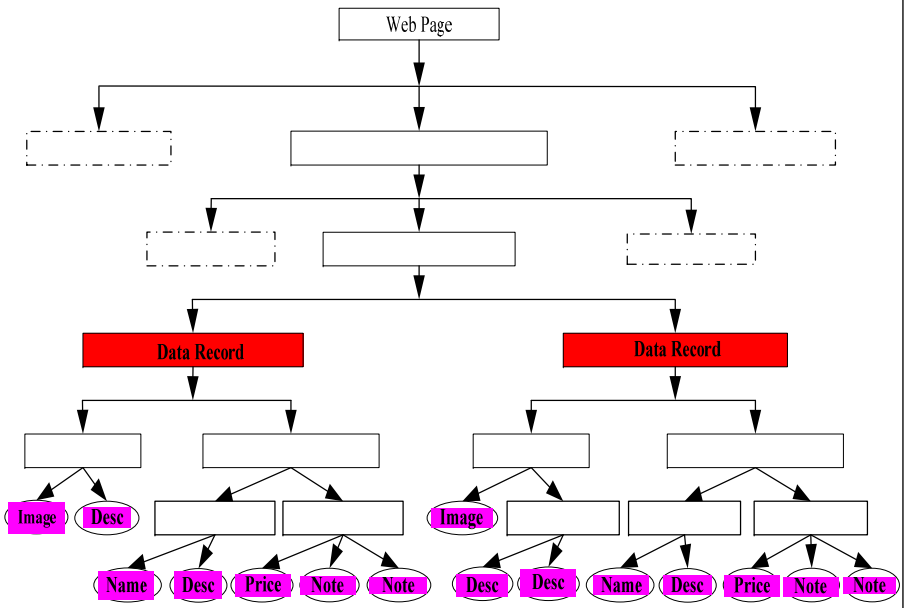
$$p(\mathbf{z}) = \frac{1}{Z(\beta)} p_0(\mathbf{z}) \exp\{\sum_{\mathbf{y}} \beta(\mathbf{y})[\mu_{\mathbf{w}}^\top \Delta \mathbf{f}_i(\mathbf{y}, \mathbf{z}) - \Delta \ell_i(\mathbf{y})]\}$$

$$\max_{\beta} -\log\Big(\sum_{\mathbf{z}} p_0(\mathbf{z}) \exp\{\sum_{\mathbf{y}} \beta(\mathbf{y})[\mu_{\mathbf{w}}^\top \Delta \mathbf{f}_i(\mathbf{y}, \mathbf{z}) - \Delta \ell_i(\mathbf{y})]\}\Big)$$
$$\text{s.t.} \;\; \sum_{\mathbf{y}} \beta(\mathbf{y}) = C, \;\; \beta(\mathbf{y}) \geq 0, \;\; \forall \mathbf{y}.$$

Equivalently reduced to a NLP with a polynomial number of constraints

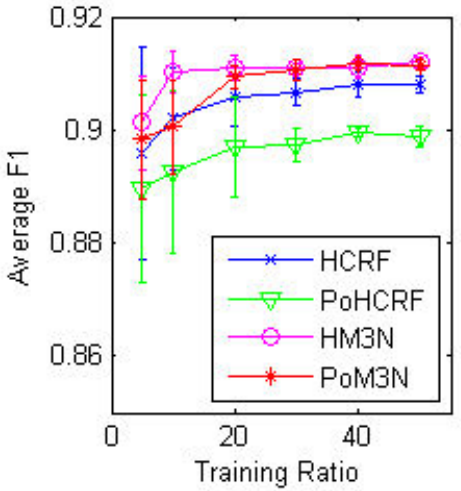
Experimental Results

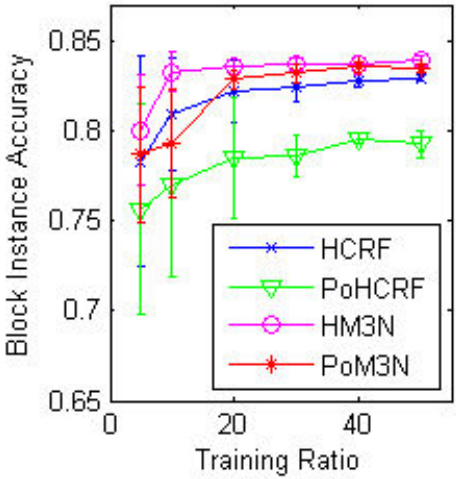
- Web data extraction:
  - *Name, Image, Price, Description*
  - Methods:
    - Hierarchical CRFs, Hierarchical M^3N
    - PoMEN, Partially observed HCRFs
  - Pages from 37 templates
    - Training: 185 (5/per template) pages, or 1585 data records
    - Testing: 370 (10/per template) pages, or 3391 data records
  - Record-level Evaluation
    - Leaf nodes are labeled
  - Page-level Evaluation
    - Supervision Level 1:
      - Leaf nodes and data record nodes are labeled
    - Supervision Level 2:
      - Level 1 + the nodes above data record nodes

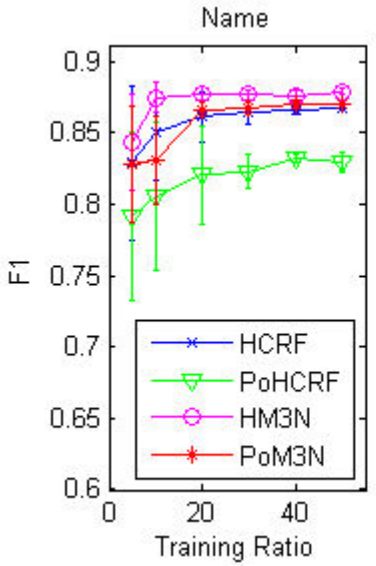


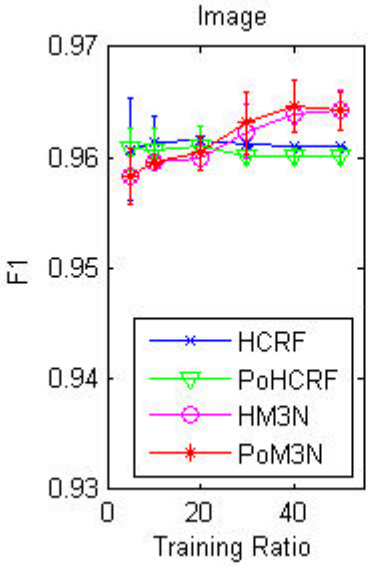
Record-Level Evaluations

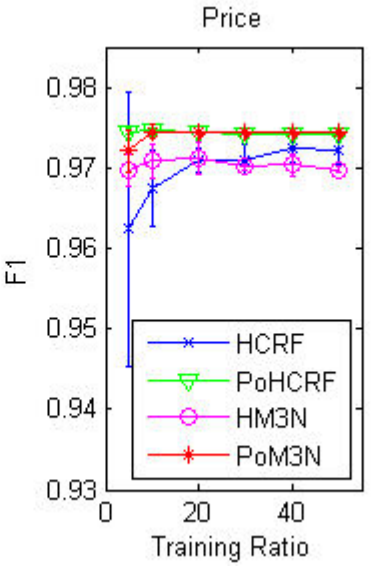
- Overall performance:
  - Avg F1:
    - avg F1 over all attributes
  - Block instance accuracy:
    - % of records whose *Name*, *Image*, and *Price* are correct
- Attribute performance:

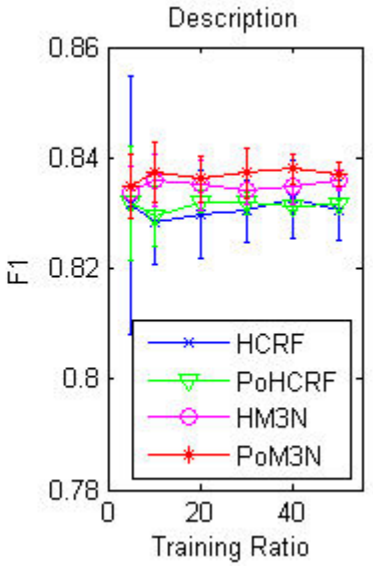










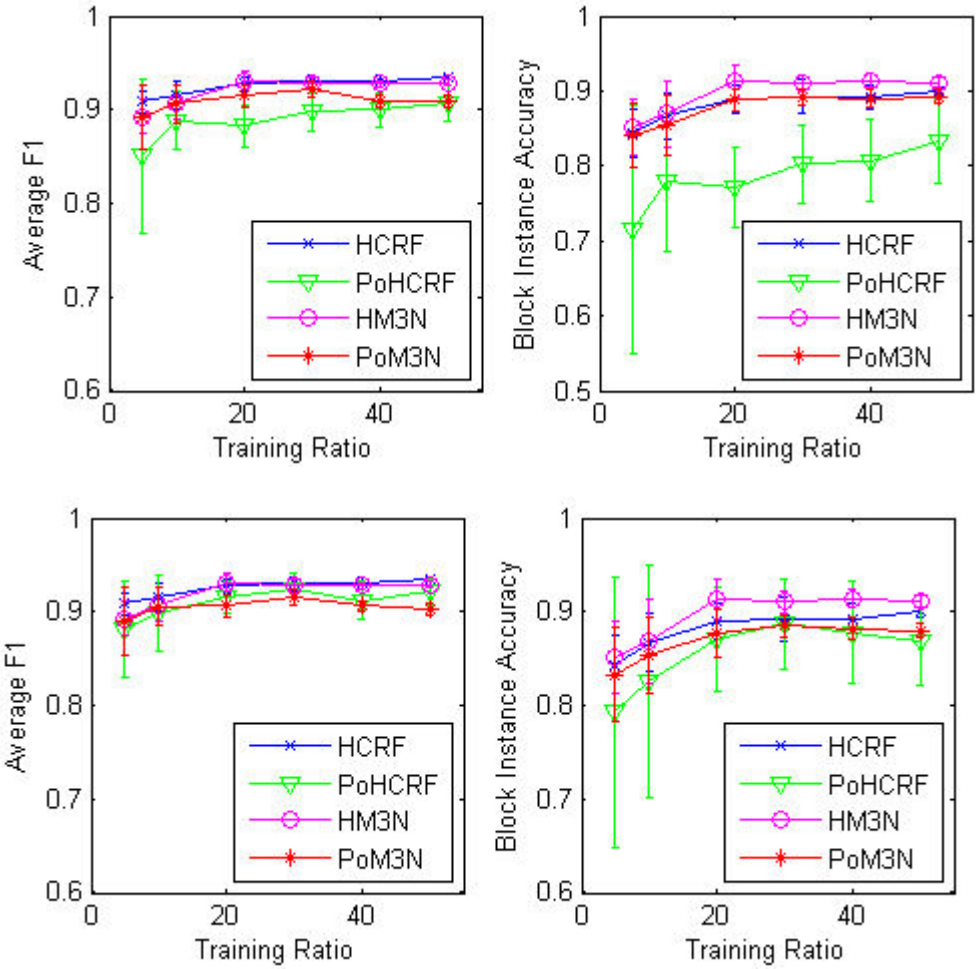


Machin

Page-Level Evaluations

- Supervision Level 1:
  - Leaf nodes and data record nodes are labeled

- Supervision Level 2:
  - Level 1 + the nodes above data record nodes

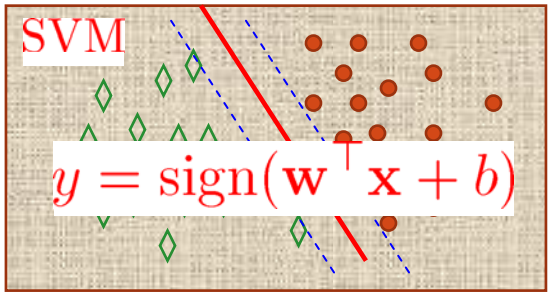


Summary

- MaxEnt Discrimination Markov Networks (MaxEnDNet)
  - PAC-Bayesian performance guarantee
  - Sparsity regularization effects
  - Incorporating latent variables and structures

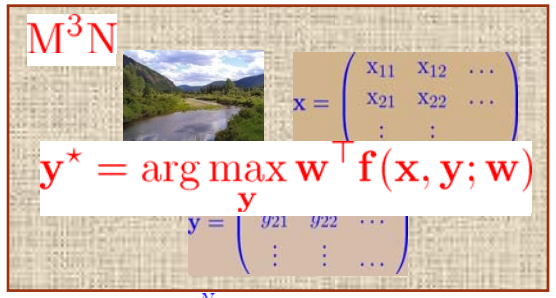
- Experimental results show the advantages of max-margin learning over likelihood methods with latent variables

Margin-based Learning Paradigms



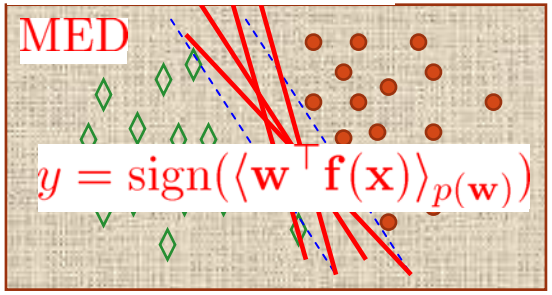
$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{N}\xi_i;$$

$$\text{s.t.} \quad y_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1 - \xi_i, \ \forall i.$$



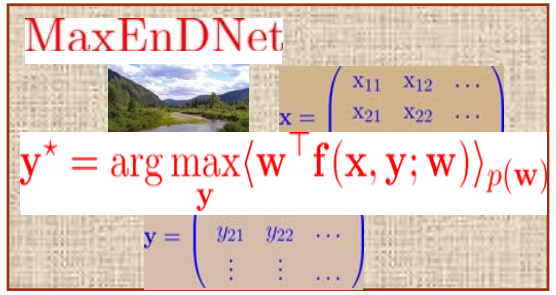
$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.} : \mathbf{w}^\top\Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i \ \ \xi_i \geq 0 \ , \forall i, \forall \mathbf{y} \neq \mathbf{y}^i$$



$$\min_{p,\xi} KL(p||p_0) + C\sum_{i=1}^{N}\xi_i;$$

$$\text{s.t.} \quad y_i\langle \mathbf{f}(\mathbf{x}_i)\rangle_{p(\mathbf{w})} \geq 1 - \xi_i, \ \forall i.$$



$$\min_{p(\mathbf{w}),\xi} \ KL(p||p_0) + U(\xi)$$

$$\text{s.t.} \quad \int p(\mathbf{w})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta\ell_i(\mathbf{y})]\,d\mathbf{w} \geq -\xi_i, \ \xi_i \geq 0,, \ \forall i, \forall \mathbf{y} \neq \mathbf{y}^i.$$