

Maximum Margin Markov Networks and Constraint Generation

Optimization - 10725

Carlos Guestrin

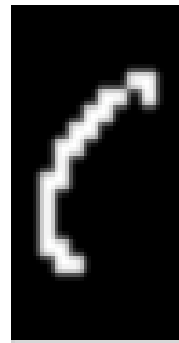
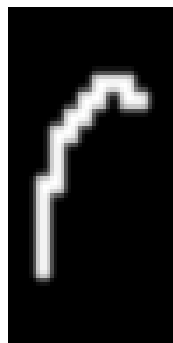
Carnegie Mellon University

February 11th, 2008

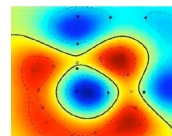
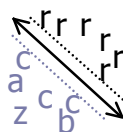
©2008 Carlos Guestrin

1

Handwriting Recognition



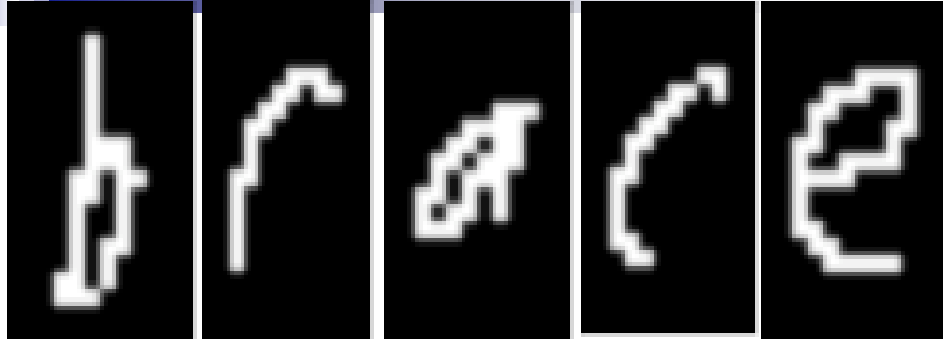
Character recognition: kernel SVMs



©2008 Carlos Guestrin

2

Handwriting Recognition 2



SVMs for sequences?

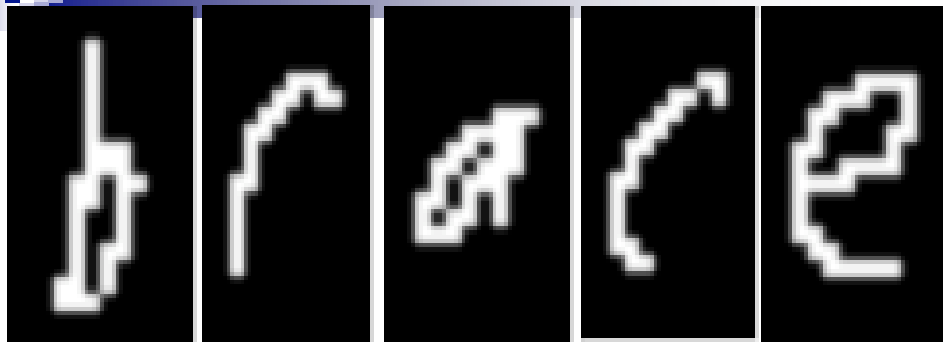
Problem: # of classes exponential in length

brare
zzzzz
aaaaa
.....
brick
brake
brace

©2008 Carlos Guestrin

3

Handwriting Recognition 2



Graphical models: HMMs, MNs

Linear in length



©2008 Carlos Guestrin

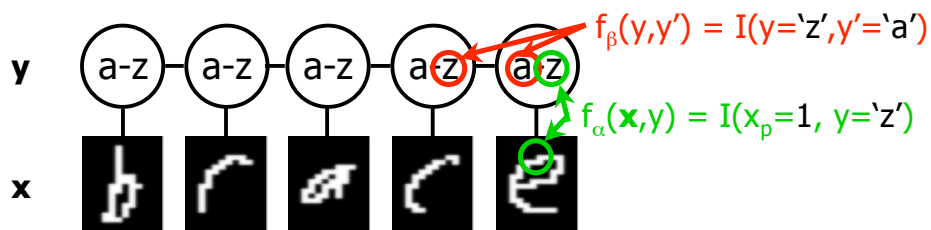
4

Chain Markov Net (aka CRF*)

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_i \phi(\mathbf{x}_i, y_i) \prod_i \phi(y_i, y_{i+1})$$

$$\phi(\mathbf{x}_i, y_i) = \exp\{\sum_{\alpha} w_{\alpha} f_{\alpha}(\mathbf{x}_i, y_i)\}$$

$$\phi(y_i, y_{i+1}) = \exp\{\sum_{\beta} w_{\beta} f_{\beta}(y_i, y_{i+1})\}$$



©2008 Carlos Guestrin

*Lafferty et al. 01

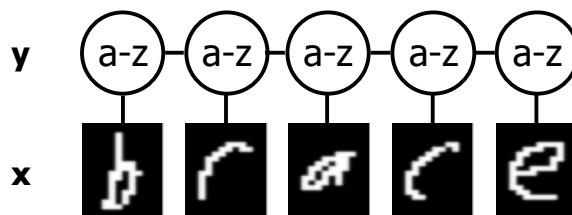
5

CRF - short notation

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_i \phi(\mathbf{x}_i, y_i) \prod_i \phi(y_i, y_{i+1}) = \frac{1}{Z(\mathbf{x})} \exp\{\mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y})\}$$

$$\phi(\mathbf{x}_i, y_i) = \exp\{\sum_{\alpha} w_{\alpha} f_{\alpha}(\mathbf{x}_i, y_i)\}$$

$$\phi(y_i, y_{i+1}) = \exp\{\sum_{\beta} w_{\beta} f_{\beta}(y_i, y_{i+1})\}$$



©2008 Carlos Guestrin

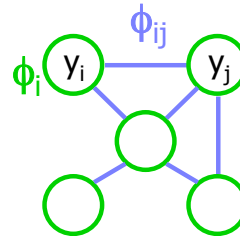
*Lafferty et al. 01

6

Pairwise Markov Nets

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_i \phi_i(\mathbf{x}, y_i) \prod_{ij} \phi_{ij}(\mathbf{x}, y_i, y_j)$$

$$= \frac{1}{Z(\mathbf{x})} \exp\{\mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y})\}$$



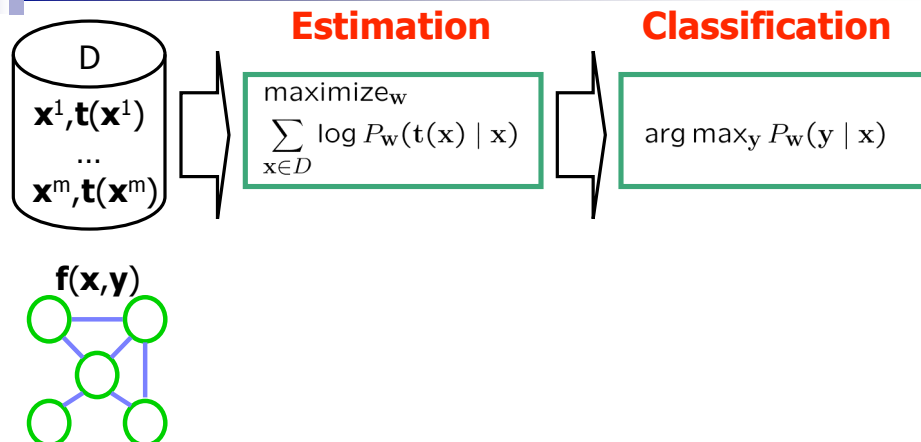
$$\mathbf{w} = [\dots, \mathbf{w}_i, \dots, \mathbf{w}_{ij}, \dots]$$

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = [\dots, \mathbf{f}_i(\mathbf{x}, y_i), \dots, \mathbf{f}_{ij}(\mathbf{x}, y_i, y_j), \dots]$$

©2008 Carlos Guestrin

7

Max (Conditional) Likelihood



©2008 Carlos Guestrin

8

OCR Example

- We want:

$$\operatorname{argmax}_{\text{word}} \mathbf{w}^T \mathbf{f}(\text{brace}, \text{word}) = \text{"brace"}$$

- Equivalently:

$$\mathbf{w}^T \mathbf{f}(\text{brace}, \text{"brace"}) > \mathbf{w}^T \mathbf{f}(\text{brace}, \text{"aaaaa"})$$

$$\mathbf{w}^T \mathbf{f}(\text{brace}, \text{"brace"}) > \mathbf{w}^T \mathbf{f}(\text{brace}, \text{"aaaab"})$$

...

$$\mathbf{w}^T \mathbf{f}(\text{brace}, \text{"brace"}) > \mathbf{w}^T \mathbf{f}(\text{brace}, \text{"zzzzz"})$$

©2008 Carlos Guestrin

9

Max Margin Estimation

- **Goal:** find $\mathbf{w}$ such that

$$\mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{t}(\mathbf{x})) > \mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}) \quad \forall \mathbf{x} \in \mathcal{D} \quad \forall \mathbf{y} \neq \mathbf{t}(\mathbf{x})$$

$$\mathbf{w}^T [\mathbf{f}(\mathbf{x}, \mathbf{t}(\mathbf{x})) - \mathbf{f}(\mathbf{x}, \mathbf{y})] > 0$$

$$\mathbf{w}^T \Delta \mathbf{f}_{\mathbf{x}}(\mathbf{y}) > 0$$

©2008 Carlos Guestrin

10

Not all margins are equal

- Goal: find $\mathbf{w}$ such that

$$\mathbf{w}^\top \Delta \mathbf{f}_x(\mathbf{y}) \geq \gamma \quad \forall \mathbf{x} \in D \quad \forall \mathbf{y} \neq \mathbf{t}(\mathbf{x})$$

- Gain over $\mathbf{y}$ grows with # of mistakes in $\mathbf{y}$: $\Delta \mathbf{t}_x(\mathbf{y})$

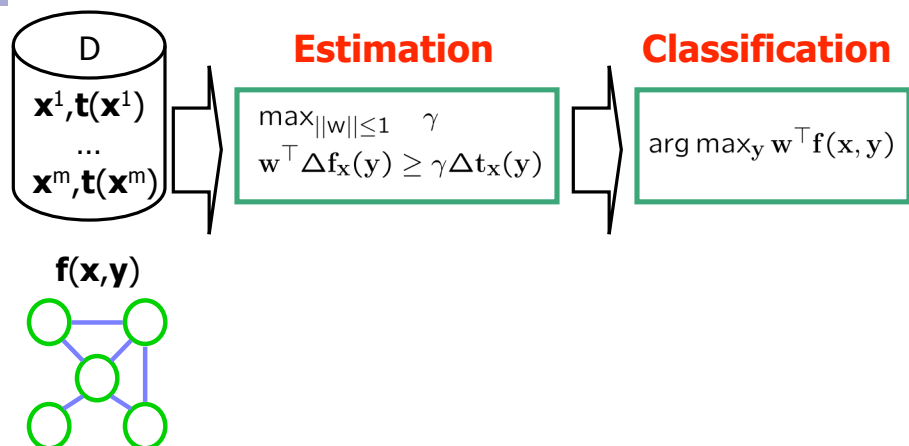
$$\begin{array}{ll} \Delta \mathbf{t}_{\text{brace}}(\text{"craze"}) & \Delta \mathbf{t}_{\text{brace}}(\text{"zzzzz"}) \\ \mathbf{w}^\top \Delta \mathbf{f}_{\text{brace}}(\text{"craze"}) & \mathbf{w}^\top \Delta \mathbf{f}_{\text{brace}}(\text{"zzzzz"}) \end{array}$$

©2008 Carlos Guestrin

11

Maximum Margin Markov Nets

[Taskar, Guestrin, Koller '03]



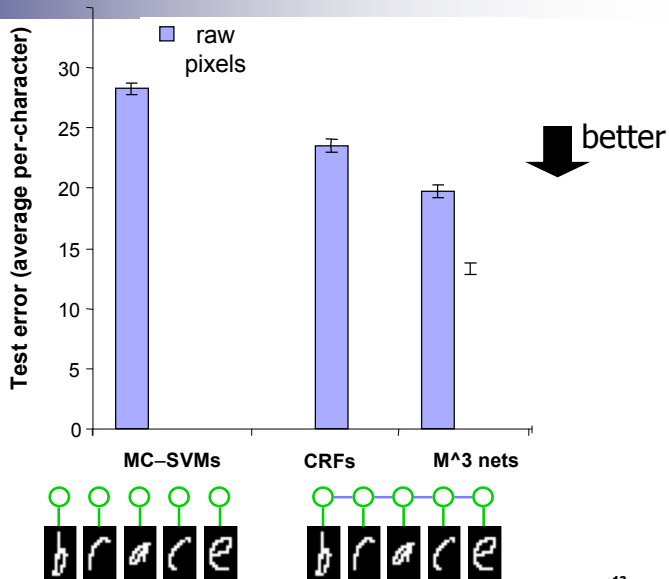
©2008 Carlos Guestrin

12

Handwriting Recognition

Length: ~8 chars
Letter: 16x8 pixels
10-fold Train/Test
5000/50000 letters
600/6000 words

Models:
Multiclass-SVMs*
CRFs
M³ nets

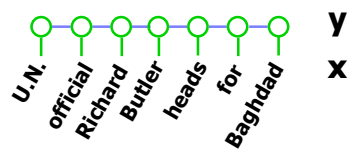


*Crammer & Singer 01

13

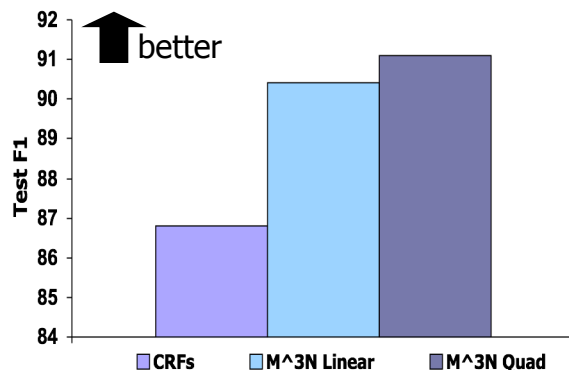
Named Entity Recognition

- Locate and classify named entities in sentences:
 - 4 categories: organization, person, location, misc.
 - e.g. "U.N. official Richard Butler heads for Baghdad".
- CoNLL 03 data set (200K words train, 50K words test)



$y_i = \text{org/per/loc/misc/none}$

$f(y_i, x) = [\dots,$
 $I(y_i=\text{org}, x_i=\text{"U.N."}),$
 $I(y_i=\text{per}, x_i=\text{capitalized}),$
 $I(y_i=\text{loc}, x_i=\text{known city}),$



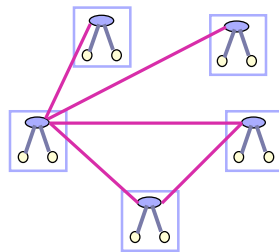
©2008 Carlos Guestrin

14

Hypertext Classification

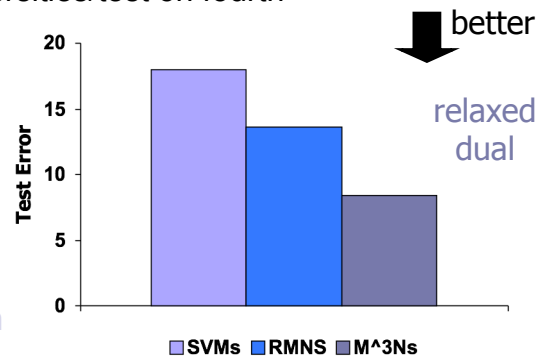
WebKB dataset

- Four CS department websites: 1300 pages/3500 links
- Classify each page: faculty, course, student, project, other
- Train on three universities/test on fourth



loopy belief propagation

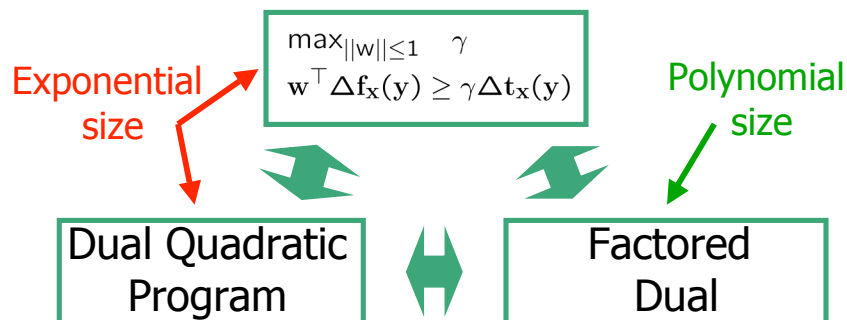
*Taskar et al 02



15

Solving M³Ns

Estimation



©2008 Carlos Guestrin

16