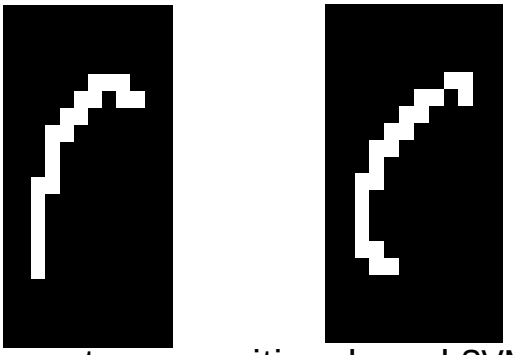
Maximum Margin Markov Networks and Constraint Generation

Optimization - 10725

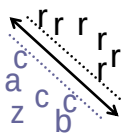
Carlos Guestrin
Carnegie Mellon University

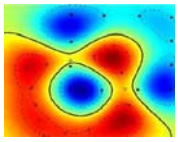
February 11th, 2008

Handwriting Recognition

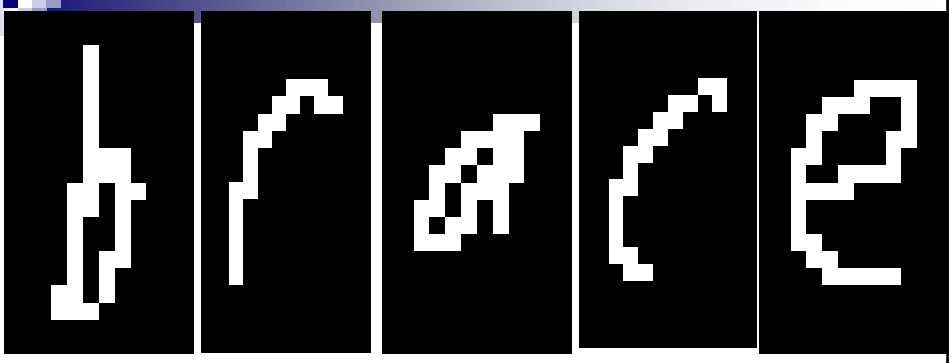


Character recognition: kernel SVMs



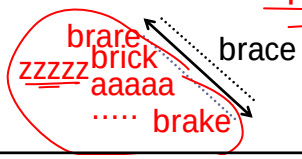


Handwriting Recognition 2

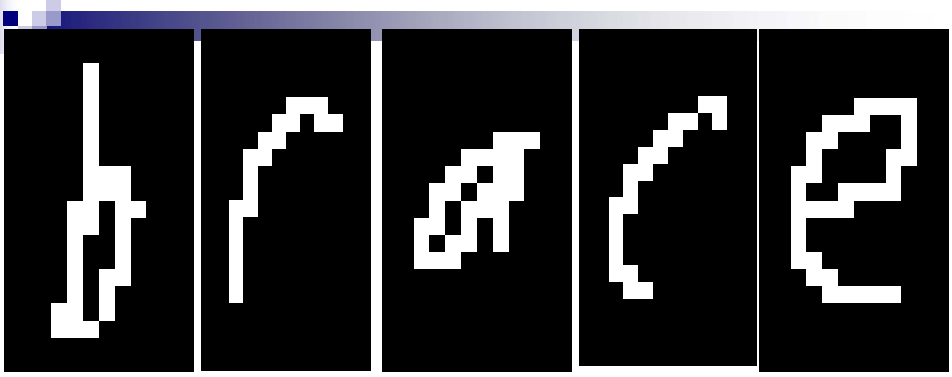


SVMs for sequences?

Problem: # of classes exponential in length

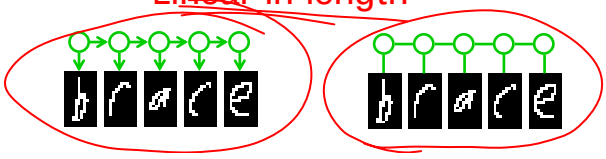


Handwriting Recognition 2



Graphical models: HMMs, MNs

Linear in length

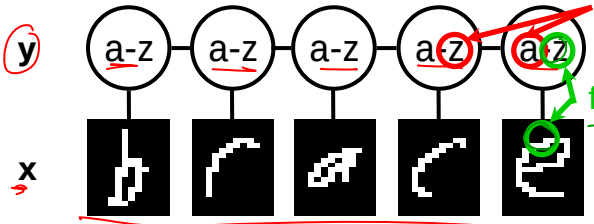


Chain Markov Net (aka CRF*)

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

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

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



$$f_\beta(y,y') = I(y='z',y='a')$$

$$f_\alpha(\mathbf{x},y) = I(x_p=1,\ y='z')$$

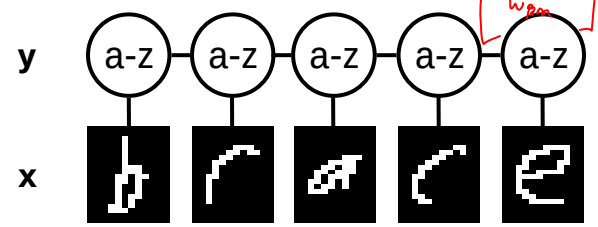
*Lafferty et al. 01

CRF - short notation

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

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

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



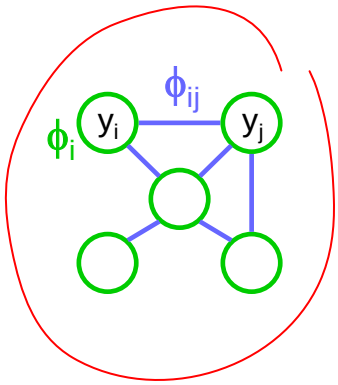
*Lafferty et al. 01

Pairwise Markov Nets

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

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

↑ normalized



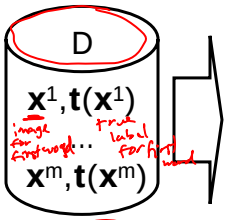
$\mathbf{w} = [\ldots, w_i, \ldots, w_{ij}, \ldots]$

$\mathbf{f}(\mathbf{x},\mathbf{y}) = [\ldots, f_i(\mathbf{x},y_i), \ldots, f_{ij}(\mathbf{x},y_i,y_j), \ldots]$

Max (Conditional) Likelihood

$\arg\max_y f(y) = \arg\max_y \log f(y)$

Estimation **Classification**

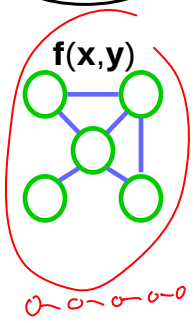


$$\text{maximize}_{\mathbf{w}} \sum_{\mathbf{x} \in D} \log P_{\mathbf{w}}(\mathbf{t}(\mathbf{x}) \mid \mathbf{x})$$

parameters w

test case $\mathbf{x}_t$

$$\arg\max_{\mathbf{y}} P_{\mathbf{w}}(\mathbf{y} \mid \mathbf{x}_t)$$



At classification time

$$\arg\max_y P_w(y|x) = \arg\max_y \frac{1}{Z(x)} e^{w^T f(y,x)}$$

doesn't depend on y, irrelevant for decision

$$= \arg\max_y \log \frac{1}{Z(x)} e^{w^T f(y,x)} = \arg\max_y w^T f(y,x) - \log Z(x)$$

$$= \arg\max_y w^T f(y,x)$$

why do I care about $Z(x)$

OCR Example

- We want:

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

- Equivalently:

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

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

  …

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

Max Margin Estimation

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

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

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

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

Not all margins are equal

- Goal: find **w** such that

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

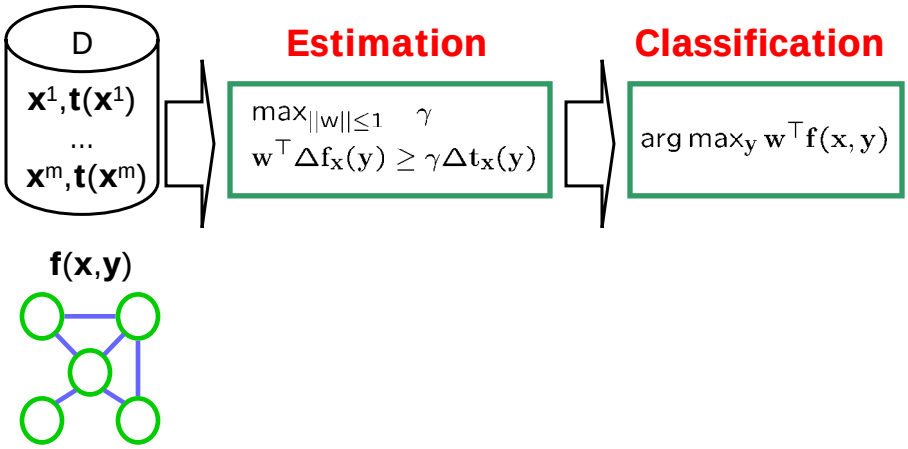
- Gain over **y** grows with # of mistakes in **y**: $\Delta \mathbf{t}_{\mathbf{x}}(\mathbf{y})$

$\Delta \mathbf{t}_{\text{brace}}$("craze") $\Delta \mathbf{t}_{\text{brace}}$("zzzzz")

$\mathbf{w}^\top \Delta \mathbf{f}_{\text{brace}}$("craze") $\mathbf{w}^\top \Delta \mathbf{f}_{\text{brace}}$("zzzzz")

Maximum Margin Markov Nets

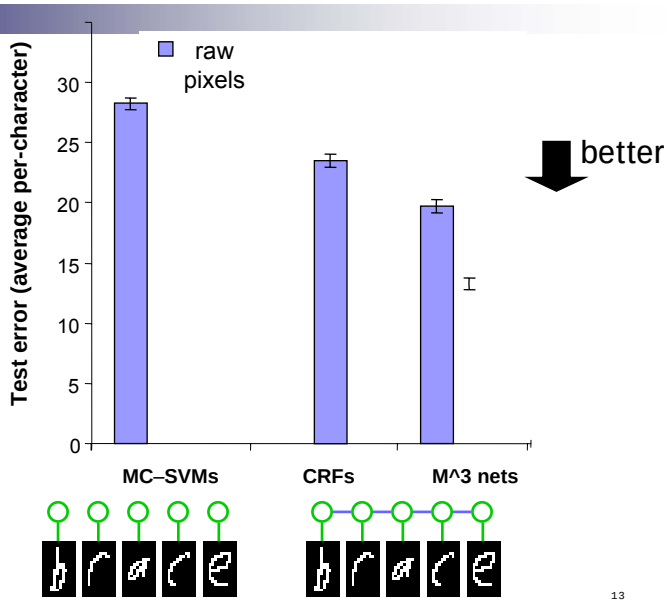
[Taskar, Guestrin, Koller '03]



Handwriting Recognition

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

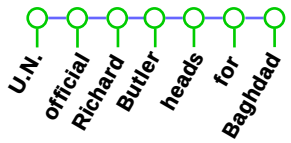
Models:
Multiclass-SVMs*
CRFs
M^3 nets



*Crammer & Singer 01

Named Entity Recognition

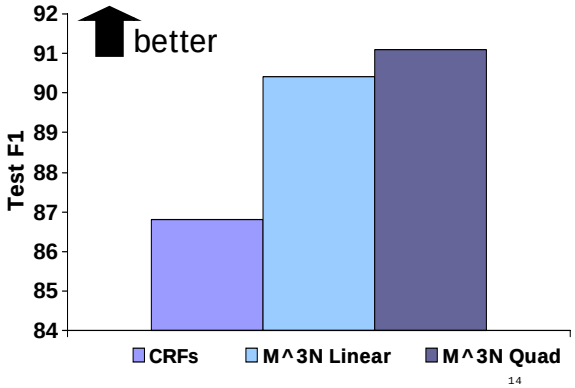
- Locate and classify named entities in sentences:
  - 4 categories: organization, person, location, misc.
  - e.g. "U.N. official Richard Butler heads for Baghdad".
- CoNLL 03 data set (200K words train, 50K words test)



y
x

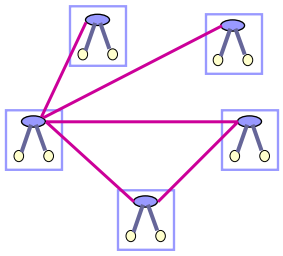
y_i = org/per/loc/misc/none

f(y_i, x) = [...,
I(y_i=org, x_i="U.N."),
I(y_i=per, x_i=capitalized),
I(y_i=loc, x_i=known city),
...]

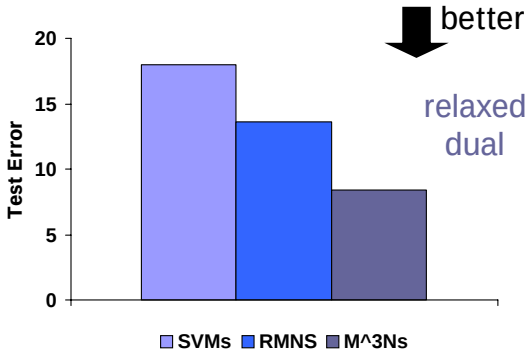


Hypertext Classification

- 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

Solving M^{3}Ns

Estimation

$$\max_{||w||\leq 1} \gamma$$
$$w^\top \Delta f_x(y) \geq \gamma \Delta t_x(y)$$

Exponential size

Polynomial size

Dual Quadratic Program

Factored Dual