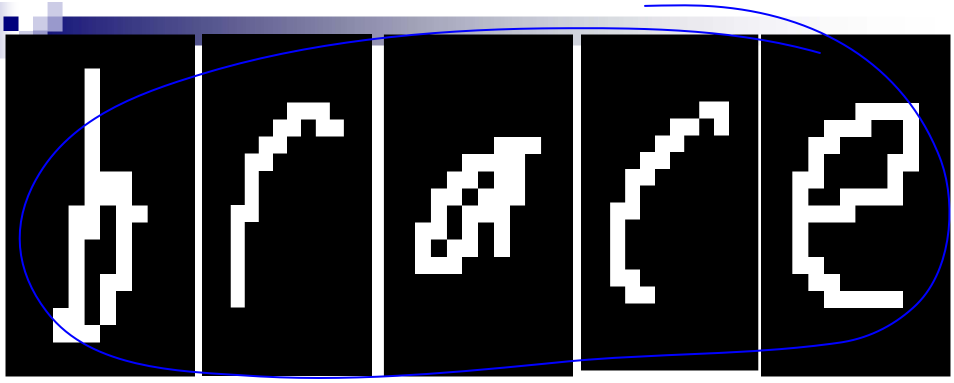
Maximum Margin Markov Networks and Constraint Generation continued

Optimization - 10725
Carlos Guestrin
Carnegie Mellon University

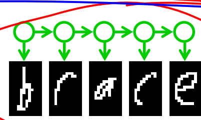
February 18th, 2008

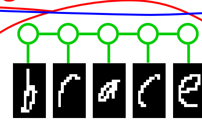
Handwriting Recognition 2



Graphical models: HMMs, MNs

Linear in length





An example of a CRF — a Conditional random field

Chain Markov Net (aka CRF*)

potentials

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

normalization (partition function) — position, pixels — want — similar to logistic regression

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

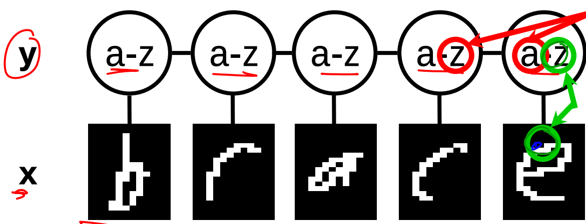
want

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

neighboring letters

letter is z — following letter is a

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



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

indexes pixel coordinate, letter, position — pixel coordinate — pixel on — when letter is z

*Lafferty et al. 01

$e^{w_1 f_1} \cdot e^{w_2 f_2} = e^{w_1 f_1 + w_2 f_2}$

CRF - short notation

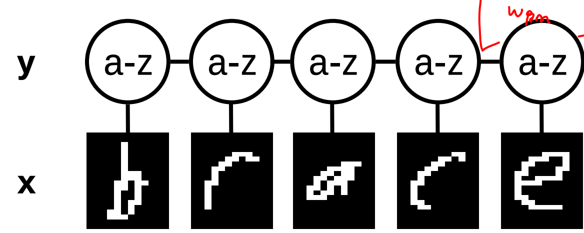
label — images

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

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

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

$$W = \begin{bmatrix} w_{\alpha_1} \\ \vdots \\ w_{\alpha_n} \\ w_{\beta_1} \\ \vdots \\ w_{\beta_m} \end{bmatrix} \qquad f(x,y) = \begin{bmatrix} f_{\alpha_1}(x_1,y_1) \\ \vdots \\ f_{\alpha_n}(x_n,y_n) \\ f_{\beta_1}(y_1,y_2) \\ \vdots \\ f_{\beta_m}(y_{n-1},y_n) \end{bmatrix}$$



*Lafferty et al. 01

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

Max (Conditional) Likelihood

Estimation

D: $\mathbf{x}^1, \mathbf{t}(\mathbf{x}^1)$ … $\mathbf{x}^m, \mathbf{t}(\mathbf{x}^m)$

image for first word; true label for first word

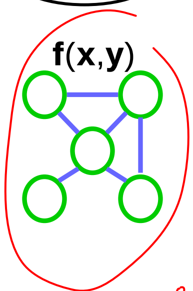
$$\text{maximize}_{\mathbf{w}} \sum_{\mathbf{x} \in D} \log P_{\mathbf{w}}(\mathbf{t}(\mathbf{x}) \mid \mathbf{x})$$

parameters w

Classification

test case x_t

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



At classification time

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

doesn't depend y, irrelevant for decision

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

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

why do I care about $Z(x)$

OCR Example

- We want:

  given input

  $\arg\max_{\text{word}}$ $\mathbf{w}^T \mathbf{f}($  $\textbf{word}) =$ **"brace"**

  true label wins over all other possibilities

- Equivalently:

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

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

  …

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

number of constraints is exponential in length of word

strictly greater than?

how???

Max Margin Estimation

- Goal: find **w** such that

$$\mathbf{w}^T\mathbf{f}(\mathbf{x},\overset{\text{true label}}{\mathbf{t}(\mathbf{x})}) > \mathbf{w}^T\mathbf{f}(\mathbf{x},\overset{\text{other labels}}{\mathbf{y}}) \qquad \forall \mathbf{x}\in D \quad \forall \mathbf{y}\neq \mathbf{t}(\mathbf{x})$$

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

$$\mathbf{w}^\top \Delta \mathbf{f}_{\mathbf{x}}(\mathbf{y}) > 0 \qquad \forall x \in D, \ \forall y \neq t(x)$$

today: "linearly separable"

can add slack vars just like SVMs

$$\max_{w,\gamma} \gamma$$

$$w^T \Delta f_x(y) \geq \gamma \leftarrow \text{"margin"}$$

$$\forall x \in D, \ \forall y \neq t(x)$$

$$\|w\|_2^2 \leq 1$$

"Solution 0"

Not all margins are equal

- Goal: find **w** such that

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

if $y = t(x) \Rightarrow$ rhs: $\gamma \cdot \Delta t_x(y) = 0$

lhs: $w^T \Delta f_{x,y} = w^T[f(x,t(x)) - f(x,y)] = 0$

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

$\Delta \mathbf{t}_{\text{brace}}$(**"craze"**) $= 2$ $\qquad$ $\Delta \mathbf{t}_{\text{brace}}$(**"zzzzz"**) $= 5$

$\mathbf{w}^\top \Delta \mathbf{f}_{\text{brace}}$(**"craze"**) $\geq 2\gamma$ $\qquad$ $\mathbf{w}^\top \Delta \mathbf{f}_{\text{brace}}$(**"zzzzz"**) $\geq 5\gamma$

Maximum Margin Markov Nets

[Taskar, Guestrin, Koller '03]

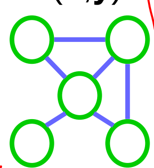
D

$\mathbf{x}^1, \mathbf{t}(\mathbf{x}^1)$

…

$\mathbf{x}^m, \mathbf{t}(\mathbf{x}^m)$

f(x,y)



Estimation

$\max_{||w|| \le 1} \quad \gamma$

$\mathbf{w}^\top \Delta \mathbf{f}_\mathbf{x}(\mathbf{y}) \geq \gamma \Delta \mathbf{t}_\mathbf{x}(\mathbf{y})$

exists a transformation

QP

$\min \|w\|_2^2 \leftarrow w^T w$

$w^T \Delta f_x(y) \geq \Delta t_x(y)$

$\forall x \in D, \forall y$

Classification

no Δ:

$\arg\max_\mathbf{y} \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})$

e.g., in chains it's called Viterbi

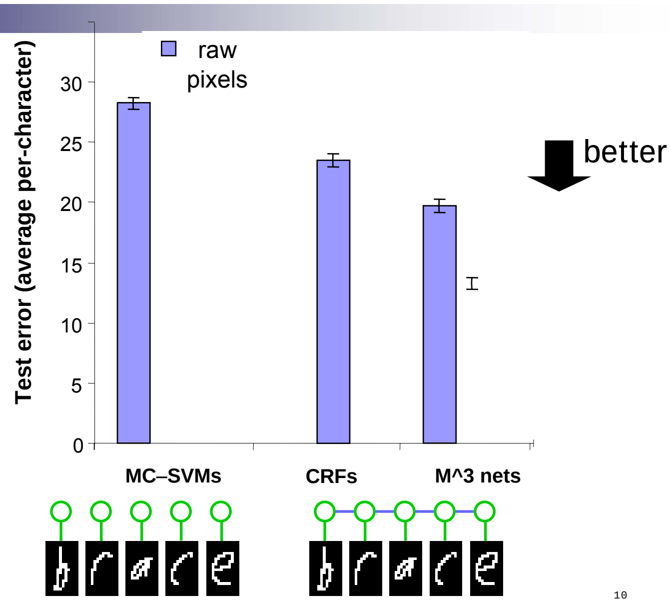
exponential in length of word!! :(

- BTW. Just like SVMs, there are "non-linearly separable" cases, must add slack variables…
  - see store for details

Handwriting Recognition

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

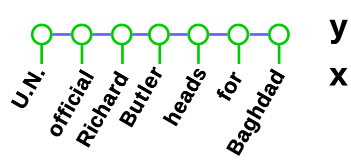
Models:
Multiclass-SVMs*
CRFs
M³ nets



*Crammer & Singer 01

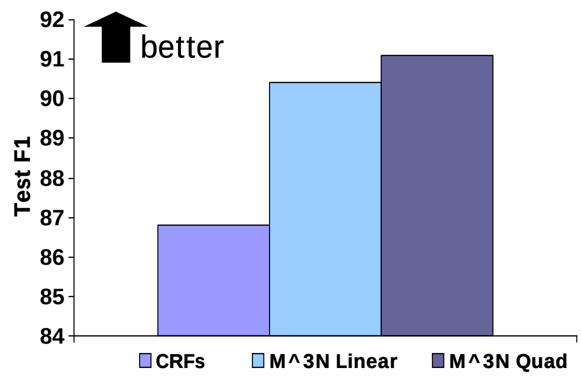
Named Entity Recognition

- 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)



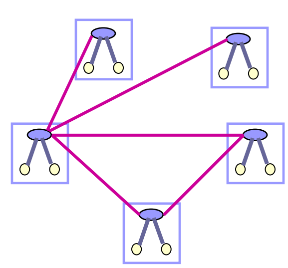
y_i = org/per/loc/misc/none

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

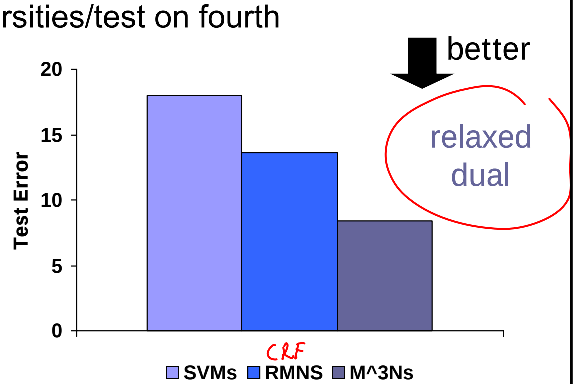


Hypertext Classification

- 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

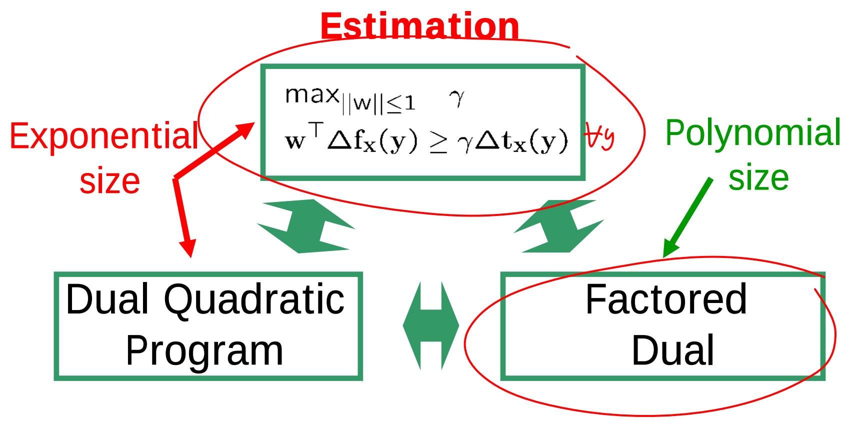


loopy belief propagation



*Taskar et al 02

Solving M³Ns [Taskar, Guestrin, Koller '03]



Other ways to solve M³Ns

- Sequential minimal optimization (SMO) [Taskar, Guestrin, Koller '03]
- Exponentiated gradient [Bartlett, Collins, Taskar, McAllester '04]
- Subgradient method [Ratliff, Bagnell, Zinkevich '07]
- …
- Today
  - Simple constraint generation
    - (Other methods will perform better in many practical problems)
    - (Other methods are better suited to adding kernels)
    - (Other methods use similar principles to simpler constraint generation)

Constraint generation overview

- Minimize $\mathbf{w}^2$ $= w^T w$
  - Subject to:

$$\mathbf{w}^T \mathbf{f}(\ \text{"brace"}) \geq \mathbf{w}^T \mathbf{f}(\ \text{"aaaaa"}) + \overbrace{\Delta(\text{brace,aaaaa})}^{4}$$

$$\ldots \quad \overbrace{}^{5}$$

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

$\Big\}\ \forall y$

- General form: $\min w^2$
  $w^T f(x, t(x)) \geq w^T f(x,y) + \Delta t_x(y) \quad \forall y \in \mathcal{Y}_x \quad \forall x$
- Subset of constraints: $\Omega_x \subset \mathcal{Y}_x$
  $\min w^2$
  $w^T f(x, t(x)) \geq w^T f(x,y) + \Delta t_x(y) \quad \forall x \quad \forall y \in \Omega_x$
- Constraint generation:
  - Solve for w for a given set of Ω_x
  - find a violated constraint: for each x:
    $w^T f(x, t(x)) \overset{?}{\geq} \underbrace{\max_y \left[w^T f(x,y) + \Delta t_x(y) \right]}_{\text{separation oracle}}$

Generating a constraint (simpler setting)

- Form of constraint
  $\mathbf{w}^T \mathbf{f}(\ \text{"brace"}) \geq \mathbf{w}^T \mathbf{f}(\ \text{aaaaa"}) + \Delta(\text{brace,aaaaa})$ $\forall y$
- Another way of expressing: $w^T f(x, t(x)) \geq \max_y \left[w^T f(x,y) + \Delta t_x(y) \right]$
- Given w, are any constraints violated?
- Separation oracle question: $\arg\max_y \left[w^T f(x,y) + \Delta t_x(y) \right]$
- **Δ(brace,aaaaa)** seems hard, simpler question:
  $\arg\max_y w^T f(x,y)$
- Exponentially many possibilities…
  same ~~??~~ question as at classification (testing) time
  - chains viterbi
  - more generally! inference for GM, e.g., Variable elimination

Generating a constraint with hamming margin part (Δ(brace,aaaaa))

$e^{\sum_i f_i} = \prod_i e^{f_i}$

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

- Without **Δ(brace,aaaaa)** ➔ standard (MAP or MPE) inference in graphical models
  - Solve with dynamic programming
  - For chains, it's called the Viterbi algorithm
- What do we do about **Δ(brace,aaaaa)** ?

$\Delta(\text{brace}, \text{aaaaa}) = 4$
$= \Delta(b,a) + \Delta(r,a) + \Delta(a,a) + \Delta(c,a) + \Delta(e,a)$
$= 4$

$\Delta t_x(y) = \sum_i \Delta t_{x_i}(y_i)$

- Reformulation:

$\Pi_i\, \phi(x_i, y_i)\, \Pi_i\, \phi(y_i, y_{i+1}) \cdot e^{\sum_i \Delta t_{x_i}(y_i)}$

$= \Pi_i\, \underbrace{\phi(x_i, y_i) \cdot e^{\Delta t_{x_i}(y_i)}}_{\phi_i'(x_i, y_i)} \cdot \Pi_i\, \phi(y_i, y_{i+1}) = e^{\overbrace{w^T f(x,y) + \Delta t_x(y)}^{\text{wanted to argmax}}}$

- Same inference algorithm!!!
  - (slightly different potentials)

Overview of constraint generation for M³Ns

- Problem we want to solve:

$\min \|w\| = w^T w$
$w^T f(x, t(x)) \geq w^T f(x,y) + \Delta t_x(y)$
$\forall x \in D, \forall y \in \mathcal{Y}_x$

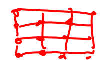
- Maintain subset of "runner up labels" for each training example:

$\Omega_x \subset \mathcal{Y}_x$

- Obtain some value for weights **w**
- Separation oracle:
  - Reformulate model to include hamming margin **Δ(brace,aaaaa)**
  - Dynamic programming (inference in graphical models)
  - Apply to each data point

Some reasons M^3Ns are cool… :)

- Often perform better
- Can use kernels easily, and get sparsity like SVMs
- Can be learned exactly in many problems where CRFs require approximate inference techniques
  - E.g., image segmentation (graph cuts)



- Can be generalized to other optimization problems
  - E.g., [Taskar, Chatalbashev, Koller, Guestrin '05]
  - Matching problems
  - Paths
  - Pretty much any optimization technique in the inner loop