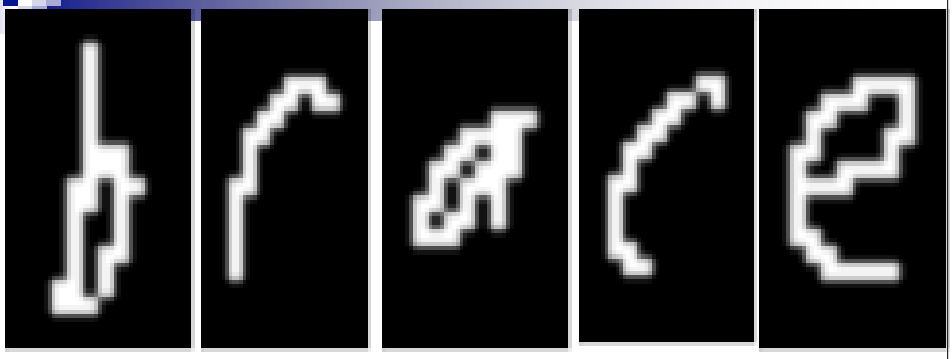
Maximum Margin Markov Networks and Constraint Generation continued

Optimization - 10725

Carlos Guestrin
Carnegie Mellon University

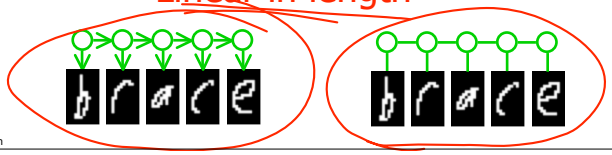
February 18th, 2008

Handwriting Recognition 2



Graphical models: HMMs, MNs

Linear in length

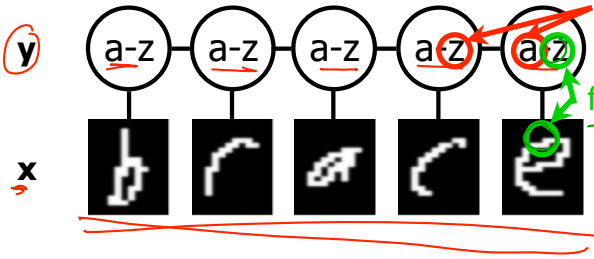


Chain Markov Net (aka CRF*)

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

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

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



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

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

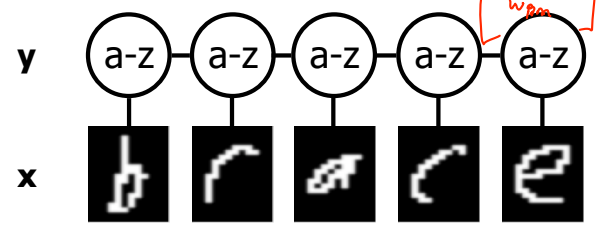
*Lafferty et al. 01

CRF - short notation

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

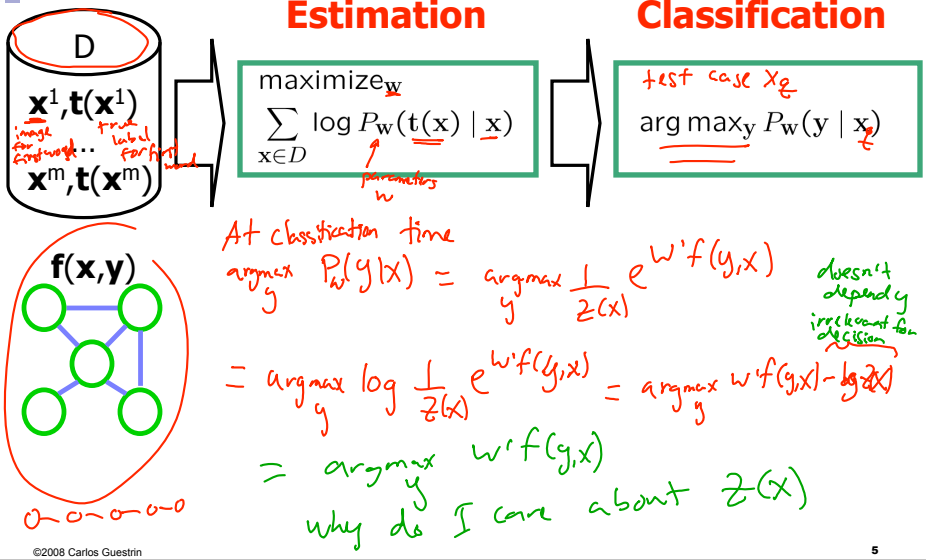
$$\phi(\mathbf{x}_i, y_i) = \exp\{\sum_\alpha w_\alpha f_\alpha(\mathbf{x}_i, y_i)\}$$

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



*Lafferty et al. 01

Max (Conditional) Likelihood



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

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:

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

- Equivalently:

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

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

  …

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

number of constraints is exponential in length of word

strictly greater than?

how w???

Max Margin Estimation

- Goal: find **w** such that

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

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

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

Not all margins are equal

- Goal: find **w** such that

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

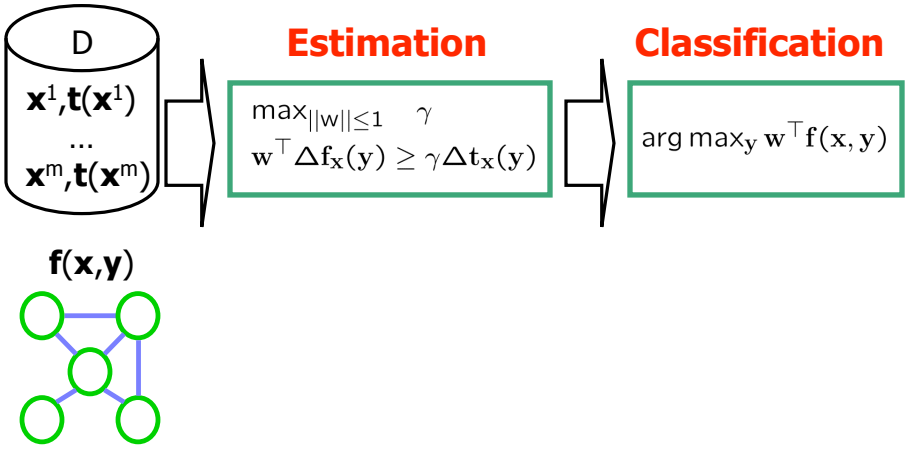
- Gain over **y** grows with # of mistakes in **y**: $\Delta\mathbf{t}_{\mathbf{x}}(\mathbf{y})$

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

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

Maximum Margin Markov Nets

[Taskar, Guestrin, Koller '03]

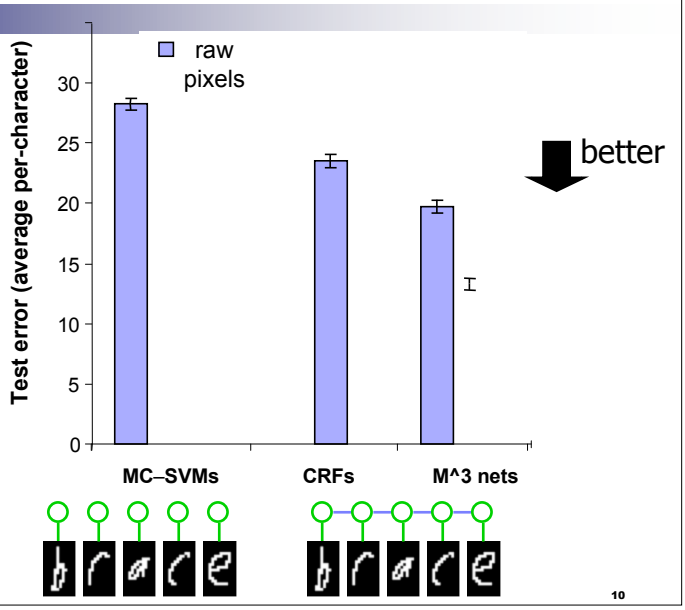


- BTW. Just like SVMs, there are "non-linearly separable" cases, must add slack variables…
- see store for details

Handwriting Recognition

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

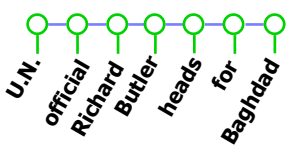
Models:
- Multiclass-SVMs*
- CRFs
- M^3 nets



*Crammer & Singer 01

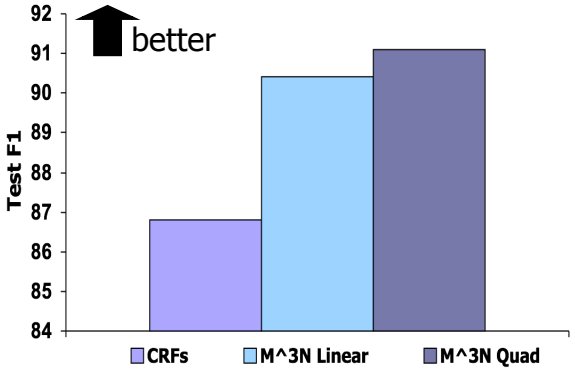
Named Entity Recognition

- 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)



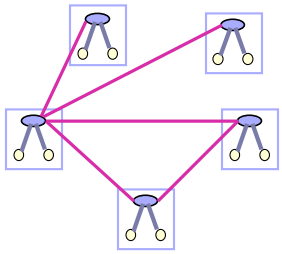
y_i = org/per/loc/misc/none

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),
...]



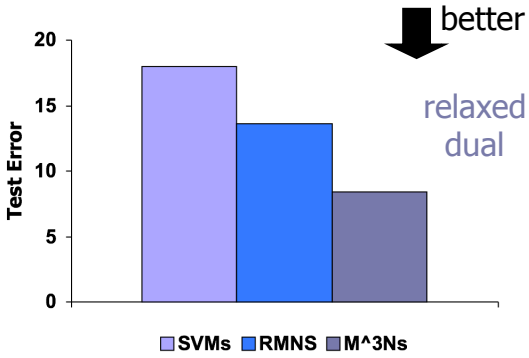
Hypertext Classification

- 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

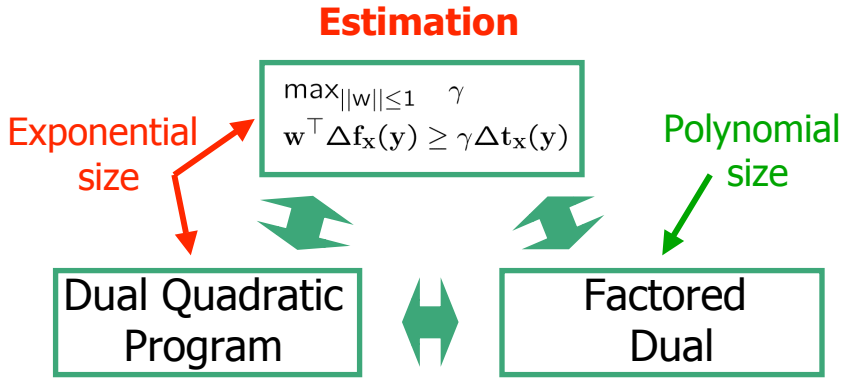


loopy belief propagation

*Taskar et al 02



Solving M^{3}Ns [Taskar, Guestrin, Koller '03]



Other ways to solve M^{3}Ns

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

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

…

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

- General form:
- Subset of constraints:
- Constraint generation:
  -
  -

Generating a constraint (simpler setting)

- Form of constraint

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

- Another way of expressing:
- Given w, are any constraints violated?
- Separation oracle question:
- $\Delta(\text{brace,aaaaa})$ seems hard, simpler question:
- Exponentially many possibilities…

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

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

- 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) ?
- Reformulation:
- Same inference algorithm!!!
  - (slightly different potentials)

Overview of constraint generation for M^{3}Ns

- Problem we want to solve:
- Maintain subset of "runner up labels" for each training example:
- 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
- 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