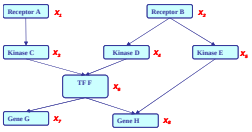
School of Computer Science
Carnegie Mellon

Large Margin Structured Models

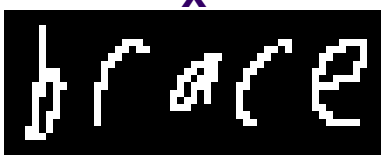
Probabilistic Graphical Models (10-708)

Lecture # 21
11/28/2007
Eric Xing



OCR example

x



y

brace

Sequential structure

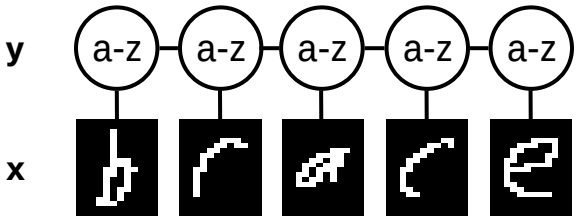
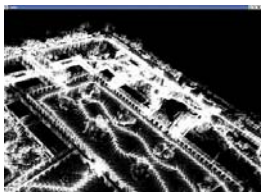
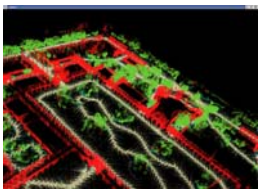


Image segmentation example

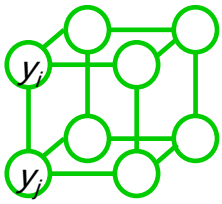
x



y



Spatial structure



Structured models

$$h(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} s(\mathbf{x},\mathbf{y})$$ ← scoring function

space of feasible outputs

Assumptions:

$$score(\mathbf{x},\mathbf{y}) = \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}) = \sum_p \mathbf{w}^\top \mathbf{f}(\mathbf{x}_p,\mathbf{y}_p)$$

linear combination of features

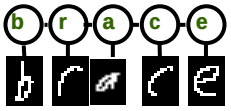
sum of part scores:
- index *p* represents a part in the structure

Learning w

- Training examples $(\mathbf{x}_i, \mathbf{y}_i)$
- Probabilistic approach:

$$P_{\mathbf{w}}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z_{\mathbf{w}}(\mathbf{x})} \exp\{\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})\}$$

- Computing $Z_w(x)$ can be NP-complete
  - Tractable models but intractable estimation
- Large margin approach:
  - Exact and efficient when prediction is tractable



Learning Strategy

- Recall that in CRF
  - We predict based on:

$$y^* \mid x = \arg\max_y p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{\sum_c \theta_c f_c(x, y_c)\right\}$$

  - And we learn based on:

$$\theta_c^* \mid \{y_n, x_n\} = \arg\max_{\theta_c} \prod_n p_\theta(y_n \mid x_n) = \prod_n \frac{1}{Z(\theta, x_n)} \exp\left\{\sum_c \theta_c f_c(x_n, y_{n,c})\right\}$$

- MaxMargin:
  - We predict based on:

$$y^* \mid x = \arg\max_y \sum_c \theta_c f_c(x, y_c) = \arg\max_y w^T F(x, y)$$

  - And we learn based on:

$$w^* \mid \{y_n, x_n\} = \arg\max_w \left(\max_{y'_n \neq y_n, \forall n} w^T \left(F(y_n, x_n) - F(y'_n, x_n)\right)\right)$$

MLE of Feature Based UGMs

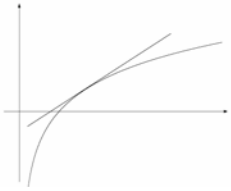
- Scaled likelihood function

$$\tilde{\ell}(\theta;\mathcal{D}) = \ell(\theta;\mathcal{D})/N = \frac{1}{N}\sum_n \log p(x_n \mid \theta)$$
$$= \sum_x \tilde{p}(x)\log p(x \mid \theta)$$
$$= \sum_x \tilde{p}(x)\sum_i \theta_i f_i(x) - \log Z(\theta)$$

- Instead of optimizing this objective directly, we attack its lower bound
  - The logarithm has a linear upper bound …

$$\log Z(\theta) \le \mu Z(\theta) - \log \mu - 1$$

  - This bound holds for all μ, in particular, for $\mu = Z^{-1}(\theta^{(t)})$



  - Thus we have

$$\tilde{\ell}(\theta;\mathcal{D}) \ge \sum_x \tilde{p}(x)\sum_i \theta_i f_i(x) - \frac{Z(\theta)}{Z(\theta^{(t)})} - \log Z(\theta^{(t)}) + 1$$

Generalized Iterative Scaling (GIS)

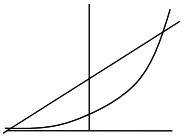
- Lower bound of scaled loglikelihood

$$\tilde{\ell}(\theta;\mathcal{D}) \ge \sum_x \tilde{p}(x)\sum_i \theta_i f_i(x) - \frac{Z(\theta)}{Z(\theta^{(t)})} - \log Z(\theta^{(t)}) + 1$$

- Define $\Delta\theta_i^{(t)} \overset{\text{def}}{=} \theta_i - \theta_i^{(t)}$

$$\tilde{\ell}(\theta;\mathcal{D}) \ge \sum_x \tilde{p}(x)\sum_i \theta_i f_i(x) - \frac{1}{Z(\theta^{(t)})}\sum_x \exp\left\{\sum_i \theta_i f_i(x)\right\} - \log Z(\theta^{(t)}) + 1$$
$$= \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \frac{1}{Z(\theta^{(t)})}\sum_x \exp\left\{\sum_i \theta_i^{(t)} f_i(x)\right\}\exp\left\{\sum_i \Delta\theta_i^{(t)} f_i(x)\right\} - \log Z(\theta^{(t)}) + 1$$
$$= \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \sum_x p(x \mid \theta^{(t)})\exp\left\{\sum_i \Delta\theta_i^{(t)} f_i(x)\right\} - \log Z(\theta^{(t)}) + 1$$

- Relax again
  - Assume $f_i(x) \ge 0, \quad \sum_i f_i(x) = 1$
  - Convexity of exponential: $\exp\left(\sum_i \pi_i x_i\right) \le \sum_i \pi_i \exp(x_i)$



- We have:

$$\tilde{\ell}(\theta;\mathcal{D}) \ge \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \sum_x p(x \mid \theta^{(t)})\sum_i f_i(x)\exp\left(\Delta\theta_i^{(t)}\right) - \log Z(\theta^{(t)}) + 1 \overset{\text{def}}{=} \Lambda(\theta)$$

GIS

- Lower bound of scaled loglikelihood

$$\tilde{\ell}(\theta;D) \geq \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \sum_x p(x \mid \theta^{(t)}) \sum_i f_i(x) \exp\left(\Delta\theta_i^{(t)}\right) - \log Z(\theta^{(t)}) + 1 \overset{\text{def}}{=} \Lambda(\theta)$$

- Take derivative: $\frac{\partial \Lambda}{\partial \theta_i} = \sum_x \tilde{p}(x) f_i(x) - \exp\left(\Delta\theta_i^{(t)}\right) \sum_x p(x \mid \theta^{(t)}) f_i(x)$
- Set to zero

$$e^{\Delta\theta_i^{(t)}} = \frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p(x \mid \theta^{(t)}) f_i(x)} = \frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} Z(\theta^{(t)})$$

  - where $p^{(t)}(x)$ is the unnormalized version of $p(x|\theta^{(t)})$

- Update

$$\theta_i^{(t+1)} = \theta_i^{(t)} + \Delta\theta_i^{(t)} \Rightarrow p^{(t+1)}(x) = p^{(t)}(x) e^{\Delta\theta_i^{(t)} f_i(x)}$$

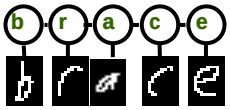
$$p^{(t+1)}(x) = \frac{p^{(t)}(x)}{Z(\theta^{(t)})} \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} Z(\theta^{(t)}) \right)^{f_i(x)}$$

$$\Rightarrow = \frac{p^{(t)}(x)}{Z(\theta^{(t)})} \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)^{f_i(x)} \left(Z(\theta^{(t)}) \right)^{\sum_i f_i(x)}$$

$$= p^{(t)}(x) \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)^{f_i(x)}$$

Learning w

- Training examples $(\mathbf{x}_i, \mathbf{y}_i)$



- Probabilistic approach:

$$P_{\mathbf{w}}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z_{\mathbf{w}}(\mathbf{x})} \exp\{\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})\}$$

- Computing $Z_w(x)$ can be NP-complete
  - Tractable models but intractable estimation
- Large margin approach:
  - Exact and efficient when prediction is tractable

OCR Example

- We want:

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

- Equivalently:

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

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

  …

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

a lot!

Large Margin Estimation

- Given training example (**x**, **y***), we want:

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

$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) > \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) \quad \forall \mathbf{y} \neq \mathbf{y}^*$$

$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \gamma \ell(\mathbf{y}^*, \mathbf{y}) \quad \forall \mathbf{y}$$

- Maximize margin γ
- Mistake weighted margin: $\gamma \ell(\mathbf{y}^*, \mathbf{y})$

$\ell(\mathbf{y}^*, \mathbf{y}) = \sum_i I(y_i^* \neq y_i)$ # of mistakes in **y**

*Taskar et al. 03

Large Margin Estimation

- Recall from SVMs:
  - Maximizing margin γ is equivalent to minimizing the square of the L2-norm of the weight vector **w**:
- New objective function:

$$\min_{\mathbf{w}} \quad \frac{1}{2}||\mathbf{w}||^2$$
$$s.t.\ \mathbf{w}^\top \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}_i, \mathbf{y}'_i) + \ell(\mathbf{y}_i, \mathbf{y}'_i), \quad \forall i, \mathbf{y}'_i \in \mathcal{Y}_i$$

Min-max Formulation

- Brute force enumeration of constraints:

$$\min \quad \frac{1}{2}||\mathbf{w}||^2$$
$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}^*, \mathbf{y}), \quad \forall \mathbf{y}$$

  - The constraints are exponential in the size of the structure
- Alternative: min-max formulation
  - add only the most violated constraint

$$\mathbf{y}' = \arg\max_{\mathbf{y} \neq \mathbf{y}^*} [\mathbf{w}^\top \mathbf{f}(\mathbf{x}^i, \mathbf{y}) + \ell(\mathbf{y}^i, \mathbf{y})]$$
$$\text{add to QP}: \ \mathbf{w}^\top \mathbf{f}(\mathbf{x}^i, \mathbf{y}^i) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}^i, \mathbf{y}') + \ell(\mathbf{y}^i, \mathbf{y}')$$

  - Handles more general loss functions
  - Only polynomial # of constraints needed

Min-max formulation

$$\min \quad \frac{1}{2}||\mathbf{w}||^2$$

$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \max_{\mathbf{y} \neq \mathbf{y}^*} \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}^*, \mathbf{y})$$

- Key step: convert the maximization in the constraint from discrete to continuous
  - This enables us to plug it into a QP

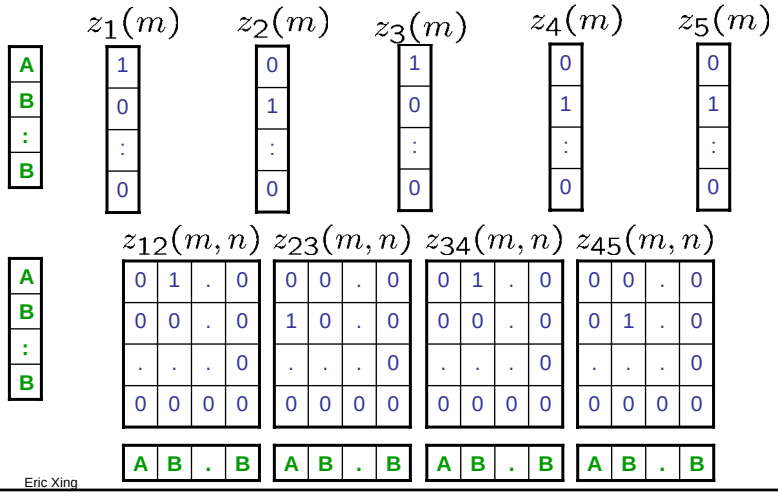
$$\max_{\mathbf{y} \neq \mathbf{y}^*} \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}^*, \mathbf{y}) \quad \Longleftrightarrow \quad \max_{\mathbf{z} \in \mathcal{Z}} \; (\mathbf{F}^\top \mathbf{w} + \ell)^\top \mathbf{z}$$

discrete optim. continuous optim.

- How to do this conversion?
  - Linear chain example in the next slides →

y ⇒ z map for linear chain structures

OCR example: y = 'ABABB';
z's are the indicator variables for the corresponding classes (alphabet)



y ⇒ z map for linear chain structures

Rewriting the maximization function in terms of indicator variables:

$$\max_{\mathbf{z}} \sum_{j,m} z_j(m)\left[\mathbf{w}^\top \mathbf{f}_{\text{node}}(\mathbf{x}_j, m) + \ell_j(m)\right] + \sum_{jk,m,n} z_{jk}(m,n)\left[\mathbf{w}^\top \mathbf{f}_{\text{edge}}(\mathbf{x}_{jk}, m, n) + \ell_{jk}(m,n)\right]$$

$$= (\mathbf{F}^\top \mathbf{w} + \ell)^\top \mathbf{z}$$

$z_k(n)$

0	1	0	0

$z_j(m)$

0
0
1
0

$z_{jk}(m,n)$

0	0	0	0
0	0	0	0
0	1	0	0
0	0	0	0

$$z_j(m) \geq 0;\ z_{jk}(m,n) \geq 0;$$

normalization $\sum_m z_j(m) = 1$

agreement $\sum_n z_{jk}(m,n) = z_j(m)$

$$\mathbf{Az} = \mathbf{b}$$

$$\max_{\mathbf{Az}=\mathbf{b}} (\mathbf{F}^\top \mathbf{w} + \ell)^\top \mathbf{z}$$

Min-max formulation

- Original problem:

$$\min \quad \frac{1}{2}||\mathbf{w}||^2$$

$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \max_{\mathbf{y}} \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}^*, \mathbf{y})$$

- Transformed problem:

$$\min \quad \frac{1}{2}||\mathbf{w}||^2$$

$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \max_{\substack{\mathbf{z}\geq 0;\\ \mathbf{Az}=\mathbf{b};}} \mathbf{q}^\top \mathbf{z} \quad \text{where } \mathbf{q}^\top = \mathbf{w}^\top \mathbf{F} + \ell^\top$$

  - Has integral solutions **z** for chains, trees
  - Can be fractional for untriangulated networks

Min-max formulation

- Using strong Lagrangian duality:
  (beyond the scope of this lecture)

$$\max_{\substack{\mathbf{z}\geq 0;\\ \mathbf{Az}=\mathbf{b};}} \mathbf{q}^\top \mathbf{z} \;=\; \min_{\mathbf{A}^\top \mu \geq \mathbf{q}} \mathbf{b}^\top \mu$$

- Use the result above to minimize jointly over **w** and **μ**:

$$\begin{aligned} &\min_{\mathbf{w},\mu} \; \frac{1}{2}||\mathbf{w}||^2 \\ &\text{s.t.} \;\; \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}^*) \geq \mathbf{b}^\top \mu; \\ &\qquad \mathbf{A}^\top \mu \geq \mathbf{q}; \end{aligned}$$

Min-max formulation

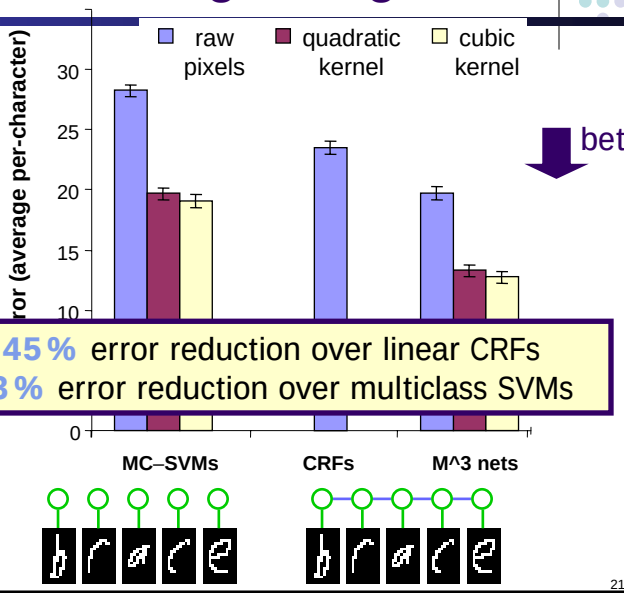
$$\begin{aligned} &\min_{\mathbf{w},\mu} \; \frac{1}{2}||\mathbf{w}||^2 \\ &\text{s.t.} \;\; \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}^*) \geq \mathbf{b}^\top \mu; \\ &\qquad \mathbf{A}^\top \mu \geq (\mathbf{w}^\top \mathbf{F} + \ell)^\top \end{aligned}$$

- Formulation produces compact QP for
  - Low-treewidth Markov networks
  - Associative Markov networks
  - Context free grammars
  - Bipartite matchings
  - Any problem with compact LP inference

Results: 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

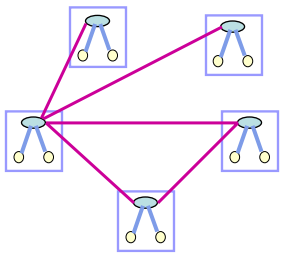


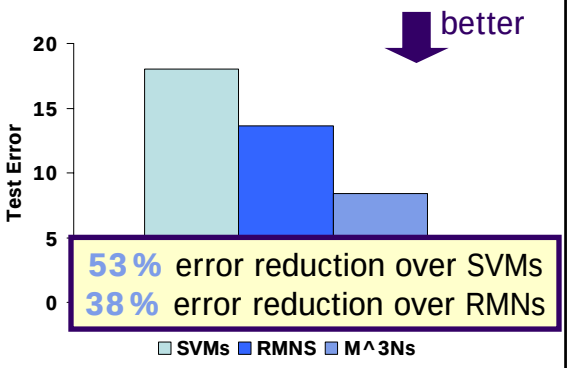
45% error reduction over linear CRFs
33% error reduction over multiclass SVMs

*Crammer & Singer 01

Results: 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





53% error reduction over SVMs
38% error reduction over RMNs

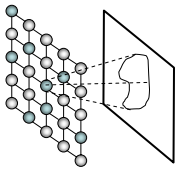
*Taskar et al 02

Summary

- Structured Maximum Margin Networks
  - Concise representation
  - Efficient QP algorithms
  - Applications:
    - Sequences
    - Trees
    - Matchings ….
  - Strong empirical results
- Acknowledgments:
  - Adopted from two different presentations given by Ben Taskar.

Open Problems

- Unsupervised CRF learning and MaxMargin Learning
  - We want to recognize a pattern that is maximally different from the rest!



  - What does margin or conditional likelihood mean in these cases? Given only $\{X_n\}$, how can we define the cost function?

$$p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{\sum_c \theta_c f_c(x, y_c)\right\}$$

$$\text{margin} = w^T \left(F(y_n, x_n) - F(y'_n, x_n)\right)$$

  - Algorithmic challenge