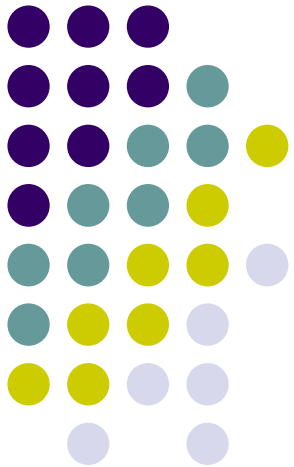
Machine Learning

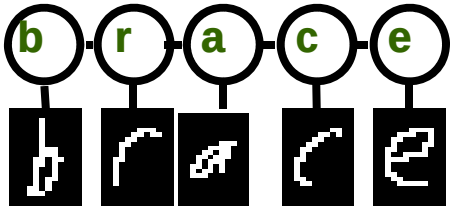
Learning Graphical Models

Max-margin learning of GM



Eric Xing

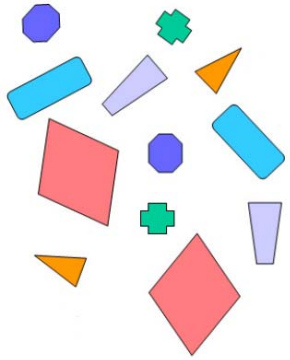
Lecture 16, August 15, 2010



Reading:

Structured Prediction Problem

- Unstructured prediction



$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \dots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 0/1 \end{pmatrix}$$

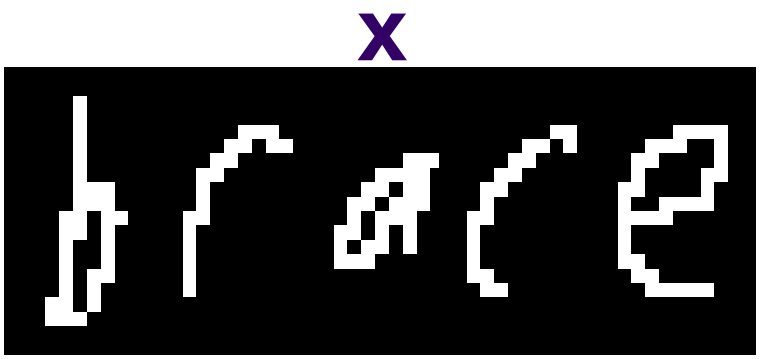
- Structured prediction
  - Part of speech tagging

$\mathbf{x}$ = "Do you want sugar in it?" $\Rightarrow$ $\mathbf{y}$ =<verb pron verb noun prep pron>

  - Image segmentation

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & \dots \\ x_{21} & x_{22} & \dots \\ \vdots & \vdots & \dots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_{11} & y_{12} & \dots \\ y_{21} & y_{22} & \dots \\ \vdots & \vdots & \dots \end{pmatrix}$$

OCR example



x

y

brace

Sequential structure

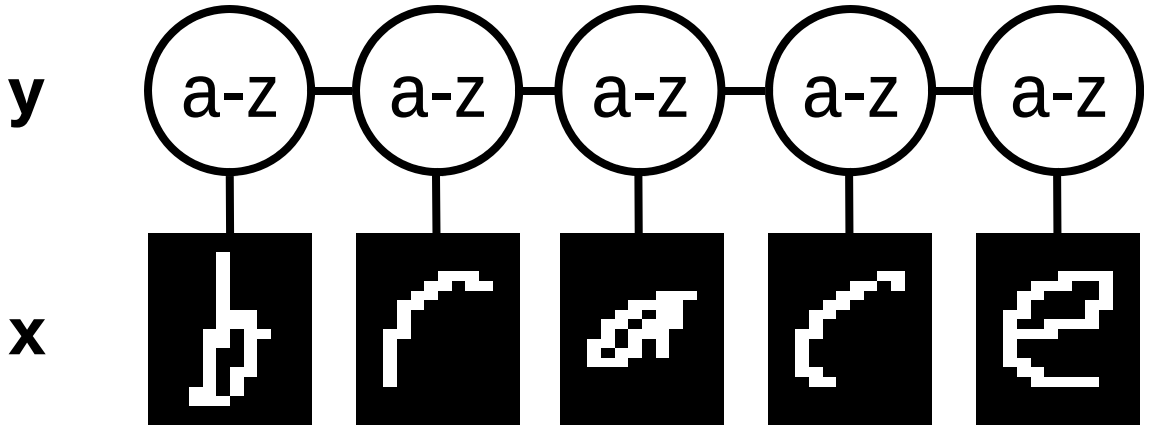
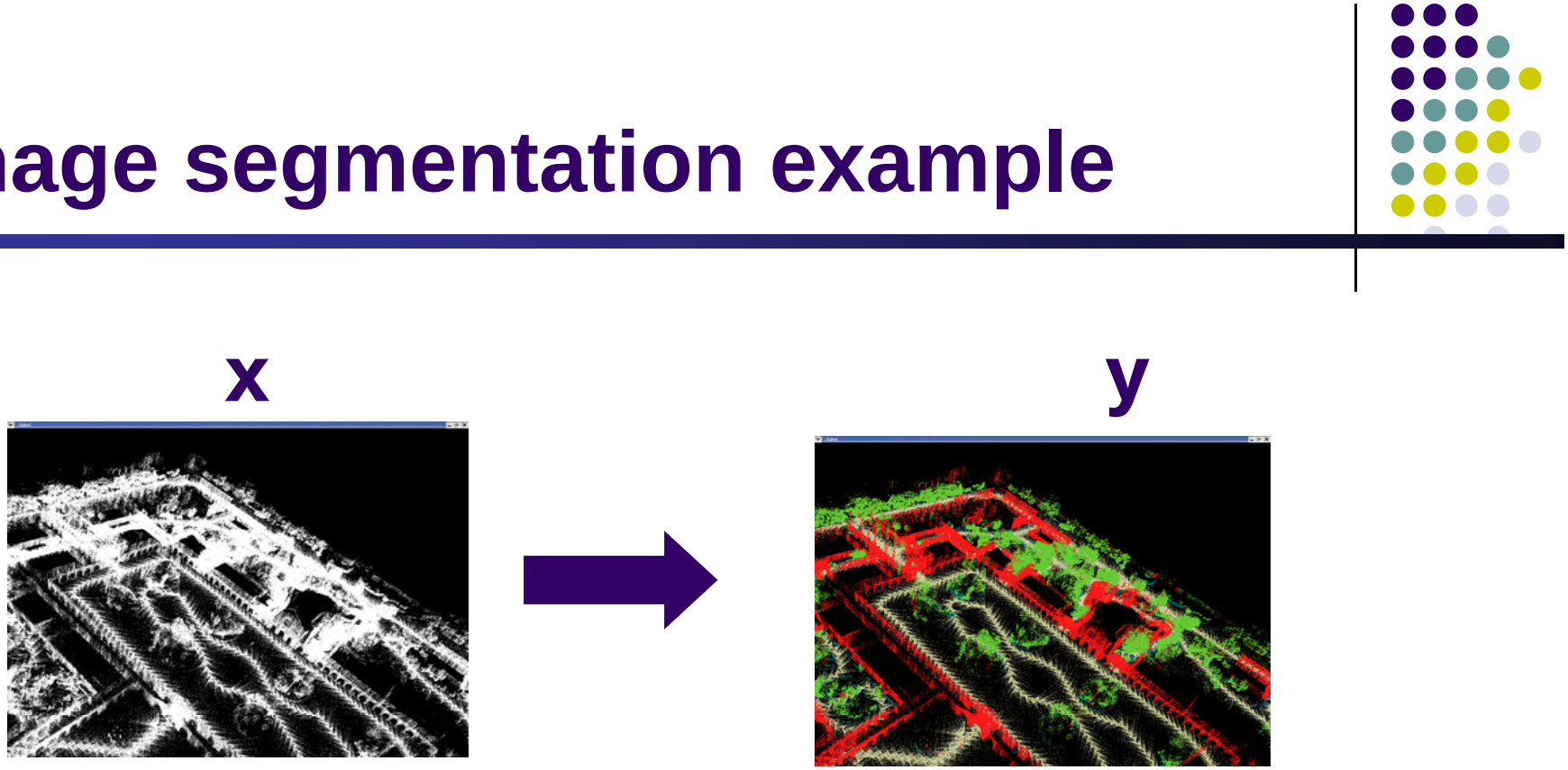
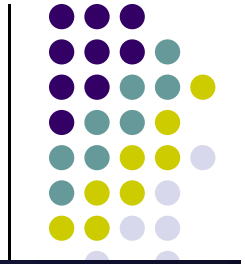


Image segmentation example



Classical Predictive Models



- Inputs:
  - a set of training samples $\mathcal{D} = \{(\mathrm{x}^i, y^i)\}_{i=1}^N$ $\mathrm{x}^i = [x_1^i, x_2^i, \cdots, x_d^i]^\top$ and $y^i \in C \triangleq \{c_1, c_2, \cdots, c_L\}$
- Outputs:
  - a predictive function $h(\mathrm{x})$ $y^\star = h(\mathrm{x}) \triangleq \arg\max_y F(\mathrm{x}, y; \mathbf{w})$
- Examples: $F(\mathrm{x}, y; \mathbf{w}) = g(\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y))$

– Logistic Regression, Bayes classifiers

- Max-likelihood estimation

E.g.: $\max_{\mathbf{w}} \mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \sum_{i=1}^N \log p(y^i|\mathrm{x}^i)$

$$p(y|\mathrm{x}) = \frac{\exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)\}}{\sum_{y'} \exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y')\}}$$

Advantages:

1.Full probabilistic semantics
2.Straightforward Bayesian or direct regularization
3.Hidden structures or generative hierarchy

– Support Vector Machines (SVM)

- Max-margin learning

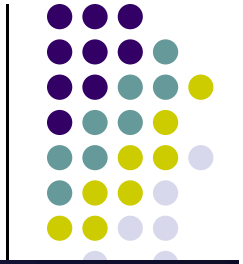
$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^N \xi_i;$$

$$\text{s.t. } \mathbf{w}^\top \Delta\mathbf{f}_i(y) \geq 1 - \xi_i, \ \forall i, \forall y \neq y^i.$$

Advantages:

1.Dual sparsity: few support vectors
2.Kernel tricks
3.Strong empirical results

Structured Prediction Models



- Conditional Random Fields (CRFs) (Lafferty et al 2001)
  - Based on Logistic Regression
  - Max-likelihood estimation (point-estimate)

$$\max_{\mathbf{w}} \mathcal{L}(\mathcal{D};\mathbf{w}) \triangleq \sum_{i=1}^{N} \log p(\mathbf{y}^i|\mathbf{x}^i)$$

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp\{\mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y})\}}{\sum_{\mathbf{y}'} \exp\{\mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}')\}}$$

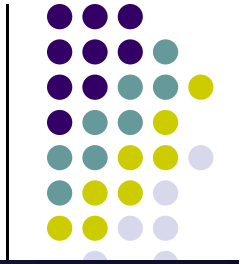
- Max-margin Markov Networks (M^3Ns) (Taskar et al 2003)
  - Based on SVM
  - Max-margin learning (point-estimate)

$$\text{P0 } (\text{M}^3\text{N}): \min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t. } \forall i, \forall \mathbf{y} \neq \mathbf{y}^i : \mathbf{w}^\top \Delta\mathbf{f}_i(\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i,\ \xi_i \geq 0\ ,$$

where $\mathbf{w}^\top \Delta\mathbf{f}_i(\mathbf{y}|\mathbf{x}_i)$ denotes the margin and $\Delta\ell_i(\mathbf{y})$ is a loss function.

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:

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

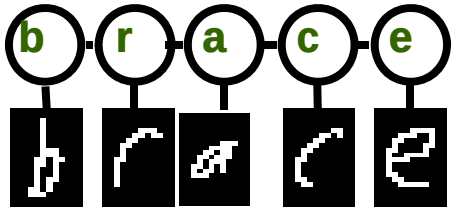
linear combination of features

sum of part scores:

- index *p* represents a part in the structure

Learning w

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



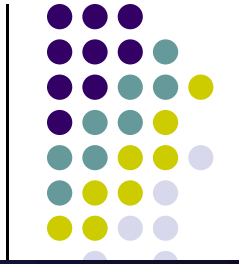
- Probabilistic approach:

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

- 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:

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

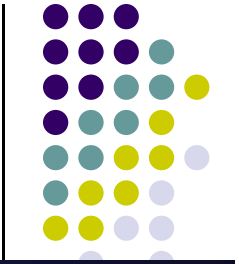
- Max-Margin:
  - We predict based on:

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

  - And we learn based on:

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

OCR Example



- We want:

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

- Equivalently:

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

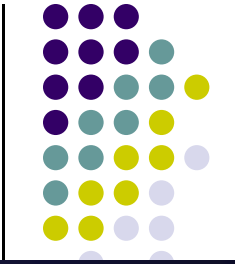
  $\mathbf{w}^T \mathbf{f}($  $, \text{“brace”}) > \mathbf{w}^T \mathbf{f}($  $, \text{“aaaab”})$

  …

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

a lot!

Large Margin Estimation



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

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

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

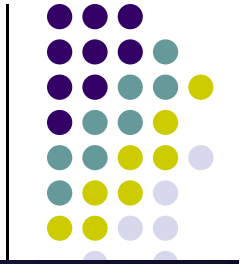
$$\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}^*) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y}) + \gamma\, \ell(\mathbf{y}^*, \mathbf{y}) \quad \forall \mathbf{y}$$

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

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

*Taskar et al. 03

Large Margin Estimation

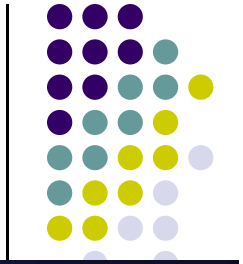


- Recall from SVMs:
  - Maximizing margin γ is equivalent to minimizing the square of the L2-norm of the weight vector **w**:

- 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

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

$$\text{add to QP}: \quad \mathbf{w}^\top \mathbf{f}(\mathbf{x}^i, \mathbf{y}^i) \geq \mathbf{w}^\top \mathbf{f}(\mathbf{x}^i, \mathbf{y}') + \ell(\mathbf{y}^i, \mathbf{y}')$$

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

Min-max Formulation

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

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

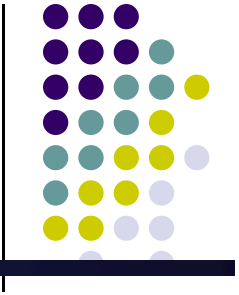
- Key step: convert the maximization in the constraint from discrete to continuous
  - This enables us to plug it into a QP

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

discrete optim. continuous optim.

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

$\mathbf{y} \Rightarrow \mathbf{z}$ map for linear chain structures



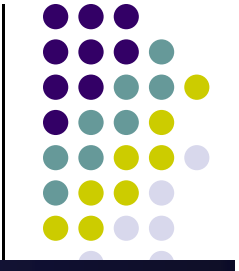
OCR example: $\mathbf{y}$ = 'ABABB';

z's are the indicator variables for the corresponding classes (alphabet)

	$z_1(m)$	$z_2(m)$	$z_3(m)$	$z_4(m)$	$z_5(m)$
A	1	0	1	0	0
B	0	1	0	1	1
:	:	:	:	:	:
B	0	0	0	0	0

	$z_{12}(m,n)$	$z_{23}(m,n)$	$z_{34}(m,n)$	$z_{45}(m,n)$
A	0 1 . 0	0 0 . 0	0 1 . 0	0 0 . 0
B	0 0 . 0	1 0 . 0	0 0 . 0	0 1 . 0
:	. . . 0	. . . 0	. . . 0	. . . 0
B	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
	A B . B	A B . B	A B . B	A B . B

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

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

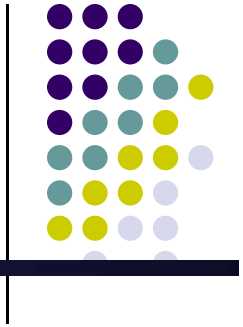
normalization $\sum_m z_j(m) = 1$

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

$$\mathbf{A}\mathbf{z} = \mathbf{b}$$

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

Min-max formulation



- Original problem:

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

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

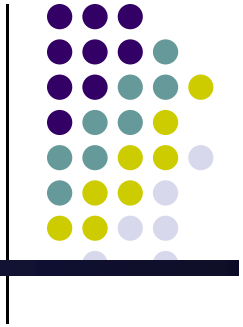
- Transformed problem:

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

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

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

Min-max formulation



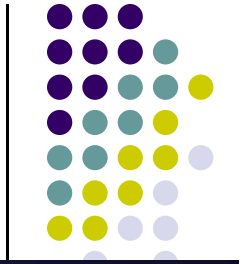
- Using strong Lagrangian duality:
  (beyond the scope of this lecture)

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

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

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

Min-max formulation



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

- 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

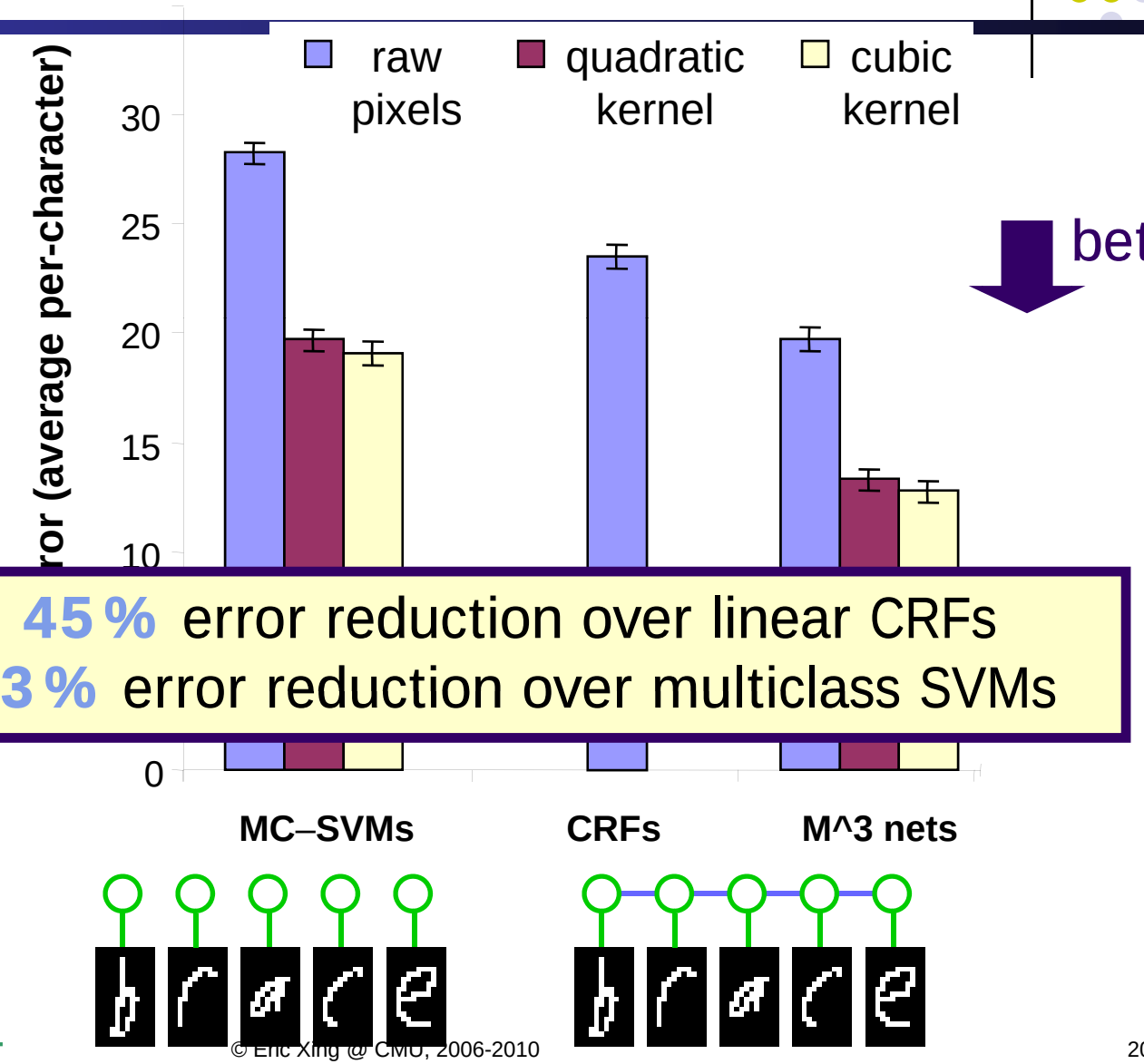
10-fold Train/Test

5000/50000 letters

600/6000 words

Models:

- Multiclass-SVMs
- CRFs
- M^3 nets



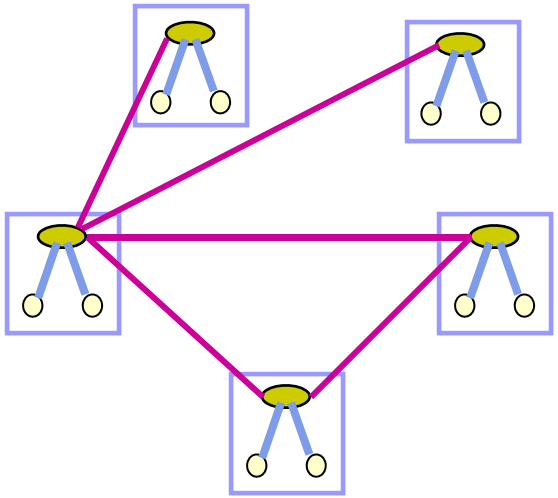
45% error reduction over linear CRFs

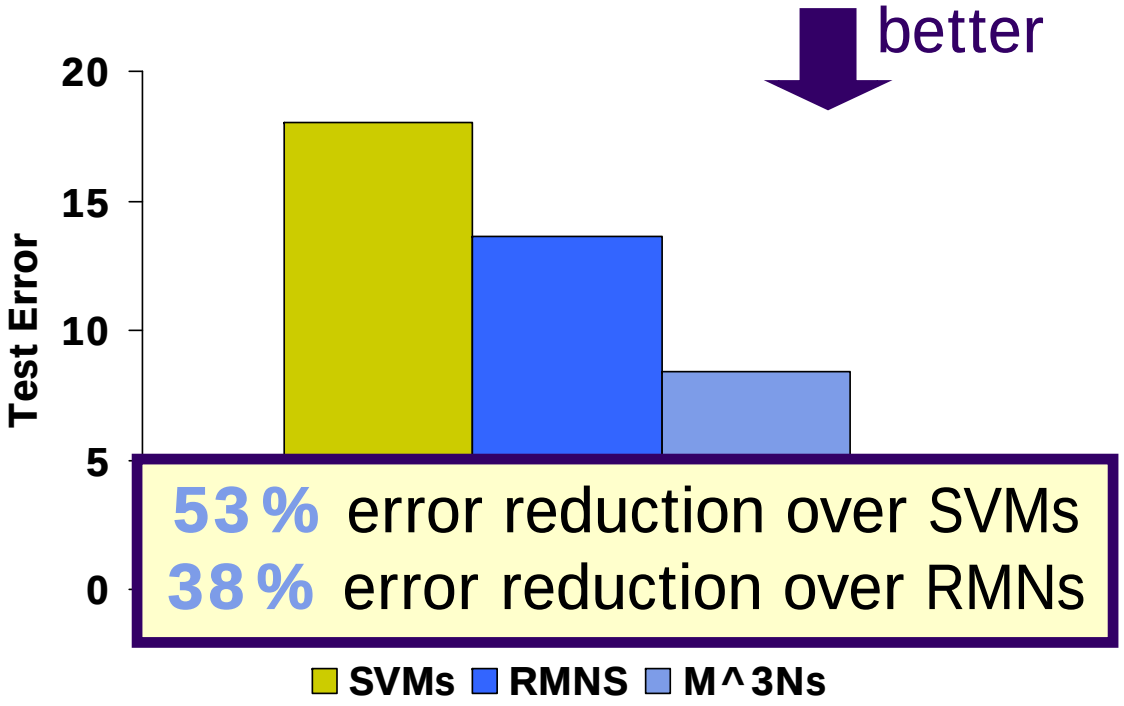
33% error reduction over multiclass SVMs

*Crammer & Singer 01

Results: Hypertext Classification

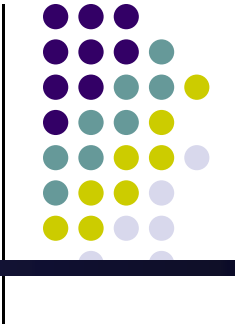
- WebKB dataset
  - Four CS department websites: 1300 pages/3500 links
  - Classify each page: faculty, course, student, project, other
  - Train on three universities/test on fourth





*Taskar et al 02

Maximum Entropy Discrimination Markov Networks



- Structured MaxEnt Discrimination (SMED):

$$\text{P1}: \quad \min_{p(\mathbf{w}),\xi} \; KL(p(\mathbf{w})||p_0(\mathbf{w})) + U(\xi)$$

$$\text{s.t.} \;\; p(\mathbf{w}) \in \mathcal{F}_1, \;\; \xi_i \geq 0, \forall i.$$

generalized maximum entropy or *regularized* KL-divergence

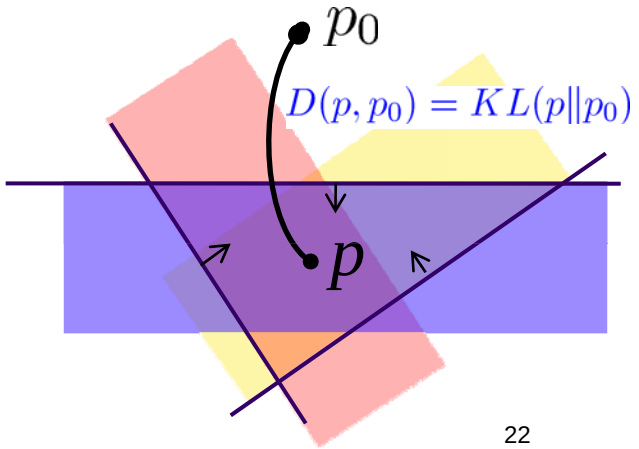
- Feasible subspace of weight distribution:

$$\mathcal{F}_1 = \left\{p(\mathbf{w}) : \int p(\mathbf{w})[\Delta F_i(\mathbf{y};\mathbf{w}) - \Delta \ell_i(\mathbf{y})]\, d\mathbf{w} \geq -\xi_i, \; \forall i, \forall \mathbf{y} \neq \mathbf{y}^i\right\},$$

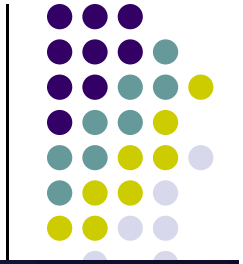
expected margin constraints.

- Average from distribution of M³Ns

$$h_1\left(\mathbf{x}; p(\mathbf{w})\right) = \arg \max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} \int p(\mathbf{w}) F(\mathbf{x}, \mathbf{y}; \mathbf{w}) d\mathbf{w}$$



Solution to MaxEnDNet



- Theorem:
  - Posterior Distribution:

$$p(\mathbf{w}) = \frac{1}{Z(\alpha)} p_0(\mathbf{w}) \exp\left\{ \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) [\Delta F_i(\mathbf{y}; \mathbf{w}) - \Delta \ell_i(\mathbf{y})] \right\}$$

  - Dual Optimization Problem:

$$\begin{aligned} \text{D1}: \quad & \max_{\alpha} \; -\log Z(\alpha) - U^\star(\alpha) \\ & \text{s.t.} \;\; \alpha_i(\mathbf{y}) \geq 0, \;\; \forall i, \;\; \forall \mathbf{y}, \end{aligned}$$

$U^\star(\cdot)$ is the conjugate of the $U(\cdot)$, i.e., $U^\star(\alpha) = \sup_\xi \left(\sum_{i,\mathrm{y}} \alpha_i(\mathrm{y}) \xi_i - U(\xi) \right)$

Gaussian MaxEnDNet (reduction to M³N)

- Theorem
  - Assume

$$F(\mathbf{x},\mathbf{y};\mathbf{w}) = \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}), U(\xi) = C\sum_i \xi_i, \text{and } p_0(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0,I)$$

  - Posterior distribution:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}, I), \text{ where } \mu_{\mathbf{w}} = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y})$$

  - Dual optimization:

$$\max_{\alpha} \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\ell_i(\mathbf{y}) - \frac{1}{2}\|\sum_{i,\mathbf{y}} \alpha_i(\mathbf{y})\Delta\mathbf{f}_i(\mathbf{y})\|^2$$

$$\text{s.t.} \quad \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C;\ \alpha_i(\mathbf{y}) \geq 0,\ \forall i,\ \forall \mathbf{y},$$

  - Predictive rule:

$$h_1(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \int p(\mathbf{w})F(\mathbf{x},\mathbf{y};\mathbf{w})\,d\mathbf{w} = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} \mu_{\mathbf{w}}^\top \mathbf{f}(\mathbf{x},\mathbf{y})$$

- Thus, MaxEnDNet subsumes M³Ns and admits all the merits of max-margin learning
- Furthermore, MaxEnDNet has at least three advantages …

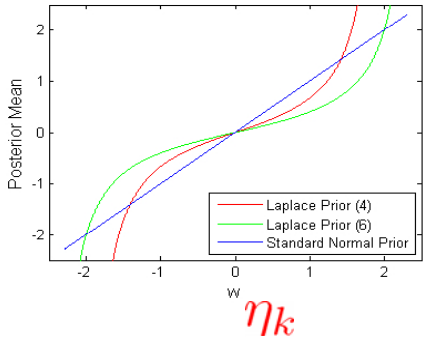
Three Advantages

- An averaging Model: PAC-Bayesian prediction error guarantee

$$\Pr_Q(M(h,\mathbf{x},\mathbf{y}) \le 0) \le \Pr_{\mathcal{D}}(M(h,\mathbf{x},\mathbf{y}) \le \gamma) + O\left(\sqrt{\frac{\gamma^{-2}KL(p||p_0)\ln(N|\mathcal{Y}|) + \ln N + \ln \delta^{-1}}{N}}\right).$$

- Entropy regularization: Introducing useful biases
  - Standard Normal prior => reduction to standard M3N (we've seen it)
  - Laplace prior => Posterior shrinkage effects (sparse M3N)

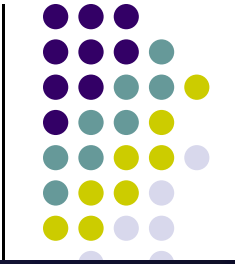
$$\forall k,\ \langle w_k \rangle_p = \frac{2\eta_k}{\lambda - \eta_k^2}$$



η_k

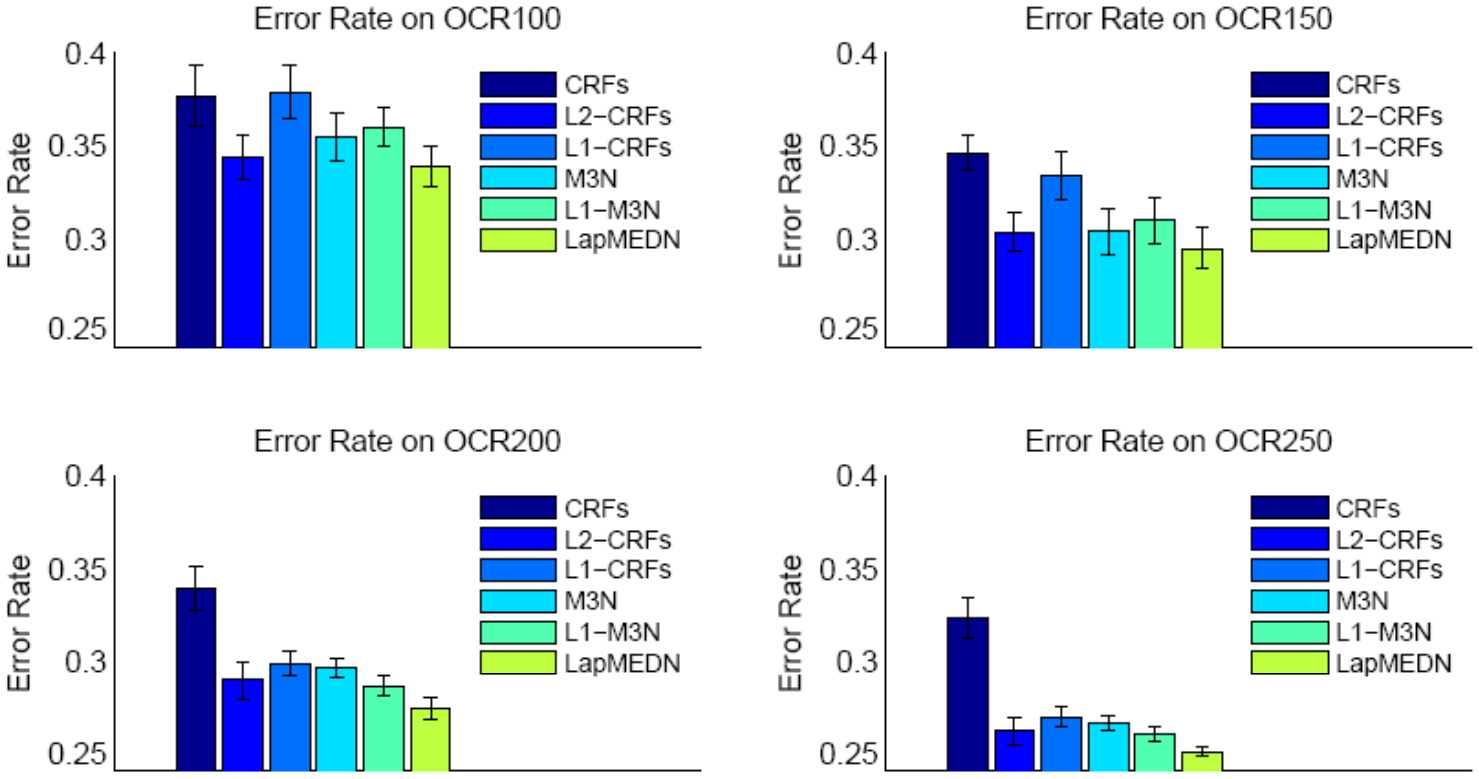
- Integration of Generative and Discriminative principles
  - Incorporate latent variables and structures (PoMEN)
  - Semisupervised learning (with partially labeled data)

Experimental results on OCR datasets

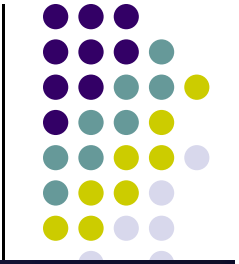


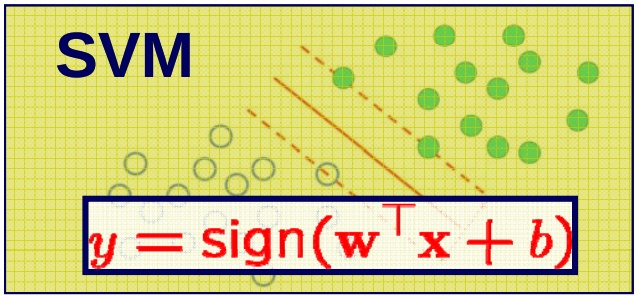
$$(\text{CRFs},\ L_1-\text{CRFs},\ L_2-\text{CRFs},\ \text{M}^3\text{Ns},\ L_1-\text{M}^3\text{Ns},\ \text{and LapMEDN})$$

- We randomly construct OCR100, OCR150, OCR200, and OCR250 for 10 fold CV.

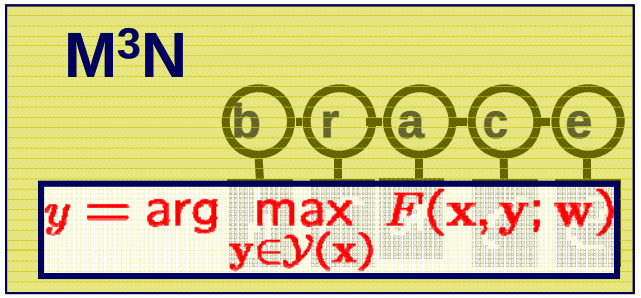


Margin-Based Discriminative Learning Paradigms

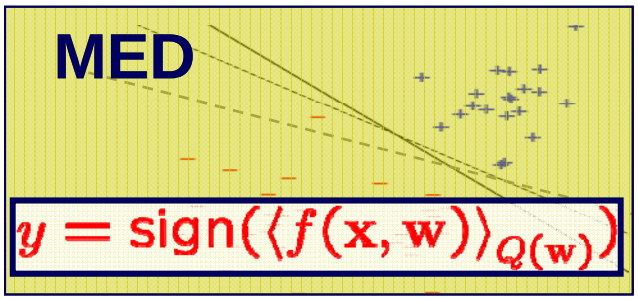




$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}\xi_i$$
$$y^i(\mathbf{w}^\top \mathbf{x}^i + b) \geq 1 - \xi_i, \quad \forall i$$



$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{m}\xi_i$$
$$\mathbf{w}^\top[\mathbf{f}(\mathbf{x}^i, \quad) - \mathbf{f}(\mathbf{x}^i,\mathbf{y})] \geq \ell(\mathbf{y}^i,\mathbf{y}) - \xi_i, \quad \forall i, \forall \mathbf{y} \neq \mathbf{y}^i$$



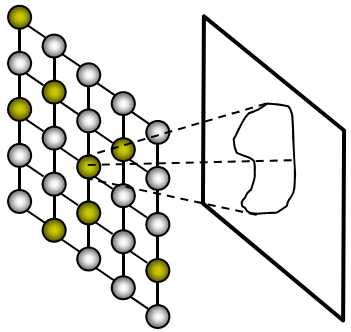
$$\min_{Q} \quad \text{KL}(Q||Q_0)$$
$$y^i\langle f(\mathbf{x}^i)\rangle_Q \geq \xi_i, \quad \forall i$$

MED-MN

= SMED + Bayesian M³N

Open Problems

- Unsupervised CRF learning and MaxMargin Learning
  - Only X, but not Y (sometimes part of Y), is available
  - We want to recognize a pattern that is maximally different from the rest!



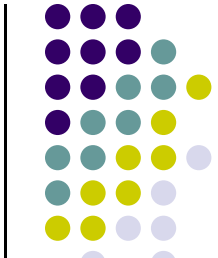
  - What does margin or conditional likelihood mean in these cases? Given only $\{X_n\}$, how can we define the cost function?

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

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

  - Algorithmic challenge
  - Stay tuned for lecture 19!

Remember: Elements of Learning



- **Here are some important elements to consider before you start:**
  - Task:
    - Embedding? Classification? Clustering? Topic extraction? …
  - Data and other info:
    - Input and output (e.g., continuous, binary, counts, …)
    - Supervised or unsupervised, of a blend of everything?
    - Prior knowledge? Bias?
  - Models and paradigms:
    - BN? MRF? Regression? SVM?
    - Bayesian/Frequents ? Parametric/Nonparametric?
  - Objective/Loss function:
    - MLE? MCLE? Max margin?
    - Log loss, hinge loss, square loss? …
  - Tractability and exactness trade off:
    - Exact inference? MCMC? Variational? Gradient? Greedy search?
    - Online? Batch? Distributed?
  - Evaluation:
    - Visualization? Human interpretability? Perperlexity? Predictive accuracy?
- **It is better to consider one element at a time!**