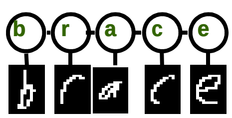
Machine Learning

10-701/15-781, Fall 2011

Max-Margin Learning of Graphical Models

Eric Xing

Lecture 21, November 28, 2011



Advanced topics in Max-Margin Learning (cont.)

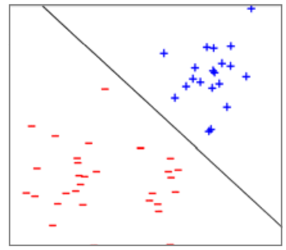
$$\max_\alpha \mathcal{J}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

$$\mathbf{w}^T \mathbf{x}_{\mathrm{new}} + b \lessgtr 0$$

- Implicit nonlinear data transformation
  - The Kernel trick
- Point rule or average rule?
  - Maximum entropy discrimination
- Can we predict vec(y)?
  - Structured SVM, aka, Maximum Margin Markov Networks
- Putting everything all together !

(2) Model averaging

- Inputs x, class y = +1, -1
- data D = { (x_1,y_1), …. (x_m,y_m) }

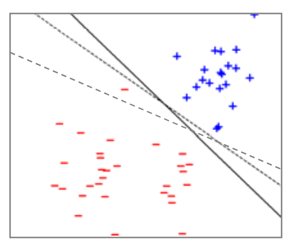


- Point Rule:
  - learn $f^{opt}(x)$ discriminant function from F = {f} family of discriminants
  - classify y = sign $f^{opt}(x)$
- E.g., SVM

$$f^{\text{opt}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}_{\text{new}} + b$$

Model averaging

- There exist many f with near optimal performance



- Instead of <u>choosing</u> f^{opt}, <u>average</u> over all f in F

  Q(f) = weight of f

$$\begin{aligned} y(x) &= \text{sign} \int_F Q(f) f(x) df \\ &= \text{sign} \langle f(x) \rangle_Q \end{aligned}$$

- How to specify:
  F = { f } family of discriminant functions?
- How to learn Q(f) distribution over F?

Recall Bayesian Inference

- Bayesian learning:

$$p_0(\mathbf{w}),\ \mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \longrightarrow \boxed{\textbf{Bayes Learner}} \longrightarrow p(\mathbf{w}|\mathcal{D})$$

$$\text{Bayes Thrm}: p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathbf{w})p(\mathcal{D}|\mathbf{w})}{p(\mathcal{D})}$$

- Bayes Predictor (model averaging):

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

Recall in SVM: $h_0(\mathbf{x}; \mathbf{w}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} F(\mathbf{x}, \mathbf{y}; \mathbf{w})$

- What p_0?

How to score distributions?

- Entropy
  - Entropy *H(X)* of a random variable *X*

$$\boxed{H(X) = -\sum_{i=1}^{N} P(x=i)\log_2 P(x=i)}$$

  - *H(X)* is the expected number of bits needed to encode a randomly drawn value of *X* (under most efficient code)
  - Why?

Information theory:

Most efficient code assigns $-\log_2 P(X=i)$ bits to encode the message *X=I*, So, expected number of bits to code one random *X* is:

$$-\sum_{i=1}^{N} P(x=i)\log_2 P(x=i)$$

Maximum Entropy Discrimination

- Given data set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, find

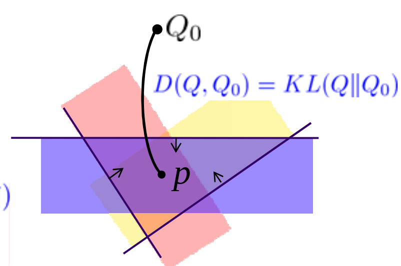
$$
\begin{aligned}
& Q_{\mathrm{ME}} = \arg\max \quad \mathrm{H}(Q) \\
\text{s.t.} \quad & y^i \langle f(\mathbf{x}^i) \rangle_{Q_{\mathrm{ME}}} \geq \xi_i, \quad \forall i \\
& \xi_i \geq 0 \quad \forall i
\end{aligned}
$$

- solution Q_{ME} correctly classifies $\mathcal{D}$
- among all admissible Q, Q_{ME} has max entropy
- max entropy ➡ "minimum assumption" about f

Introducing Priors

- Prior $Q_0(f)$
- Minimum Relative Entropy Discrimination

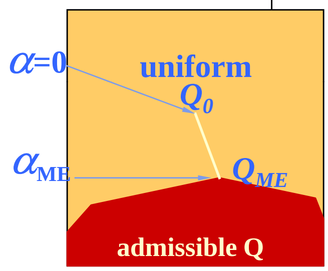
$$
\begin{aligned}
Q_{\mathrm{MRE}} \quad = \quad & \arg\min \quad \mathrm{KL}(Q \| Q_0) + U(\xi) \\
\text{s.t.} \quad & y^i \langle f(\mathbf{x}^i) \rangle_{Q_{\mathrm{ME}}} \geq \xi_i, \quad \forall i \\
& \xi_i \geq 0 \quad \forall i
\end{aligned}
$$



- Convex problem: Q_{MRE} unique solution
- MER ➡ "minimum *additional* assumption" over Q_0 about f

Solution: Q_{ME} as a projection

- Convex problem: Q_{ME} unique



- Theorem:

$$Q_{\text{MRE}} \propto \exp\{\sum_{i=1}^{N} \alpha_i y_i f(x_i; w)\} Q_0(w)$$

$\alpha_i \geq 0$ Lagrange multipliers

- finding Q_M : start with $\alpha_i = 0$ and follow gradient of unsatisfied constraints

 9

Solution to MED

- Theorem (Solution to MED):
  - Posterior Distribution:

$$Q(\mathbf{w}) = \frac{1}{Z(\alpha)} Q_0(\mathbf{w}) \exp\{\sum_i \alpha_i y_i [f(\mathbf{x}_i; \mathbf{w})]\}$$

  - Dual Optimization Problem:

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

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

- Algorithm: to computer α_t , $t = 1, \ldots \mathrm{T}$
  - start with $\alpha_i = 0$ (uniform distribution)
  - iterative ascent on $J(\alpha)$ until convergence

 10

Examples: SVMs

- **Theorem**

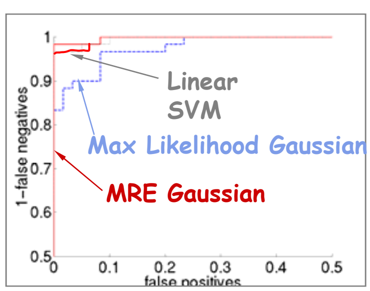
For $f(x) = w^T x + b$, $Q_0(w)$ = Normal(0, I), $Q_0(b)$ = non-informative prior, the Lagrange multipliers α are obtained by maximizing $J(\alpha)$ subject to $0 \leq \alpha_t \leq C$ and $\sum_t \alpha_t y_t = 0$, where

$$J(\alpha) = \sum_t \left[\alpha_t + \log(1 - \alpha_t / C)\right] - \frac{1}{2} \sum_{s,t} \alpha_s \alpha_t y_s y_t x_s^T x_t$$

- Separable D ➡ SVM recovered exactly
- Inseparable D ➡ SVM recovered with different misclassification penalty

SVM extensions

- Example: Leptograpsus Crabs (5 inputs, T_{train}=80, T_{test}=120)



(3) Structured Prediction

- Unstructured prediction

$$\mathbf{x} = \begin{pmatrix} \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots \\ \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots \\ \vdots & \vdots & \ldots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}$$

- Structured prediction
  - Part of speech tagging

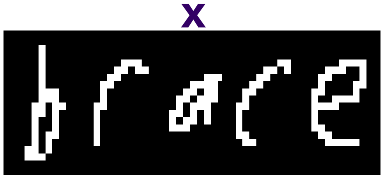
$\mathbf{x} =$ **"Do you want sugar in it?"** $\Rightarrow$ $\mathbf{y} =$ **<verb pron verb noun prep pron>**

  - Image segmentation

$$\mathbf{x} = \begin{pmatrix} \mathrm{x}_{11} & \mathrm{x}_{12} & \cdots \\ \mathrm{x}_{21} & \mathrm{x}_{22} & \cdots \\ \vdots & \vdots & \ldots \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_{11} & y_{12} & \cdots \\ y_{21} & y_{22} & \cdots \\ \vdots & \vdots & \ldots \end{pmatrix}$$

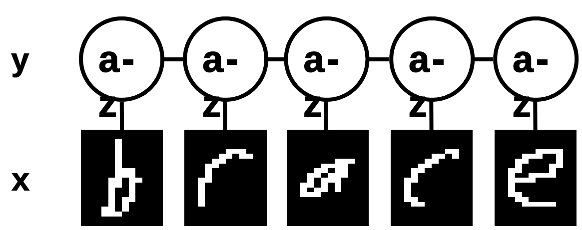
OCR example

x **y**



$\rightarrow$ **brace**

Sequential structure



Classical Predictive Models

- Input and output space: $\mathcal{X} \triangleq \mathbb{R}^{M_x}$ $\quad \mathcal{Y} \triangleq \{-1, +1\}$
- Predictive function $h(\mathrm{x})$: $y^\star = h(\mathrm{x}) \triangleq \arg\max_{y \in \mathcal{Y}} F(\mathrm{x}, y; \mathbf{w})$
- Examples: $F(\mathrm{x}, y; \mathbf{w}) = g(\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y))$
- Learning: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathcal{W}} \ell(\mathrm{x}, y; \mathbf{w}) + \lambda R(\mathbf{w})$

where $\ell(\cdot)$ represents a convex loss, and $R(\mathbf{w})$ is a regularizer preventing overfitting

– Logistic Regression
  - Max-likelihood (or MAP) estimation

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

  - Corresponds to a Log loss with L2 R

$$\ell_{LL}(\mathrm{x}, y; \mathbf{w}) \triangleq \ln \sum_{y' \in \mathcal{Y}} \exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y')\} - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)$$

– Support Vector Machines (SVM)
  - Max-margin learning

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

$$\text{s.t. } \forall i, \forall y' \neq y^i : \mathbf{w}^\top \Delta \mathbf{f}_i(y') \geq 1 - \xi_i, \ \xi_i \geq 0.$$

  - Corresponds to a hinge loss with L2 R

$$\ell_{MM}(\mathrm{x}, y; \mathbf{w}) \triangleq \max_{y' \in \mathcal{Y}} \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y') - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y) + \ell'(y', y)$$

Classical Predictive Models

- Inputs:
  - a set of training samples $\mathcal{D} = \{(\mathrm{x}^i, y^i)\}_{i=1}^N$

    where $\mathrm{x}^i = [x_1^i, x_2^i, \cdots, x_d^i]^\top$ and $y^i \in C \triangleq \{c_1, c_2, \cdots, c_L\}$
- Outputs:
  - a predictive function $h(\mathrm{x})$ $\ y^\star = h(\mathrm{x}) \triangleq \arg\max_y F(\mathrm{x}, y; \mathbf{w})$

– Logistic Regression
  - Max-likelihood (or MAP) estimation

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

  - Corresponds to a Log loss with L2 R

$$\ell_{LL}(\mathrm{x}, y; \mathbf{w}) \triangleq \ln \sum_{y' \in \mathcal{Y}} \exp\{\mathbf{w}^\top \mathbf{f}(\mathrm{x}, y')\} - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y)$$

Advantages:

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

– Support Vector Machines (SVM)
  - Max-margin learning

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

$$\text{s.t. } \forall i, \forall y' \neq y^i : \mathbf{w}^\top \Delta \mathbf{f}_i(y') \geq 1 - \xi_i, \ \xi_i \geq 0.$$

  - Corresponds to a hinge loss with L2 R

$$\ell_{MM}(\mathrm{x}, y; \mathbf{w}) \triangleq \max_{y' \in \mathcal{Y}} \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y') - \mathbf{w}^\top \mathbf{f}(\mathrm{x}, y) + \ell'(y', y)$$

Advantages:

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

Structured Prediction Graphical Models

- **Input and output space** $\mathcal{X} \triangleq \mathbb{R}_{X_1} \times, \dots, \mathbb{R}_{X_K} \quad \mathcal{Y} \triangleq \mathbb{R}_{Y_1} \times, \dots, \mathbb{R}_{Y_{K'}}$

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

$$\mathcal{L}(\mathcal{D};\mathbf{w}) \triangleq \log \sum_{\mathbf{y}'} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}')) - \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}) + R(\mathbf{w})$$

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

$$\mathcal{L}(\mathcal{D};\mathbf{w}) \triangleq \log \max_{\mathbf{y}'} \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}') - \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}) + \ell(\mathbf{y}',\mathbf{y}) + R(\mathbf{w})$$

Structured Models

$$h(\mathbf{x}) = \arg\max_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} F(\mathbf{x},\mathbf{y})$$

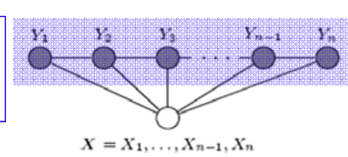
space of feasible outputs **discriminant function**

- Assumptions:

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

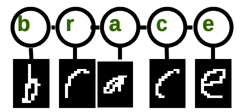
  - Linear combination of features
  - Sum of partial scores: index *p* represents a part in the structure
  - Random fields or Markov network features:

- Markov properties are encoded in the feature fun$\mathbf{f}(\mathbf{x},\mathbf{y})$s



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

Discriminative Learning Strategies

- Recall that in CRF (Max Conditional Likelihood):
  - We predict based on:

$$\mathbf{y}^* \mid \mathbf{x} = \arg\max_{\mathbf{y}} p_{\mathbf{w}}(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{w}, \mathbf{x})} \exp\left\{\sum_c w_c f_c(\mathbf{x}, \mathbf{y}_c)\right\}$$

  - And we learn based on:

$$\mathbf{w}^* \mid \{\mathbf{y}_i, \mathbf{x}_i\} = \arg\max_{\mathbf{w}} \prod_i p_{\mathbf{w}}(\mathbf{y}_i \mid \mathbf{x}_i) = \prod_i \frac{1}{Z(\mathbf{w}, \mathbf{x}_i)} \exp\left\{\sum_c w_c f_c(\mathbf{x}_i, \mathbf{y}_i)\right\}$$

- Max Margin:
  - We predict based on:

$$\mathbf{y}^* \mid \mathbf{x} = \arg\max_{\mathbf{y}} \sum_c w_c f_c(\mathbf{x}, \mathbf{y}_c) = \arg\max_{y} \mathbf{w}^T f(\mathbf{x}, \mathbf{y})$$

  - And we learn based on:

$$\mathbf{w}^* \mid \{\mathbf{y}_i, \mathbf{x}_i\} = \arg\max_{\mathbf{w}} \left(\min_{\mathbf{y} \neq \mathbf{y}^i, \forall i} \mathbf{w}^T \left(f(\mathbf{y}_i, \mathbf{x}_i) - f(\mathbf{y}, \mathbf{x}_i) \right) \right)$$

E.g. Max-Margin Markov Networks

- Convex Optimization Problem:

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

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

- Feasible subspace of weights:

$$\mathcal{F}_0 = \{\mathbf{w} : \mathbf{w}^\top \Delta\mathbf{f}_i(\mathbf{x},\mathbf{y}) \geq \Delta\ell_i(\mathbf{y}) - \xi_i;\ \forall i, \forall \mathbf{y} \neq \mathbf{y}_i\}$$

- Predictive Function:

$$h_0(\mathbf{x};\mathbf{w}) = \arg\max_{\mathbf{y}\in\mathcal{Y}(\mathbf{x})} F(\mathbf{x},\mathbf{y};\mathbf{w})$$

OCR Example

- We want:

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

- Equivalently:

  w^T f( ,"brace") > w^T f( ,"aaaaa")
  w^T f( ,"brace") > w^T f( ,"aaaab")
  …
  w^T f( ,"brace") > w^T f( ,"zzzzz")

  a lot!

Large Margin Estimation

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

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

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

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

- **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
  - Several algorithms exist …

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

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

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';

$\mathbf{z}$'s are the indicator variables for the corresponding classes (alphabet)

$$\begin{array}{c} A \\ B \\ : \\ B \end{array} \quad z_1(m) = \begin{bmatrix} 1 \\ 0 \\ : \\ 0 \end{bmatrix} \quad z_2(m) = \begin{bmatrix} 0 \\ 1 \\ : \\ 0 \end{bmatrix} \quad z_3(m) = \begin{bmatrix} 1 \\ 0 \\ : \\ 0 \end{bmatrix} \quad z_4(m) = \begin{bmatrix} 0 \\ 1 \\ : \\ 0 \end{bmatrix} \quad z_5(m) = \begin{bmatrix} 0 \\ 1 \\ : \\ 0 \end{bmatrix}$$

$$\begin{array}{c} A \\ B \\ : \\ B \\ \\ \end{array} \quad \begin{array}{c} z_{12}(m,n) \\ \begin{bmatrix} 0 & 1 & . & 0 \\ 0 & 0 & . & 0 \\ . & . & . & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ \begin{array}{cccc} A & B & . & B \end{array} \end{array} \quad \begin{array}{c} z_{23}(m,n) \\ \begin{bmatrix} 0 & 0 & . & 0 \\ 1 & 0 & . & 0 \\ . & . & . & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ \begin{array}{cccc} A & B & . & B \end{array} \end{array} \quad \begin{array}{c} z_{34}(m,n) \\ \begin{bmatrix} 0 & 1 & . & 0 \\ 0 & 0 & . & 0 \\ . & . & . & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ \begin{array}{cccc} A & B & . & B \end{array} \end{array} \quad \begin{array}{c} z_{45}(m,n) \\ \begin{bmatrix} 0 & 0 & . & 0 \\ 0 & 1 & . & 0 \\ . & . & . & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ \begin{array}{cccc} A & B & . & B \end{array} \end{array}$$

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

Rewriting the maximization function in terms of indicator variables:

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

$$z_k(n) = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$$

$$z_j(m) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad z_{jk}(m,n) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\left.\begin{array}{l} z_j(m) \geq 0; \; z_{jk}(m,n) \geq 0; \\ \textbf{normalization} \quad \sum_m z_j(m) = 1 \\ \textbf{agreement} \quad \sum_n z_{jk}(m,n) = z_j(m) \end{array}\right\} \mathbf{A}\mathbf{z} = \mathbf{b}$$

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

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)

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

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

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

Min-max formulation

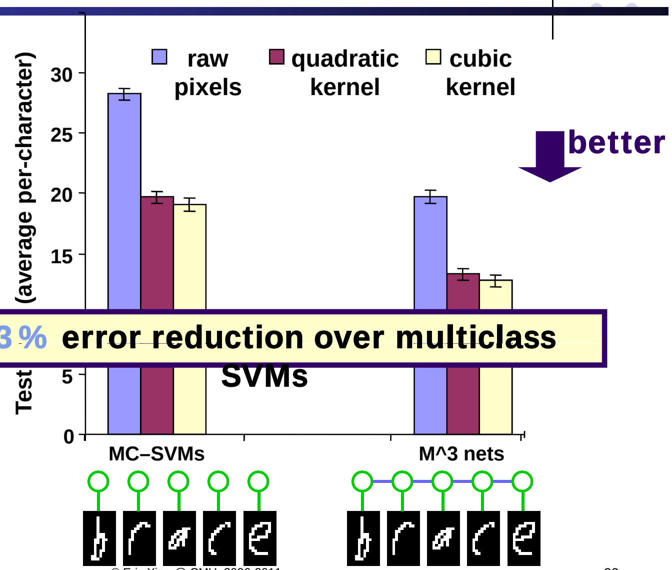
$$\boxed{\begin{aligned} &\min_{\mathbf{w},\mu} \ \frac{1}{2}||\mathbf{w}||^2 \\ &\text{s.t.} \ \ \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}^*) \geq \mathbf{b}^\top \mu; \\ &\qquad \mathbf{A}^\top \mu \geq (\mathbf{w}^\top \mathbf{F} + \ell)^\top \end{aligned}}$$

- 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-SV
M^3 nets



Crammer & Singer 01

MLE versus max-margin learning

- **Likelihood-based estimation**
  - Probabilistic (joint/conditional likelihood model)
  - Easy to perform Bayesian learning, and incorporate prior knowledge, latent structures, missing data
  - Bayesian or direct regularization
  - Hidden structures or generative hierarchy

- **Max-margin learning**
  - Non-probabilistic (concentrate on input-output mapping)
  - Not obvious how to perform Bayesian learning or consider prior, and missing data
  - Support vector property, sound theoretical guarantee with limited samples
  - Kernel tricks

- **Maximum Entropy Discrimination (MED) (Jaakkola, et al., 1999)**
  - Model averaging $\hat{y} = \text{sign} \int p(\mathbf{w}) F(x; \mathbf{w})\, d\mathbf{w}$ $\quad (y \in \{+1, -1\})$
  - The optimization problem (binary classification)

$$\min_{p(\Theta)} KL(p(\Theta) || p_0(\Theta))$$
$$\text{s.t.} \int p(\Theta)[y_i F(x; \mathbf{w}) - \xi_i]\, d\Theta \geq 0, \forall i,$$

where Θ is the parameter $\mathbf{w}$ when ξ are kept fixed or the pair $(\mathbf{w}, \xi)$ when we want to optimize over ξ

Maximum Entropy Discrimination Markov Networks

- Structured MaxEnt Discrimination (SMED):

$$\text{P1}: \quad \min_{p(\mathbf{w}), \xi} \boxed{KL(p(\mathbf{w}) || p_0(\mathbf{w})) + U(\xi)}$$
$$\text{s.t. } p(\mathbf{w}) \in \mathcal{F}_1, \; \xi_i \geq 0, \forall i.$$

generalized maximum entropy or *regularized* KL-divergence

- Feasible subspace of weight distribution:

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

expected margin constraints.

- Average from distribution of M³Ns

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

Solution to MaxEnDNet

- Theorem:
  - Posterior Distribution:

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

  - Dual Optimization Problem:

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

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

Gaussian MaxEnDNet (reduction to M³N)

- Theorem
  - Assume $F(\mathbf{x},\mathbf{y};\mathbf{w}) = \mathbf{w}^\top \mathbf{f}(\mathbf{x},\mathbf{y}), U(\xi) = C \sum_i \xi_i$, and $p_0(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, I)$
    - Posterior distribution: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu_{\mathbf{w}}, I)$, where $\mu_{\mathbf{w}} = \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) \Delta \mathbf{f}_i(\mathbf{y})$
    - Dual optimization:

$$\begin{aligned} & \max_{\alpha} \; \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) \Delta \ell_i(\mathbf{y}) - \frac{1}{2} \Big\| \sum_{i,\mathbf{y}} \alpha_i(\mathbf{y}) \Delta \mathbf{f}_i(\mathbf{y}) \Big\|^2 \\ & \text{s.t.} \;\; \sum_{\mathbf{y}} \alpha_i(\mathbf{y}) = C; \; \alpha_i(\mathbf{y}) \geq 0, \; \forall i, \; \forall \mathbf{y}, \end{aligned}$$

    - 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 …

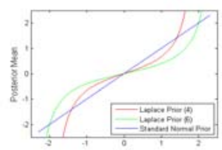
Three Advantages

- An averaging Model: PAC-Bayesian prediction error guarantee

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

- Er
  - Standard Normal prior => reduction to standard M³N (we've seen it)
  - Laplace prior => Posterior shrinkage effects (sparse M³N)

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



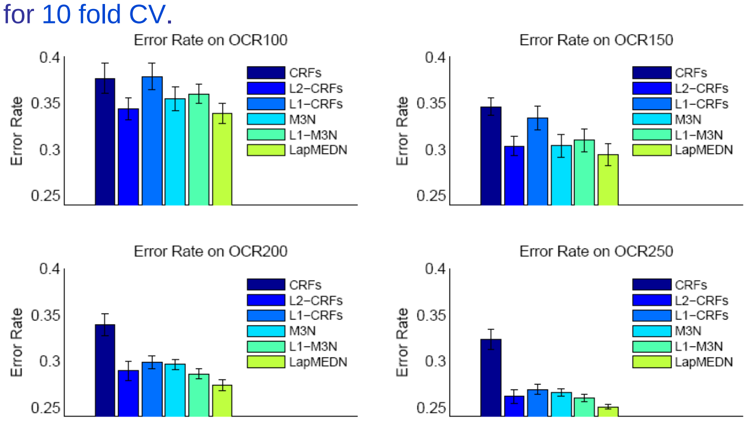
η_k

- Integratio e and Dis les
  - Incorporate latent variables and structures (PoMEN)
  - Semisupervised learning (with partially labeled data)

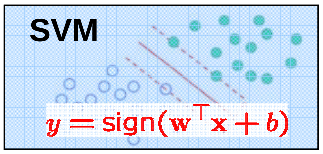
Experimental results on OCR datasets

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

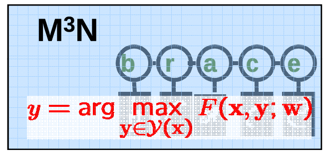
- We randomly construct OCR100, OCR150, OCR200, and OCR250 for 10 fold CV.



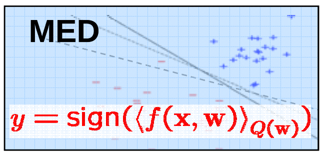
Discriminative Learning Paradigms



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



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



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

MED-MN

= SMED + Bayesian M³N

See [Zhu and Xing, 2008]

Summary

- Maximum margin nonlinear separator
  - Kernel trick
  - Project into linearly separatable space (possibly high or infinite dimensional)
  - No need to know the explicit projection function
- Max-entropy discrimination
  - Average rule for prediction,
  - Average taken over a posterior distribution of w who defines the separation hyperplane
  - P(w) is obtained by max-entropy or min-KL principle, subject to expected marginal constraints on the training examples
- Max-margin Markov network
  - Multi-variate, rather than uni-variate output Y
  - Variable in the outputs are not independent of each other (structured input/output)
  - Margin constraint over every possible configuration of Y (exponentially many!)