Probabilistic Graphical Models

Max-margin learning of GM

Eric Xing
Lecture 23, Apr 8, 2015

Reading:
Classical Predictive Models

- Input and output space: \( \mathcal{X} \triangleq \mathbb{R}^{M_x} \quad \mathcal{Y} \triangleq \{-1, +1\} \)

- Predictive function \( h(x) : y^* = h(x) \triangleq \arg\max_{y \in \mathcal{Y}} F(x, y; w) \)

- Examples:
  \[
  F(x, y; w) = g(w^T f(x, y))
  \]

- Learning:
  \[
  \hat{w} = \arg\min_{w \in \mathcal{W}} \ell(x, y; w) + \lambda R(w)
  \]
  where \( \ell(\cdot) \) represents a convex loss, and \( R(w) \) is a regularizer preventing overfitting

- **Logistic Regression**
  - Max-likelihood (or MAP) estimation
  \[
  \max_w \mathcal{L}(D; w) \triangleq \sum_{i=1}^{N} \log p(y^i|x^i; w) + \mathcal{N}(w)
  \]
  \[
  \ell_{LL}(x, y; w) \triangleq \ln \sum_{y' \in \mathcal{Y}} \exp\{w^T f(x, y')\} - w^T f(x, y)
  \]

- **Support Vector Machines (SVM)**
  - Max-margin learning
  \[
  \min_{w, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \\
  \text{s.t. } \forall i, \forall y' \neq y^i : w^T \Delta f_i(y') \geq 1 - \xi_i, \ \xi_i \geq 0.
  \]
  \[
  \ell_{MM}(x, y; w) \triangleq \max_{y' \in \mathcal{Y}} w^T f(x, y') - w^T f(x, y) + \ell'(y', y)
  \]

© Eric Xing @ CMU, 2005-2015
Classical Predictive Models

- Input and output space:  $X \triangleq \mathbb{R}^M_x$  $Y \triangleq \{-1, +1\}$

- Learning:  
  $$\hat{w} = \arg \min_{w \in \mathcal{W}} \ell(x, y; w) + \lambda R(w)$$

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

  - Logistic Regression
    - Max-likelihood (or MAP) estimation
    $\max_w \mathcal{L}(\mathcal{D}; w) \triangleq \sum_{i=1}^N \log p(y^i|x^i; w) + \mathcal{N}(w)$
    - Correlates to a Log loss with L2 R
    $\ell_{LL}(x, y; w) \triangleq \ln \sum_{y' \in \mathcal{Y}} \exp\{w^T f(x, y')\} - w^T f(x, y)$

  - Support Vector Machines (SVM)
    - Max-margin learning
    $$\min_{w, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i; \quad \text{s.t. } \forall i, \forall y' \neq y^i : w^T \Delta f_i(y') \geq 1 - \xi_i, \quad \xi_i \geq 0.$$  
    - Correlates to a hinge loss with L2 R
    $\ell_{MM}(x, y; w) \triangleq \max_{y' \in \mathcal{Y}} w^T f(x, y') - w^T f(x, y) + \ell(y', y)$

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

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

© Eric Xing @ CMU, 2005-2015
Structured Prediction Problem

- Unstructured prediction

\[ x = \begin{pmatrix} x_{11} & x_{12} & \ldots \end{pmatrix} \quad y = \begin{pmatrix} 0/1 \end{pmatrix} \]

- Structured prediction
  - Part of speech tagging
    \[ x = \text{“Do you want sugar in it?”} \quad \Rightarrow \quad y = \langle \text{verb pron verb noun prep pron} \rangle \]
  - Image segmentation

\[ x = \begin{pmatrix} x_{11} & x_{12} & \ldots \\ x_{21} & x_{22} & \ldots \\ \vdots & \vdots & \vdots \end{pmatrix} \quad y = \begin{pmatrix} y_{11} & y_{12} & \ldots \\ y_{21} & y_{22} & \ldots \\ \vdots & \vdots & \vdots \end{pmatrix} \]
OCR example

Sequential structure
Image Segmentation

- Jointly segmenting/annotating images
- Image-image matching, image-text matching
- Problem:
  - Given structure (feature), learning $\theta$
  - Learning sparse, interpretable, predictive structures/features

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

Challenge:
Structured outputs, and globally constrained to be a valid tree
# Structured Prediction Graphical Models

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conditional Random Fields (CRFs)</strong> (Lafferty et al 2001)</td>
<td>Based on a Logistic Loss (LR)</td>
</tr>
<tr>
<td></td>
<td>Max-likelihood estimation (point-estimate)</td>
</tr>
<tr>
<td>( \mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \log \sum_{\mathbf{y}'} \exp(\mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}')) )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-\mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}))</td>
</tr>
</tbody>
</table>

| **Max-margin Markov Networks (M³Ns)** (Taskar et al 2003) | Based on a Hinge Loss (SVM)                                                   |
|                                                        | Max-margin learning (point-estimate)                                      |
| \( \mathcal{L}(\mathcal{D}; \mathbf{w}) \triangleq \log \max_{\mathbf{y}'} \mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}') \)  |                                                                             |
|                                                        | \(-\mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}', \mathbf{y})\) | |

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

© Eric Xing @ CMU, 2005-2015
Structured Prediction Graphical Models

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

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

- **Max-margin Markov Networks (M3Ns)** (Taskar et al 2003)
  - Based on a Hinge Loss (SVM)
  - Max-margin learning (point-estimate)

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

**Challenges:**
- **SPARSE** “Interpretable” prediction model
- **Prior** information of structures
- **Latent** structures/variables
- **Time** series and non-stationarity
- **Scalable** to large-scale problems (e.g., $10^4$ input/output dimension)
Comparing to unstructured predictive models

- **Input and output space:** \( \mathcal{X} \triangleq \mathbb{R}^{M_x} \quad \mathcal{Y} \triangleq \{-1, 1\} \)

- **Learning:**
  \[
  \hat{w} = \arg \min_{w \in \mathcal{W}} \ell(x, y; w) + \lambda R(w)
  \]

  where \( \ell(\cdot) \) represents a convex loss, and \( R(w) \) is a regularizer preventing overfitting.

---

- **Logistic Regression**
  - Max-likelihood (or MAP) estimation
    \[
    \max_w \mathcal{L}(\mathcal{D}; w) \triangleq \sum_{i=1}^{N} \log p(y^i|x^i; w) + \mathcal{N}(w)
    \]
  - Corresponds to a Log loss with L2 R
    \[
    \ell_{LL}(x, y; w) \triangleq \ln \sum_{y' \in \mathcal{Y}} \exp\{w^T f(x, y')\} - w^T f(x, y)
    \]

- **Support Vector Machines (SVM)**
  - Max-margin learning
    \[
    \min_{w, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i;
    \]
    \[
    \text{s.t. } \forall i, \forall y' \neq y^i : w^T \Delta f_i(y') \geq 1 - \xi_i, \quad \xi_i \geq 0.
    \]
  - Corresponds to a hinge loss with L2 R
    \[
    \ell_{MM}(x, y; w) \triangleq \max_{y' \in \mathcal{Y}} w^T f(x, y') - w^T f(x, y) + \ell'(y', y)
    \]
Structured models

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

space of feasible outputs

Assumptions:

\[ score(x, y) = w^\top f(x, y) = \sum_p w^\top f(x_p, y_p) \]

linear combination of features

sum of part scores:
• index $p$ represents a part in the structure
Large Margin Estimation

- Given training example \((x, y^*)\), we want:
  \[ \arg \max_y w^T f(x, y) = y^* \]
  \[ w^T f(x, y^*) > w^T f(x, y) \quad \forall y \neq y^* \]
  \[ w^T f(x, y^*) \geq w^T f(x, y) + \gamma \ell(y^*, y) \quad \forall y \]

- Maximize margin \(\gamma\)
- Mistake weighted margin \(\gamma \ell(y^*, y)\)

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

*Taskar et al. 03*
Large Margin Estimation

- Recall from SVMs:
  - Maximizing margin $\gamma$ is equivalent to minimizing the square of the L2-norm of the weight vector $w$:

- New objective function:

$$
\min_w \frac{1}{2} \|w\|^2 \\
\text{s.t. } w^T f(x_i, y_i) \geq w^T f(x_i, y'_i) + \ell(y_i, y'_i), \quad \forall i, y'_i \in \mathcal{Y}_i
$$
OCR Example

- We want:
\[
\arg\max_{\text{word}} w^T f(\text{brace}, \text{word}) = \text{“brace”}
\]

- Equivalently:
\[
\begin{align*}
w^T f(\text{brace}, \text{“brace”}) &> w^T f(\text{brace}, \text{“aaaaa”}) \\
w^T f(\text{brace}, \text{“brace”}) &> w^T f(\text{brace}, \text{“aaaab”}) \\
\vdots \\
w^T f(\text{brace}, \text{“brace”}) &> w^T f(\text{brace}, \text{“zzzzz”})
\end{align*}
\]
Brute force enumeration of constraints:

\[
\min \frac{1}{2}||w||^2 \\
 w^T f(x, y^*) \geq w^T f(x, y) + \ell(y^*, y), \quad \forall y
\]

- The constraints are exponential in the size of the structure

Alternative: min-max formulation

- add only the most violated constraint

\[
y' = \arg \max_{y \neq y^*} [w^T f(x^i, y) + \ell(y^i, y)]
\]

add to QP: \[w^T f(x^i, y^i) \geq w^T f(x^i, y') + \ell(y^i, y')\]

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

\[
\min \quad \frac{1}{2} \|w\|^{2}
\]

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

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

\[
\max_{y \neq y^*} \ w^{\top} f(x, y) + \ell(y^*, y) \quad \leftrightarrow \quad \max_{z \in \mathcal{Z}} \ (F^{\top} w + \ell)^{\top} z
\]

**discrete optim.** \quad **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)

<table>
<thead>
<tr>
<th></th>
<th>z₁(m)</th>
<th>z₂(m)</th>
<th>z₃(m)</th>
<th>z₄(m)</th>
<th>z₅(m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>z₁₂(m, n)</th>
<th>z₂₃(m, n)</th>
<th>z₃₄(m, n)</th>
<th>z₄₅(m, n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0 1 . 0</td>
<td>0 0 . 0</td>
<td>0 1 . 0</td>
<td>0 0 . 0</td>
</tr>
<tr>
<td>B</td>
<td>0 0 . 0</td>
<td>1 0 . 0</td>
<td>0 0 . 0</td>
<td>0 1 . 0</td>
</tr>
<tr>
<td>:</td>
<td>. . . 0</td>
<td>. . . 0</td>
<td>. . . 0</td>
<td>. . . 0</td>
</tr>
<tr>
<td>B</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
</tr>
</tbody>
</table>
Rewriting the maximization function in terms of indicator variables:

\[
\begin{align*}
\max_z & \quad \sum_{j,m} z_j(m) \left[ w^\top f_{\text{node}}(x_j, m) + \ell_j(m) \right] \\
& \quad + \sum_{jk,m,n} z_{jk}(m, n) \left[ w^\top f_{\text{edge}}(x_{jk}, m, n) + \ell_{jk}(m, n) \right] \\
& \quad \text{subject to} \quad z_j(m) \geq 0; \quad z_{jk}(m, n) \geq 0; \\
& \quad \sum_m z_j(m) = 1 \\
& \quad \sum_n z_{jk}(m, n) = z_j(m) \\
\end{align*}
\]

\[
\begin{bmatrix}
0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]
Min-max formulation

- **Original problem:**
  \[
  \min \quad \frac{1}{2} ||w||^2 \\
  w^T f(x, y^*) \geq \max_y w^T f(x, y) + \ell(y^*, y)
  \]

- **Transformed problem:**
  \[
  \min \quad \frac{1}{2} ||w||^2 \\
  w^T f(x, y^*) \geq \max_{z \geq 0; \ A z = b} q^T z \quad \text{where} \quad q^T = w^T F + \ell^T
  \]

  - 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)

\[
\max_{\substack{z \geq 0; \\
A z = b;}} q^T z = \min_{\mu \geq q} b^T \mu
\]

- Use the result above to minimize jointly over \( w \) and \( \mu \):

\[
\min_{w, \mu} \frac{1}{2} ||w||^2 \\
\text{s.t. } w^T f(x, y^*) \geq b^T \mu; \\
A^T \mu \geq q;
\]
Min-max formulation

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

- 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

Error (average per-character)

45% error reduction over linear CRFs
33% error reduction over multiclass SVMs
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*
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)
  \[
  \min_{p(\Theta)} KL(p(\Theta)\|p_0(\Theta))
  \]
  \[
  \text{s.t.} \quad \int p(\Theta)[y_i F(x; \mathbf{w}) - \xi_i] \, d\Theta \geq 0, \forall i,
  \]

where \( \Theta \) is the parameter \( \mathbf{w} \) when \( \xi \) are kept fixed or the pair \( (\mathbf{w}, \xi) \) when we want to optimize over \( \xi \)
Maximum Entropy Discrimination Markov Networks

- Structured MaxEnt Discrimination (SMED):

\[ P_1 : \min_{p(w), \xi} KL(p(w) \| p_0(w)) + U(\xi) \]

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

generalized maximum entropy or regularized KL-divergence

- Feasible subspace of weight distribution:

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

expected margin constraints.

- Average from distribution of \( M^3 \)Ns

\[ h_1(x; p(w)) = \arg \max_{y \in \mathcal{Y}(x)} \int p(w) F(x, y; w) \, dw \]
Solution to MaxEnDNet

- **Theorem:**
  
  - **Posterior Distribution:**

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

  - **Dual Optimization Problem:**

    \[
    \text{D1 : } \max_{\alpha} - \log Z(\alpha) - U^*(\alpha)
    \]
    \[
    \text{s.t. } \alpha_i(y) \geq 0, \ \forall i, \ \forall y,
    \]

    \[U^*(\cdot) \text{ is the conjugate of the } U(\cdot), \text{ i.e., } U^*(\alpha) = \sup_{\xi} \left( \sum_{i,y} \alpha_i(y)\xi_i - U(\xi) \right)\]
Gaussian MaxEnDNet (reduction to $M^3N$)

- **Theorem**
  
  Assume

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

  $p(w) = \mathcal{N}(w|\mu_w, I)$, where $\mu_w = \sum_{i,y} \alpha_i(y) \Delta f_i(y)$

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

  s.t. $\sum_y \alpha_i(y) = C; \alpha_i(y) \geq 0, \forall i, \forall y$.

- Thus, MaxEnDNet subsumes $M^3N$s 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 (Theorem 3)

\[
P_{\mathcal{Q}}(M(h, x, y) \leq 0) \leq P_{\mathcal{D}}(M(h, x, y) \leq \gamma) + O\left(\sqrt{\frac{\gamma^{-2}KL(p||p_0)\ln(\mathcal{N}|\mathcal{Y}|) + \ln N + \ln \delta^{-1}}{N}}\right).
\]

- Entropy regularization: Introducing useful biases
  - Standard Normal prior => reduction to standard M^3N (we’ve seen it)
  - Laplace prior => Posterior shrinkage effects (sparse M^3N)

\[
\min_{\mu, \xi} \sqrt{\lambda} \sum_{k=1}^{K} \left(\sqrt{\mu_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}} \log \frac{\sqrt{\lambda} \mu_k^2 + 1 + 1}{2}\right) + C \sum_{i=1}^{N} \xi_i
\]

  s.t. \( \mu^T \Delta f_i(y) \geq \Delta f_i(y) - \xi_i; \quad \xi_i \geq 0, \quad \forall i, \forall y \neq y^i. \)

- Integrating Generative and Discriminative principles (next class)
  - Incorporate latent variables and structures (PoMEN)
  - Semisupervised learning (with partially labeled data)
Laplace MaxEnDNet (primal sparse M³N)  
(Zhu and Xing, ICML 2009)

- Laplace Prior:
  \[ p_0(w) = \prod_{k=1}^{K} \frac{\sqrt{\lambda}}{2} e^{-\sqrt{\lambda}|w_k|} = \left( \frac{\sqrt{\lambda}}{2} \right)^K e^{-\sqrt{\lambda}\|w\|} \]

- Corollary 4:
  - Under a Laplace MaxEnDNet, the posterior mean of parameter vector \( w \) is:
    \[ \forall k, \langle w_k \rangle_p = \frac{2\eta_k}{\lambda - \eta_k^2} \]
    where the vector \( \eta \) is a linear combination of "support vectors":
    \[ \eta = \sum_{i} \alpha_i(y)\Delta f_i(y) \]

- The Gaussian MaxEnDNet and the regular M³N has no such shrinkage
  - there, we have
    \[ \langle w \rangle_p = \eta \iff \forall k, \langle w_k \rangle_p = \eta_k \]
Corollary 5: LapMEDN corresponding to solving the following primal optimization problem:

$$\min_{\mu, \xi} |\mu| + C \sum_{i=1}^{N} \xi_i$$

s.t. $\mu^T \Delta f_i(y) \geq \Delta \ell_i(y) - \xi_i; \ \xi_i \geq 0, \ \forall i, \ \forall y \neq y^i.$

- KL norm:
  $$\|\mu\|_{KL} \triangleq \sum_{k=1}^{K} \left( \sqrt{\mu_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}} \log \frac{\sqrt{\lambda \mu_k^2 + 1} + 1}{2} \right)$$

L_1 and L_2 norms

KL norms
Recall Primal and Dual Problems of $\text{M}^3\text{Ns}$

- **Primal problem:**
  
  \[ \text{P0 (M}^3\text{N)} : \min_{\mathbf{w}, \xi} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{N} \xi_i \]
  
  s.t. \( \forall i, \forall \mathbf{y} \neq \mathbf{y}^i : \mathbf{w}^\top \Delta \mathbf{f}_i(y) \geq \Delta \ell_i(\mathbf{y}) - \xi_i, \quad \xi_i \geq 0 \),

- **Algorithms**
  - Cutting plane
  - Sub-gradient
  - ...

- **Dual problem:**
  
  \[ \text{D0 (M}^3\text{N)} : \max_{\alpha} \sum_{i,y} \alpha_i(\mathbf{y}) \Delta \ell_i(\mathbf{y}) - \frac{1}{2} \eta^\top \eta \]
  
  s.t. \( \forall i, \forall \mathbf{y} : \sum \alpha_i(\mathbf{y}) = C; \quad \alpha_i(\mathbf{y}) \geq 0. \)

  where \( \eta = \sum_{i,y} \alpha_i(\mathbf{y}) \Delta \mathbf{f}_i(y). \)

- **Algorithms:**
  - SMO
  - Exponentiated gradient
  - ...

\[ \mathbf{w}^* = \eta^* = \sum_{i,y} \alpha_i^*(\mathbf{y}) \Delta \mathbf{f}_i(\mathbf{y}). \]

- So, $\text{M}^3\text{N}$ is dual sparse!

\[ \mathbf{y}^* = h(\mathbf{x}) \triangleq \arg \max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}; \mathbf{w}) \]

© Eric Xing @ CMU, 2005-2015
Variational Learning of LapMEDN

- Exact primal or dual function is hard to optimize

\[
\min_{\mu, \xi} \sum_{k=1}^{K} \left( \sqrt{\mu_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}} \log \frac{\sqrt{\mu_k^2 + 1} + 1}{2} \right) + C \sum_{i=1}^{N} \xi_i
\]

s.t. \( \mu_i^T \Delta \xi_i(y) \geq \Delta \xi_i(y) - \xi_i; \xi_i \geq 0, \forall i, \forall y \neq y^i. \)

\[
\max_{\sum \alpha_i(y) = C} \sum_{i,y} \alpha_i(y) \Delta \xi_i(y) - \sum_{k=1}^{K} \log \frac{\lambda}{\lambda - \eta_k^2}
\]

- Use the hierarchical representation of Laplace prior, we get:

\[
KL(p||p_0) = -H(p) - \langle \log \int p(w|\tau)p(\tau|\lambda) \, d\tau \rangle_p
\]

\[
\leq -H(p) - \left\langle \int q(\tau) \log \frac{p(w|\tau)p(\tau|\lambda)}{q(\tau)} \, d\tau \right\rangle_p \triangleq \mathcal{L}(p(w), q(\tau))
\]

- We optimize an upper bound:

\[
\min_{p(w) \in \mathcal{F}_1; q(\tau); \xi} \mathcal{L}(p(w), q(\tau)) + U(\xi)
\]

- Why is it easier?

  - Alternating minimization leads to nicer optimization problems

<table>
<thead>
<tr>
<th>Keep (q(\tau)) fixed</th>
<th>Keep (p(w)) fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>- The effective prior is normal</td>
<td>- Closed form solution (\bar{q}(\tau)) and its expectation</td>
</tr>
<tr>
<td>(\forall k: p_0(w_k</td>
<td>\tau_k) = \mathcal{N}(w_k</td>
</tr>
</tbody>
</table>

An M^3N optimization problem! Closed-form solution!
Algorithmic issues of solving M³Ns

- **Primal problem:**
  \[
  \text{P0 (M³N)} : \min_{w, \xi} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i \\
  \text{s.t. } \forall i, \forall y \neq y^i : w^\top \Delta f_i(y) \geq \Delta \ell_i(y) - \xi_i, \\
  \xi_i \geq 0,
  \]

- **Algorithms**
  - Cutting plane
  - Sub-gradient
  - ...

- **Dual problem:**
  \[
  \text{D0 (M³N)} : \max_{\alpha} \sum_{i,y} \alpha_i(y) \Delta \ell_i(y) - \frac{1}{2} \eta^\top \eta \\
  \text{s.t. } \forall i, \forall y : \sum_{y} \alpha_i(y) = C; \ \alpha_i(y) \geq 0.
  \]
  where \( \eta = \sum_{i,y} \alpha_i(y) \Delta f_i(y). \)

- **Algorithms:**
  - SMO
  - Exponentiated gradient
  - ...

- **Nonlinear Features with Kernels**
  - Generative entropic kernels [Martins et al, JMLR 2009]
  - Nonparametric RKHS embedding of rich distributions [on going]

- **Approximate decoders for global features**
  - LP-relaxed Inference (polyhedral outer approx.) [Martins et al, ICML 09, ACL 09]
  - Balancing Accuracy and Runtime: Loss-augmented inference
Experimental results on OCR datasets

Structured output

© Eric Xing @ CMU, 2005-2015
Experimental results on OCR datasets

(CRFs, $L_1$–CRFs, $L_2$–CRFs, $M^3$Ns, $L_1$–$M^3$Ns, and LapMEDN)

- We randomly construct OCR100, OCR150, OCR200, and OCR250 for 10 fold CV.
Feature Selection
Sensitivity to Regularization Constants

- $L_1$-CRFs are much sensitive to regularization constants; the others are more stable
- LapM$^3$N is the most stable one

$L_1$-CRF and $L_2$-CRF:
- 0.001, 0.01, 0.1, 1, 4, 9, 16

M$^3$N and LapM$^3$N:
- 1, 4, 9, 16, 25, 36, 49, 64, 81
Summary:
Margin-based Learning Paradigms

**SVM**
\[
y = \text{sign}(w^T x + b)
\]
\[
\min_{w, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i;
\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \forall i.
\]

**MED**
\[
y = \text{sign}(\langle w | f(x) \rangle_{p(w)})
\]
\[
\min_{p, \xi} KL(p||p_0) + C \sum_{i=1}^{N} \xi_i;
\text{s.t. } y_i(f(x_i))_{p(w)} \geq 1 - \xi_i, \forall i.
\]

**Structured prediction**
\[
y^* = \arg \max_y w^T f(x, y; w)
\]
\[
\min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i;
\text{s.t. } w^T \Delta f_\xi(y) \geq \Delta f_\xi(y) - \xi_i \geq 0, \forall i, \forall y \neq y^i
\]

**MaxEnDNet**
\[
y^* = \arg \max_y (w^T f(x, y; w))_{p(w)}
\]
\[
\min_{p(w), \xi} KL(p||p_0) + U(\xi);
\text{s.t. } \int p(w)[\Delta f_\xi(y; w) - \Delta f_\xi(y)] dw \geq -\xi_i, \xi_i \geq 0, \forall i, \forall y \neq y^i.
\]
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?

\[
\text{margin} = w^T (F(y_n, x_n) - F(y'_n, x_n))
\]

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

- Algorithmic challenge
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:**
    - Online? Batch? Distributed?
  - **Evaluation:**
    - Visualization? Human interpretability? Perplexity? Predictive accuracy?
- It is better to consider one element at a time!