Handwriting Recognition 2

Graphical models: HMMs, MNs

Linear in length
Chain Markov Net (aka CRF*)

\[ P(y|x) = \frac{1}{Z(x)} \prod_i \phi(x_i, y_i) \prod_i \phi(y_i, y_{i+1}) \]

\[ \phi(x_i, y_i) = \exp\{\sum_\alpha w_\alpha f_\alpha(x_i, y_i)\} \]

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

\[ P(y|x) = \frac{1}{Z(x)} \prod_i \phi(x_i, y_i) \prod_i \phi(y_i, y_{i+1}) \]

\*Lafferty et al. 01

CRF - short notation

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

\[ \phi(x_i, y_i) = \exp\{\sum_\alpha w_\alpha f_\alpha(x_i, y_i)\} \]

\[ \phi(y_i, y_{i+1}) = \exp\{\sum_\beta w_\beta f_\beta(y_i, y_{i+1})\} \]
Max (Conditional) Likelihood

**Estimation**

\[ \max_{\theta} \sum_{x \in D} \log P_W(t(x) | x) \]

**Classification**

\[ \text{test case } x \]
\[ \arg \max_y P_W(y | x) \]

OCR Example

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

- Equivalently:
  \[ w^T f(\text{brace “brace”}) > w^T f(\text{brace “aaaaa”}) \]
  \[ w^T f(\text{brace, brace”}) > w^T f(\text{brace “aaaab”}) \]
  \[ \ldots \]
  \[ w^T f(\text{brace “brace”}) > w^T f(\text{brace “zzzzz”}) \]
Max Margin Estimation

- **Goal:** find \( w \) such that
  \[
  w^T f(x, t(x)) > w^T f(x, y) \quad \forall x \in D \quad \forall y \neq t(x)
  \]
  \[
  w^T [f(x, t(x)) - f(x, y)] > 0
  \]
  \[
  w^T \Delta f_x (y) > 0 \quad \forall x \in D, \forall y \neq t(x)
  \]
  \[
  \max_{y \neq t(x)} \gamma \quad w^T \Delta f_x (y) \geq \gamma \quad \text{margin}
  \]
  \[
  \forall x \in D, \forall y \neq t(x)
  \]
  \[
  ||w||^2 \leq 1
  \]

Not all margins are equal

- **Goal:** find \( w \) such that
  \[
  w^T \Delta f_x (y) \geq \gamma \quad \forall x \in D \quad \forall y \neq t(x)
  \]
  \[
  \text{if } y = t(x) \Rightarrow \text{rhs: } \gamma \cdot \Delta f_x (y) = 0
  \]
  \[
  \text{lhs: } w^T \Delta f_x (y) = w^T [f(x, t(x)) - f(x, y)] = \gamma
  \]

- Gain over \( y \) grows with # of mistakes in \( y \): \( \Delta t_x (y) \)
  \[
  \Delta t_{\text{craze}} \left(\text{"craze"}\right) = 2
  \]
  \[
  \Delta t_{\text{zzzzz}} \left(\text{"zzzzz"}\right) = 5
  \]
  \[
  w^T \Delta f_{\text{craze}} \left(\text{"craze"}\right) \geq 2 \gamma
  \]
  \[
  w^T \Delta f_{\text{zzzzz}} \left(\text{"zzzzz"}\right) \geq 5 \gamma
  \]
Maximum Margin Markov Nets

(Taskar, Guestrin, Koller '03)

\[
D = \{ x_1, t(x_1) \}
\]

\[
\ldots
\]

\[
x_m, t(x_m) \}
\]

\[
f(x,y)
\]

**Estimation**

\[
\max_{||w|| \leq 1} \gamma
w^T \Delta f_x(y) \geq \gamma \Delta t_x(y)
\]

**Classification**

\[
\arg \max_y w^T f(x,y)
\]

BTW. Just like SVMs, there are “non-linearly separable” cases, must add slack variables…

Handwriting Recognition

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

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

* Crammer & Singer 01
### Named Entity Recognition

- Locate and classify named entities in sentences:
  - 4 categories: organization, person, location, misc.
  - e.g. “U.N. official Richard Butler heads for Baghdad”.
- CoNLL 03 data set (200K words train, 50K words test)

\[ y_i = \text{org/per/loc/misc/none} \]

**Test F1**

- M^3N Linear
- M^3N Quad

### 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

**Test Error**

- SVMs
- RMNS
- M^3Ns

*Taskar et al. 02*
Solving M³Ns [Taskar, Guestrin, Koller '03]

Estimation

\[
\max_{\|w\| \leq 1} \gamma \quad w^T \Delta f_X(y) \geq \gamma \Delta t_X(y)
\]

Exponential size

Dual Quadratic Program

Factored Dual

Polynomial size

Other ways to solve M³Ns

- Sequential minimal optimization (SMO) [Taskar, Guestrin, Koller '03]
- Exponentiated gradient [Bartlett, Collins, Taskar, McAllester '04]
- Subgradient method [Ratliff, Bagnell, Zinkevich '07]
- …

Today

- Simple constraint generation
  - (Other methods will perform better in many practical problems)
  - (Other methods are better suited to adding kernels)
  - (Other methods use similar principles to simpler constraint generation)
Constraint generation overview

- Minimize $\mathbf{w}^T \mathbf{v}$
  - Subject to:
    - $\mathbf{w}^T f(\text{brace}) \geq \mathbf{w}^T f(\text{aaaaa}) + \Delta(\text{brace,aaaaa})$
    - $\ldots$
    - $\mathbf{w}^T f(\text{brace}) \geq \mathbf{w}^T f(\text{zzzzz}) + \Delta(\text{brace,zzzzz})$

  General form:
  $$\min_w \mathbf{w}^T \mathbf{w} \quad \text{s.t. } \mathbf{w}^T f(x_i + \epsilon(x_i)) \geq \mathbf{w}^T f(x_i) + \Delta x_i(y) \quad \forall \epsilon \neq 0$$

- Subset of constraints:
  $$\mathbf{w}^T f(x_i) \geq \mathbf{w}^T f(x_i + \epsilon) \quad \forall \epsilon \neq 0$$

- Constraint generation:
  - Solve for $w$ for a given set of $x_i$
  - Find a violated constraint: $\forall x_i

Generating a constraint
(simpler setting)

- Form of constraint
  $$\mathbf{w}^T f(\text{brace}) \geq \mathbf{w}^T f(\text{aaaaa}) + \Delta(\text{brace,aaaaa})$$

- Another way of expressing:
  $$\mathbf{w}^T f(x_i + \epsilon) \geq \mathbf{w}^T f(x_i) + \Delta x_i(y) \quad \forall \epsilon \neq 0$$

- Given $\mathbf{w}$, are any constraints violated?

- Separation oracle question:
  $$\arg\max_y [\mathbf{w}^T f(x_i) + \Delta x_i(y)]$$

- $\Delta(\text{brace,aaaaa})$ seems hard, simpler question:
  $$\arg\max_y [\mathbf{w}^T f(x_i) + \Delta x_i(y)]$$

- Exponentially many possibilities...
  $$\text{same question as at classification (testing)}$$
  $$\text{time \& variable}$$
  $$\text{inference for CM, e.g.}$$
Generating a constraint with hamming margin part ($\Delta(brace,aaaaa)$)

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

- Without $\Delta(brace,aaaaa)$: standard (MAP or MPE) inference in graphical models
  - Solve with dynamic programming
  - For chains, it’s called the Viterbi algorithm

- What do we do about $\Delta(brace,aaaaa)$?

$$\Delta f(y) = \sum_i \Delta \phi_i (y_i)$$

$$s = +$$

- Reformulation:

$$\prod_i \phi(x_i,y_i) \prod_i \phi(y_i,y_{i+1}) \exp\{\sum_i \Delta \phi_i (y_i)\}$$

$$\prod_i \phi(x_i,y_i) \prod_i \phi(y_i,y_{i+1}) = e^{\Delta f(y)}$$

- Same inference algorithm!!!
  - (slightly different potentials)

---

Overview of constraint generation for M3Ns

- Problem we want to solve:

$$\min \|w\|_1 = \min \sum_i w_i \Delta f(x_i,y_i) + \Delta f(x)$$

- Maintain subset of “runner up labels” for each training example:

$$\bigwedge_{x \in X} C \bigvee_{y \in Y} x$$

- Obtain some value for weights $w$

- Separation oracle:
  - Reformulate model to include hamming margin $\Delta(brace,aaaaa)$
  - Dynamic programming (inference in graphical models)
  - Apply to each data point
Some reasons $M^3Ns$ are cool… :)  

- Often perform better  
- Can use kernels easily, and get sparsity  
- Can be learned exactly in many problems where CRFs require approximate inference techniques  
  - E.g., image segmentation (graph cuts)  
- Can be generalized to other optimization problems  
  - E.g., [Taskar, Chatalbashev, Koller, Guestrin '05]  
  - Matching problems  
  - Paths  
  - Pretty much any optimization technique in the inner loop