Maximum Margin Markov Networks and Constraint Generation

Optimization - 10725
Carlos Guestrin
Carnegie Mellon University
February 11th, 2008

Handwriting Recognition

Character recognition: kernel SVMs
SVMs for sequences?  
Problem: # of classes exponential in length

Graphical models: HMMs, MNs
Linear in length
Chain Markov Net (aka CRF*)

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

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

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

\[
f_{\beta}(y, y') = I(y='z', y'='a')
\]

\[
f_{\alpha}(x, y) = I(x_p=1, y='z')
\]

CRF - short notation

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

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

\[
\phi(y_i y_{i+1}) = \exp\{\sum \beta w_{\beta} f_{\beta}(y_i y_{i+1})\}
\]
Pairwise Markov Nets

\[ P(y|x) = \frac{1}{Z(x)} \prod_i \phi_i(x,y_i) \prod_{ij} \phi_{ij}(x,y_i,y_j) \]

\[ = \frac{1}{Z(x)} \exp\{w^T f(x,y)\} \]

\[ w = [..., w_i, ..., w_{ij}, ...] \]

\[ f(x,y) = [..., f_i(x,y_i), ..., f_{ij}(x,y_i,y_j), ...] \]

Max (Conditional) Likelihood

\[ D \]

\[ x^1, t(x^1) \]

\[ ... \]

\[ x^m, t(x^m) \]

\[ f(x,y) \]

\[ \text{Estimation} \]

\[ \text{Classification} \]

\[ \text{maximize}_w \sum_{x \in D} \log P_w(t(x) | x) \]

\[ \arg \max_y P_w(y | x) \]
OCR Example

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

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

Max Margin Estimation

- Goal: find \( w \) such that
  \[ w^T f(x, t(x)) > w^T f(x, y) \quad \forall x \in D \quad \forall y \neq t(x) \]
  \[ w^T [f(x, t(x)) - f(x, y)] > 0 \]
  \[ w^T \Delta f_x(y) > 0 \]
Not all margins are equal

- **Goal:** find $w$ such that
  \[ w^\top \Delta f_x(y) \geq \gamma \quad \forall x \in D \quad \forall y = t(x) \]

- Gain over $y$ grows with # of mistakes in $y$: $\Delta t_x(y)$
  \[ \Delta t_x \text{ ("craze") } \quad \Delta t_x \text{ ("zzzzz") } \]
  \[ w^\top \Delta f_x \text{ ("craze") } \quad w^\top \Delta f_x \text{ ("zzzzz") } \]

Maximum Margin Markov Nets

(Taskar, Guestrin, Koller '03)

- **Estimation**
  \[ \max_{||w|| \leq 1} \quad \gamma \quad w^\top \Delta f_x(y) \geq \gamma \Delta t_x(y) \]

- **Classification**
  \[ \arg \max_y w^\top f(x, y) \]
Handwriting Recognition

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

Models:
- Multiclass-SVMs*
- CRFs
- M³ nets

*Crammer & Singer 01

![Graph showing test error for different models](image)

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)

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

$$f(y_i, x) = \ldots$$
- $$I(y_i=\text{org}, x_i=\text{"U.N."})$$
- $$I(y_i=\text{per}, x_i=\text{capitalized})$$
- $$I(y_i=\text{loc}, x_i=\text{known city})$$

![Graph showing test F1 scores for different models](image)
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

Solving $M^3Ns$

- Estimation
  - Exponential size
  - Polynomial size

- Dual Quadratic Program
- Factored Dual