6.864 (Fall 2007)

Global Linear Models: Part II

Overview

- Recap: global linear models
- Log-linear models for parameter estimation
- Global and local features
  - The perceptron revisited
  - Log-linear models revisited

Three Components of Global Linear Models

- $\mathbf{f}$ is a function that maps a structure (x, y) to a **feature vector** $\mathbf{f}(x, y) \in \mathbb{R}^d$
- $\mathbf{GEN}$ is a function that maps an input x to a set of **candidates** $\mathbf{GEN}(x)$
- $\mathbf{w}$ is a parameter vector (also a member of $\mathbb{R}^d$)
- Training data is used to set the value of $\mathbf{w}$

Putting it all Together

- $\mathcal{X}$ is set of sentences, $\mathcal{Y}$ is set of possible outputs (e.g. trees)
- Need to learn a function $F : \mathcal{X} \rightarrow \mathcal{Y}$
- $\mathbf{GEN}$, $\mathbf{f}$, $\mathbf{w}$ define

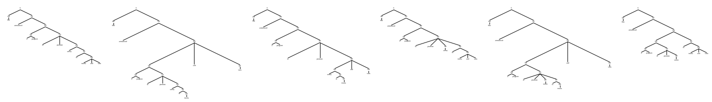
$$F(x) = \arg\max_{y \in \mathbf{GEN}(x)} \mathbf{f}(x, y) \cdot \mathbf{w}$$

Choose the highest scoring candidate as the most plausible structure

- Given examples (x_i, y_i), how to set $\mathbf{w}$?

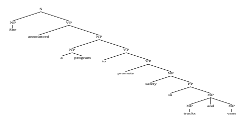
She announced a program to promote safety in trucks and vans

$\Downarrow$ **GEN**



$\Downarrow \mathbf{f}$	$\Downarrow \mathbf{f}$	$\Downarrow \mathbf{f}$	$\Downarrow \mathbf{f}$	$\Downarrow \mathbf{f}$	$\Downarrow \mathbf{f}$
$\langle 1, 1, 3, 5 \rangle$	$\langle 2, 0, 0, 5 \rangle$	$\langle 1, 0, 1, 5 \rangle$	$\langle 0, 0, 3, 0 \rangle$	$\langle 0, 1, 0, 5 \rangle$	$\langle 0, 0, 1, 5 \rangle$
$\Downarrow \mathbf{f} \cdot \mathbf{w}$	$\Downarrow \mathbf{f} \cdot \mathbf{w}$	$\Downarrow \mathbf{f} \cdot \mathbf{w}$	$\Downarrow \mathbf{f} \cdot \mathbf{w}$	$\Downarrow \mathbf{f} \cdot \mathbf{w}$	$\Downarrow \mathbf{f} \cdot \mathbf{w}$
13.6	12.2	12.1	3.3	9.4	11.1

$\Downarrow \arg\max$



A Variant of the Perceptron Algorithm

Inputs: Training set (x_i, y_i) for $i = 1 \ldots n$

Initialization: $\mathbf{w} = 0$

Define: $F(x) = \text{argmax}_{y \in \mathbf{GEN}(x)} \mathbf{f}(x, y) \cdot \mathbf{w}$

Algorithm: For $t = 1 \ldots T$, $i = 1 \ldots n$

$z_i = F(x_i)$

If $(z_i \neq y_i)$ $\quad \mathbf{w} = \mathbf{w} + \mathbf{f}(x_i, y_i) - \mathbf{f}(x_i, z_i)$

Output: Parameters $\mathbf{w}$

Overview

- Recap: global linear models
- Log-linear models for parameter estimation
- Global and local features
  - The perceptron revisited
  - Log-linear models revisited

Back to Maximum Likelihood Estimation [Johnson et. al 1999]

- We can use the parameters to define a probability for each parse:

$$P(y \mid x, \mathbf{w}) = \frac{e^{\mathbf{f}(x,y) \cdot \mathbf{w}}}{\sum_{y' \in \mathbf{GEN}(x)} e^{\mathbf{f}(x,y') \cdot \mathbf{w}}}$$

- Log-likelihood is then

$$L(\mathbf{w}) = \sum_i \log P(y_i \mid x_i, \mathbf{w})$$

- A first estimation method: take maximum likelihood estimates, i.e.,

$$\mathbf{w}_{ML} = \text{argmax}_{\mathbf{w}} L(\mathbf{w})$$

Adding Gaussian Priors [Johnson et. al 1999]

- A first estimation method: take maximum likelihood estimates, i.e., $\mathbf{w}_{ML} = \text{argmax}_{\mathbf{w}} L(\mathbf{w})$
- Unfortunately, very likely to "overfit"
- A way of preventing overfitting: choose parameters as

$$\mathbf{w}_{MAP} = \text{argmax}_{\mathbf{w}} \left(L(\mathbf{w}) - C \sum_k \mathbf{w}_k^2 \right)$$

for some constant C

- Intuition: adds a penalty for large parameter values

Summary

Choose parameters as:

$$\mathbf{w}_{MAP} = \text{argmax}_{\mathbf{w}} \left(L(\mathbf{w}) - C \sum_k \mathbf{w}_k^2 \right)$$

where

$$\begin{aligned} L(\mathbf{w}) &= \sum_i \log P(y_i \mid x_i, \mathbf{w}) \\ &= \sum_i \log \frac{e^{\mathbf{f}(x_i, y_i) \cdot \mathbf{w}}}{\sum_{y' \in \mathbf{GEN}(x_i)} e^{\mathbf{f}(x_i, y') \cdot \mathbf{w}}} \end{aligned}$$

Can use (conjugate) gradient ascent

(see previous lectures on log-linear models)

Overview

- Recap: global linear models
- Log-linear models for parameter estimation
- Global and local features
  - The perceptron revisited
  - Log-linear models revisited

Global and Local Features

- So far: algorithms have depended on size of $\mathbf{GEN}$
- Strategies for keeping the size of $\mathbf{GEN}$ manageable:
  - Reranking methods: use a baseline model to generate its top N analyses

Global and Local Features

- Global linear models are “global” in a couple of ways:
  - Feature vectors are defined over entire structures
  - Parameter estimation methods explicitly related to errors on entire structures
- Next topic: **global** training methods with **local features**
  - Our “global” features will be defined through *local* features
  - Parameter estimates will be global
  - **GEN** will be large!
  - Dynamic programming used for search and parameter estimation: **this is possible for some combinations of GEN and f**

Tagging Problems

TAGGING: Strings to Tagged Sequences

a b e e a f h j ⇒ a/C b/D e/C e/C a/D f/C h/D j/C

Example 1: Part-of-speech tagging

Profits/N soared/V at/P Boeing/N Co./N ,/, easily/ADV topping/V forecasts/N on/P Wall/N Street/N ,/, as/P their/POSS CEO/N Alan/N Mulally/N announced/V first/ADJ quarter/N results/N ./.

Example 2: Named Entity Recognition

Profits/NA soared/NA at/NA Boeing/SC Co./CC ,/NA easily/NA topping/NA forecasts/NA on/NA Wall/SL Street/CL ,/NA as/NA their/NA CEO/NA Alan/SP Mulally/CP announced/NA first/NA quarter/NA results/NA ./NA

Tagging

Going back to tagging:

- Inputs x are sentences $w_{[1:n]} = \{w_1 \dots w_n\}$
- **GEN**$(w_{[1:n]}) = \mathcal{T}^n$ i.e. all tag sequences of length n
- Note: **GEN** has an exponential number of members
- How do we define **f**?

Representation: Histories

- A **history** is a 4-tuple $\langle t_{-2}, t_{-1}, w_{[1:n]}, i\rangle$
- t_{-2}, t_{-1} are the previous two tags.
- $w_{[1:n]}$ are the n words in the input sentence.
- i is the index of the word being tagged

Hispaniola/NNP quickly/RB became/VB an/DT important/JJ base/?? from which Spain expanded its empire into the rest of the Western Hemisphere .

- $t_{-2}, t_{-1} =$ DT, JJ
- $w_{[1:n]} = \langle Hispaniola, quickly, became, \dots, Hemisphere, .\rangle$
- $i = 6$

Local Feature-Vector Representations

- Take a history/tag pair (h, t).
- $g_s(h, t)$ for $s = 1 \dots d$ are **local features** representing tagging decision t in context h.

Example: POS Tagging

- **Word/tag features**

$$g_{100}(h,t) = \begin{cases} 1 & \text{if current word } w_i \text{ is } \texttt{base} \text{ and } t = \texttt{VB} \\ 0 & \text{otherwise} \end{cases}$$

$$g_{101}(h,t) = \begin{cases} 1 & \text{if current word } w_i \text{ ends in } \texttt{ing} \text{ and } t = \texttt{VBG} \\ 0 & \text{otherwise} \end{cases}$$

- **Contextual Features**

$$g_{103}(h,t) = \begin{cases} 1 & \text{if } \langle t_{-2}, t_{-1}, t \rangle = \langle \text{DT, JJ, VB} \rangle \\ 0 & \text{otherwise} \end{cases}$$

A tagged sentence with n words has n history/tag pairs

Hispaniola/NNP quickly/RB became/VB an/DT important/JJ base/NN

History				Tag
t_{-2}	t_{-1}	$w_{[1:n]}$	i	t
*	*	$\langle Hispaniola, quickly, \dots, \rangle$	1	NNP
*	NNP	$\langle Hispaniola, quickly, \dots, \rangle$	2	RB
NNP	RB	$\langle Hispaniola, quickly, \dots, \rangle$	3	VB
RB	VB	$\langle Hispaniola, quickly, \dots, \rangle$	4	DT
VP	DT	$\langle Hispaniola, quickly, \dots, \rangle$	5	JJ
DT	JJ	$\langle Hispaniola, quickly, \dots, \rangle$	6	NN

A tagged sentence with n words has n history/tag pairs

Hispaniola/NNP quickly/RB became/VB an/DT important/JJ base/NN

History				Tag
t_{-2}	t_{-1}	$w_{[1:n]}$	i	t
*	*	$\langle Hispaniola, quickly, \dots, \rangle$	1	NNP
*	NNP	$\langle Hispaniola, quickly, \dots, \rangle$	2	RB
NNP	RB	$\langle Hispaniola, quickly, \dots, \rangle$	3	VB
RB	VB	$\langle Hispaniola, quickly, \dots, \rangle$	4	DT
VP	DT	$\langle Hispaniola, quickly, \dots, \rangle$	5	JJ
DT	JJ	$\langle Hispaniola, quickly, \dots, \rangle$	6	NN

Define global features through local features:

$$\mathbf{f}(t_{[1:n]}, w_{[1:n]}) = \sum_{i=1}^{n} \mathbf{g}(h_i, t_i)$$

where t_i is the i'th tag, h_i is the i'th history

Global and Local Features

- Typically, local features are indicator functions, e.g.,

$$g_{101}(h,t) = \begin{cases} 1 & \text{if current word } w_i \text{ ends in } \texttt{ing} \text{ and } t = \texttt{VBG} \\ 0 & \text{otherwise} \end{cases}$$

- and global features are then counts,

$f_{101}(w_{[1:n]}, t_{[1:n]})$ = Number of times a word ending in ing is tagged as VBG in $(w_{[1:n]}, t_{[1:n]})$

Putting it all Together

- $\mathbf{GEN}(w_{[1:n]})$ is the set of all tagged sequences of length n
- $\mathbf{GEN}$, $\mathbf{f}$, $\mathbf{w}$ define

$$
\begin{aligned}
F(w_{[1:n]}) &= \arg \max_{t_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} \mathbf{w} \cdot \mathbf{f}(w_{[1:n]}, t_{[1:n]}) \\
&= \arg \max_{t_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} \mathbf{w} \cdot \sum_{i=1}^{n} \mathbf{g}(h_i, t_i) \\
&= \arg \max_{t_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} \sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{g}(h_i, t_i)
\end{aligned}
$$

- Some notes:
  - Score for a tagged sequence is a sum of local scores
  - **Dynamic programming can be used to find the** argmax**!** (because history only considers the previous two tags)

A Variant of the Perceptron Algorithm

Inputs: Training set (x_i, y_i) for $i = 1 \ldots n$

Initialization: $\mathbf{w} = 0$

Define: $F(x) = \text{argmax}_{y \in \mathbf{GEN}(x)} \mathbf{f}(x, y) \cdot \mathbf{w}$

Algorithm: For $t = 1 \ldots T$, $i = 1 \ldots n$
$z_i = F(x_i)$
If $(z_i \neq y_i)$ $\quad \mathbf{w} = \mathbf{w} + \mathbf{f}(x_i, y_i) - \mathbf{f}(x_i, z_i)$

Output: Parameters $\mathbf{w}$

Training a Tagger Using the Perceptron Algorithm

Inputs: Training set $(w^i_{[1:n_i]}, t^i_{[1:n_i]})$ for $i = 1 \ldots n$.

Initialization: $\mathbf{w} = 0$

Algorithm: For $t = 1 \ldots T$, $i = 1 \ldots n$

$$z_{[1:n_i]} = \arg \max_{u_{[1:n_i]} \in \mathcal{T}^{n_i}} \mathbf{w} \cdot \mathbf{f}(w^i_{[1:n_i]}, u_{[1:n_i]})$$

$z_{[1:n_i]}$ can be computed with the dynamic programming (Viterbi) algorithm

If $z_{[1:n_i]} \neq t^i_{[1:n_i]}$ then

$$\mathbf{w} = \mathbf{w} \quad + \quad \mathbf{f}(w^i_{[1:n_i]}, t^i_{[1:n_i]}) - \mathbf{f}(w^i_{[1:n_i]}, z_{[1:n_i]})$$

Output: Parameter vector $\mathbf{w}$.

An Example

Say the correct tags for i'th sentence are

the/DT man/NN bit/VBD the/DT dog/NN

Under current parameters, output is

the/DT man/NN bit/NN the/DT dog/NN

Assume also that features track: (1) all bigrams; (2) word/tag pairs

Parameters incremented:

$\langle$NN, VBD$\rangle$, $\langle$VBD, DT$\rangle$, $\langle$VBD $\rightarrow$ bit$\rangle$

Parameters decremented:

$\langle$NN, NN$\rangle$, $\langle$NN, DT$\rangle$, $\langle$NN $\rightarrow$ bit$\rangle$

Experiments

- Wall Street Journal part-of-speech tagging data

  Perceptron = 2.89%, Max-ent = 3.28%
  (11.9% relative error reduction)

- [Ramshaw and Marcus, 1995] NP chunking data

  Perceptron = 93.63%, Max-ent = 93.29%
  (5.1% relative error reduction)

How Does this Differ from Log-Linear Taggers?

- Log-linear taggers (in an earlier lecture) used very similar *local representations*

- How does the perceptron model differ?

- Why might these differences be important?

Log-Linear Tagging Models

- Take a history/tag pair (h, t).

- $g_s(h, t)$ for $s = 1 \ldots d$ are **features**
  $\mathbf{w}_s$ for $s = 1 \ldots d$ are **parameters**

- Conditional distribution:

$$P(t|h) = \frac{e^{\mathbf{w}\cdot\mathbf{g}(h,t)}}{Z(h, \mathbf{w})}$$

  where $Z(h, \mathbf{w}) = \sum_{t' \in \mathcal{T}} e^{\mathbf{w}\cdot\mathbf{g}(h,t')}$

- Parameters estimated using maximum-likelihood

Log-Linear Tagging Models

- Word sequence $w_{[1:n]} = [w_1, w_2 \ldots w_n]$
- Tag sequence $t_{[1:n]} = [t_1, t_2 \ldots t_n]$
- Histories $h_i = \langle t_{i-1}, t_{i-2}, w_{[1:n]}, i \rangle$

$$\log P(t_{[1:n]} \mid w_{[1:n]}) = \sum_{i=1}^{n} \log P(t_i \mid h_i) = \underbrace{\sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{g}(h_i, t_i)}_{\text{Linear Score}} - \underbrace{\sum_{i=1}^{n} \log Z(h_i, \mathbf{w})}_{\text{Local Normalization Terms}}$$

- Compare this to the perceptron, where **GEN**, **f**, $\mathbf{w}$ define

$$F(w_{[1:n]}) = \arg \max_{t_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} \underbrace{\sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{g}(h_i, t_i)}_{\text{Linear score}}$$

Problems with Locally Normalized models

- "Label bias" problem [Lafferty, McCallum and Pereira 2001]
  See also [Klein and Manning 2002]
- Example of a conditional distribution that locally normalized models can't capture (under bigram tag representation):

$$\text{a b c} \Rightarrow \begin{array}{ccccc} \text{A} & — & \text{B} & — & \text{C} \\ | & & | & & | \\ \text{a} & & \text{b} & & \text{c} \end{array} \quad \text{with } P(\text{A B C} \mid \text{a b c}) = 1$$

$$\text{a b e} \Rightarrow \begin{array}{ccccc} \text{A} & — & \text{D} & — & \text{E} \\ | & & | & & | \\ \text{a} & & \text{b} & & \text{e} \end{array} \quad \text{with } P(\text{A D E} \mid \text{a b e}) = 1$$

- Impossible to find parameters that satisfy

$$P(A \mid a) \times P(B \mid b, A) \times P(C \mid c, B) = 1$$

$$P(A \mid a) \times P(D \mid b, A) \times P(E \mid e, D) = 1$$

Overview

- Recap: global linear models, and boosting
- Log-linear models for parameter estimation
- An application: LFG parsing
- Global and local features
  - The perceptron revisited
  - Log-linear models revisited

Global Log-Linear Models

- We can use the parameters to define a probability for each tagged sequence:

$$P(t_{[1:n]} \mid w_{[1:n]}, \mathbf{w}) = \frac{e^{\sum_i \mathbf{w} \cdot \mathbf{g}(h_i, t_i)}}{Z(w_{[1:n]}, \mathbf{w})}$$

  where

$$Z(w_{[1:n]}, \mathbf{w}) = \sum_{t_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} e^{\sum_i \mathbf{w} \cdot \mathbf{g}(h_i, t_i)}$$

  is a **global** normalization term

- This is a global log-linear model with

$$\mathbf{f}(w_{[1:n]}, t_{[1:n]}) = \sum_i \mathbf{g}(h_i, t_i)$$

Now we have:

$$\log P(t_{[1:n]} \mid w_{[1:n]})$$

$$= \underbrace{\sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{g}(h_i, t_i)}_{\text{Linear Score}} - \underbrace{\log Z(w_{[1:n]}, \mathbf{w})}_{\text{Global Normalization Term}}$$

When finding highest probability tag sequence, the global term is irrelevant:

$$\text{argmax}_{t_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} \sum_{i=1}^{n} \left(\mathbf{w} \cdot \mathbf{g}(h_i, t_i) - \log Z(w_{[1:n]}, \mathbf{w}) \right)$$

$$= \text{argmax}_{t_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} \sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{g}(h_i, t_i)$$

Parameter Estimation

- For parameter estimation, we must calculate the gradient of

$$\log P(t_{[1:n]} \mid w_{[1:n]}) = \sum_{i=1}^{n} \mathbf{w} \cdot \mathbf{g}(h_i, t_i) - \log \sum_{t'_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} e^{\sum_i \mathbf{w} \cdot \mathbf{g}(h'_i, t'_i)}$$

with respect to $\mathbf{w}$

- Taking derivatives gives

$$\frac{dL}{d\mathbf{w}} = \sum_{i=1}^{n} \mathbf{g}(h_i, t_i) - \sum_{t'_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} P(t'_{[1:n]} \mid w_{[1:n]}, \mathbf{w}) \sum_{i=1}^{n} \mathbf{g}(h'_i, t'_i)$$

- Can be calculated using dynamic programming!

Summary of Perceptron vs. Global Log-Linear Model

- Both are global linear models, where

$$\begin{aligned} \mathbf{GEN}(w_{[1:n]}) &= \text{the set of all possible tag sequences for } w_{[1:n]} \\ \mathbf{f}(w_{[1:n]}, t_{[1:n]}) &= \sum_i \mathbf{g}(h_i, t_i) \end{aligned}$$

- In both cases,

$$\begin{aligned} F(w_{[1:n]}) &= \text{argmax}_{t_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} \mathbf{w} \cdot \mathbf{f}(w_{[1:n]}, t_{[1:n]}) \\ &= \text{argmax}_{t_{[1:n]} \in \mathbf{GEN}(w_{[1:n]})} \sum_i \mathbf{w} \cdot \mathbf{g}(h_i, t_i) \end{aligned}$$

can be computed using dynamic programming

- Dynamic programming is also used in training:
  - Perceptron requires highest-scoring tag sequence for each training example
  - Global log-linear model requires gradient, and therefore "expected counts"

Results

From [Sha and Pereira, 2003]

- Task = shallow parsing (base noun-phrase recognition)

Model	Accuracy
SVM combination	94.39%
Conditional random field (global log-linear model)	94.38%
Generalized winnow	93.89%
Perceptron	94.09%
Local log-linear model	93.70%