Lecture 2: COMS E6998, Spring 2012

Log-Linear Models

Michael Collins

The Language Modeling Problem

- w_i is the i'th word in a document
- Estimate a distribution $P(w_i|w_1, w_2, \ldots w_{i-1})$ given previous "history" $w_1, \ldots, w_{i-1}$.
- E.g., $w_1, \ldots, w_{i-1} =$

Third, the notion "grammatical in English" cannot be identified in any way with the notion "high order of statistical approximation to English". It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical

A Second Example: Part-of-Speech Tagging

INPUT:

Profits soared at Boeing Co., easily topping forecasts on Wall Street, as their CEO Alan Mulally announced first quarter results.

OUTPUT:

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 ./.

N = Noun
V = Verb
P = Preposition
Adv = Adverb
Adj = Adjective
. . .

A Second Example: Part-of-Speech Tagging

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

- There are many possible tags in the position ??
{NN, NNS, Vt, Vi, IN, DT, ...}

- The task: model the distribution

$$P(t_i | t_1, \ldots, t_{i-1}, w_1 \ldots w_n)$$

where t_i is the i'th tag in the sequence, w_i is the i'th word

A Second Example: Part-of-Speech Tagging

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

- The task: model the distribution

$$P(t_i | t_1, \ldots, t_{i-1}, w_1 \ldots w_n)$$

where t_i is the i'th tag in the sequence, w_i is the i'th word

- Many "features" of $t_1, \ldots, t_{i-1}, w_1 \ldots w_n$ may be relevant

$$\begin{array}{lll} P(t_i = \text{NN} & | & w_i = \text{base}) \\ P(t_i = \text{NN} & | & t_{i-1} \text{ is } \texttt{JJ}) \\ P(t_i = \text{NN} & | & w_i \text{ ends in "e"}) \\ P(t_i = \text{NN} & | & w_i \text{ ends in "se"}) \\ P(t_i = \text{NN} & | & w_{i-1} \text{ is "important"}) \\ P(t_i = \text{NN} & | & w_{i+1} \text{ is "from"}) \end{array}$$

The General Problem

- We have some **input domain** $\mathcal{X}$

- Have a finite **label set** $\mathcal{Y}$

- Aim is to provide a **conditional probability** $P(y \mid x)$ for any x, y where $x \in \mathcal{X}$, $y \in \mathcal{Y}$

Language Modeling

- x is a "history" $w_1, w_2, \ldots w_{i-1}$, e.g.,

  Third, the notion "grammatical in English" cannot be identified in any way with the notion "high order of statistical approximation to English". It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical

- y is an "outcome" w_i

Feature Vector Representations

- Aim is to provide a conditional probability $P(y \mid x)$ for "decision" y given "history" x

- A **feature** is a function $\phi(x, y) \in \mathbb{R}$
(Often **binary features** or **indicator functions** $\phi(x, y) \in \{0, 1\}$).

- Say we have m features ϕ_k for $k = 1 \ldots m$
$\Rightarrow$ A **feature vector** $\underline{\phi}(x, y) \in \mathbb{R}^m$ for any x, y

Language Modeling

- x is a "history" $w_1, w_2, \ldots w_{i-1}$, e.g.,

  Third, the notion "grammatical in English" cannot be identified in any way with the notion "high order of statistical approximation to English". It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical

- y is an "outcome" w_i

- Example features:

$$\phi_1(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_2(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model} \text{ and } w_{i-1} = \texttt{statistical} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_3(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model},\ w_{i-2} = \texttt{any},\ w_{i-1} = \texttt{statistical} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_4(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model},\ w_{i-2} = \texttt{any} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_5(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model},\ w_{i-1} \text{ is an adjective} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_6(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model},\ w_{i-1} \text{ ends in ``ical''} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_7(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model}\text{, author = Chomsky} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_8(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model}\text{, ``model'' is not in } w_1, \ldots w_{i-1} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_9(x,y) = \begin{cases} 1 & \text{if } y = \texttt{model}\text{, ``grammatical'' is in } w_1, \ldots w_{i-1} \\ 0 & \text{otherwise} \end{cases}$$

Defining Features in Practice

- We had the following "trigram" feature:

$$\phi_3(x, y) = \begin{cases} 1 & \text{if } y = \texttt{model}, w_{i-2} = \texttt{any}, w_{i-1} = \texttt{statistical} \\ 0 & \text{otherwise} \end{cases}$$

- In practice, we would probably introduce one trigram feature for every trigram seen in the training data: i.e., for all trigrams (u, v, w) seen in training data, create a feature

$$\phi_{N(u,v,w)}(x, y) = \begin{cases} 1 & \text{if } y = w, w_{i-2} = u, w_{i-1} = v \\ 0 & \text{otherwise} \end{cases}$$

where $N(u, v, w)$ is a function that maps each (u, v, w) trigram to a different integer

The POS-Tagging Example

- Each x is a "history" of the form $\langle t_1, t_2, \ldots, t_{i-1}, w_1 \ldots w_n, i \rangle$
- Each y is a POS tag, such as $NN, NNS, Vt, Vi, IN, DT, \ldots$
- We have m features $\phi_k(x, y)$ for $k = 1 \ldots m$

For example:

$$\phi_1(x, y) = \begin{cases} 1 & \text{if current word } w_i \text{ is } \texttt{base} \text{ and } y = \texttt{Vt} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_2(x, y) = \begin{cases} 1 & \text{if current word } w_i \text{ ends in } \texttt{ing} \text{ and } y = \texttt{VBG} \\ 0 & \text{otherwise} \end{cases}$$

$$\ldots$$

The Full Set of Features in [?]

- Word/tag features for all word/tag pairs, e.g.,

$$\phi_{100}(x, y) = \begin{cases} 1 & \text{if current word } w_i \text{ is } \texttt{base} \text{ and } y = \texttt{Vt} \\ 0 & \text{otherwise} \end{cases}$$

- Spelling features for all prefixes/suffixes of length ≤ 4, e.g.,

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

$$\phi_{102}(h, t) = \begin{cases} 1 & \text{if current word } w_i \text{ starts with } \texttt{pre} \text{ and } y = \texttt{NN} \\ 0 & \text{otherwise} \end{cases}$$

The Full Set of Features in [?]

- Contextual Features, e.g.,

$$\phi_{103}(x,y) = \begin{cases} 1 & \text{if } \langle t_{i-2}, t_{i-1}, y \rangle = \langle \text{DT, JJ, Vt} \rangle \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{104}(x,y) = \begin{cases} 1 & \text{if } \langle t_{i-1}, y \rangle = \langle \text{JJ, Vt} \rangle \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{105}(x,y) = \begin{cases} 1 & \text{if } \langle y \rangle = \langle \text{Vt} \rangle \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{106}(x,y) = \begin{cases} 1 & \text{if previous word } w_{i-1} = \textit{the} \text{ and } y = \text{Vt} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{107}(x,y) = \begin{cases} 1 & \text{if next word } w_{i+1} = \textit{the} \text{ and } y = \text{Vt} \\ 0 & \text{otherwise} \end{cases}$$

The Final Result

- We can come up with practically any questions (*features*) regarding history/tag pairs.

- For a given history $x \in \mathcal{X}$, each label in $\mathcal{Y}$ is mapped to a different feature vector

$$
\begin{aligned}
\underline{\phi}(\langle \text{JJ, DT}, \langle \text{ Hispaniola}, \ldots \rangle, 6\rangle, \text{Vt}) &= 1001011001001100110 \\
\underline{\phi}(\langle \text{JJ, DT}, \langle \text{ Hispaniola}, \ldots \rangle, 6\rangle, \text{JJ}) &= 0110010101011110010 \\
\underline{\phi}(\langle \text{JJ, DT}, \langle \text{ Hispaniola}, \ldots \rangle, 6\rangle, \text{NN}) &= 0001111101001100100 \\
\underline{\phi}(\langle \text{JJ, DT}, \langle \text{ Hispaniola}, \ldots \rangle, 6\rangle, \text{IN}) &= 0001011011000000010 \\
&\ldots
\end{aligned}
$$