Processamento Estatístico da Linguagem Natural

Aula 16

Professora Bianca
(Sala 302 – Bloco E)
bianca@ic.uff.br
http://www.ic.uff.br/~bianca/peln/

Aula de Hoje

- Cap. 6 – Jurafsky & Martin – Hidden Markov and Maximum Entropy Models
  - Seção 6.7 e 6.8

Modelo de Entropia Máxima

- Regressão logística com mais de 2 classes = regressão logística multinomial.
  - Também conhecida como Modelo de Entropia Máxima (MaxEnt).
  - A equação do MaxEnt é:

$$p(c|x) = \frac{exp\left(\sum_{i=0}^{N} w_{ci} f_i\right)}{\sum_{c' \in C} exp\left(\sum_{i=0}^{N} w_{c'i} f_i\right)}$$

  - O denominador é um fator de normalização comumente chamado de Z:

$$p(c|x) = \frac{1}{Z} exp \sum_{i=0}^{N} w_{ci} f_i$$

Função indicadora

- Em processamento de linguagem é comum usar atributos binários (0 ou 1).
  - Também chamados de função indicadora.
  - Indicam a presença ou ausência de alguma propriedade da observação e da classe em questão.
  - Usamos a notação $f_i(c,x)$.
    - Atributo i para a classe c e observação x.
- MaxEnt com função indicadora:

$$p(c|x) = \frac{\exp\left(\sum_{i=0}^{N} w_{ci} f_i(c,x)\right)}{\sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i(c',x)\right)}$$

Exemplo: Rotulação POS

- Rotular a palavra "race" em

  Secretariat/NNP is/BEZ expected/VBN to/TO **race**/?? tomorrow/

- Atributos úteis

$$f_1(c,x) = \begin{cases} 1 & \text{if } word_i = \text{"race"} \ \& \ c = \text{NN} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(c,x) = \begin{cases} 1 & \text{if } t_{i-1} = \text{TO} \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

$$f_3(c,x) = \begin{cases} 1 & \text{if suffix}(word_i) = \text{"ing"} \ \& \ c = \text{VBG} \\ 0 & \text{otherwise} \end{cases}$$

$$f_4(c,x) = \begin{cases} 1 & \text{if is\_lower\_case}(word_i) \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

Exemplo: Rotulação POS

- Atributos úteis (cont.)

$$f_5(c,x) = \begin{cases} 1 & \text{if } word_i = \text{"race"} \ \& \ c = \text{VB} \\ 0 & \text{otherwise} \end{cases}$$

$$f_6(c,x) = \begin{cases} 1 & \text{if } t_{i-1} = \text{TO} \ \& \ c = \text{NN} \\ 0 & \text{otherwise} \end{cases}$$

- Alguns valores e pesos (para a palavra "race")

		f1	f2	f3	f4	f5	f6
VB	f	0	1	0	1	1	0
VB	w		.8		.01	.1	
NN	f	1	0	0	0	0	1
NN	w	.8					-1.3

Exemplo: Rotulação POS

- Cálculo usando a equação do MaxEnt

$$P(NN|x) = \frac{e^{.8}e^{-1.3}}{e^{.8}e^{-1.3} + e^{.8}e^{.01}e^{.1}} = .20$$

$$P(VB|x) = \frac{e^{.8}e^{.01}e^{.1}}{e^{.8}e^{-1.3} + e^{.8}e^{.01}e^{.1}} = .80$$

Atributos "complexos"

- MaxEnt não considera automaticamente interações entre os atributos.
- Mas podemos criar atributos "complexos" usando expressões booleanas.
  - Exemplo:

$$f_{125}(c,x) = \begin{cases} 1 & \text{if } word_{i-1} = \texttt{<s>} \ \& \ \text{isupperfirst}(word_i) \ \& \ c = \text{NNP} \\ 0 & \text{otherwise} \end{cases}$$

- O sucesso do modelo dependerá da criação de atributos apropriados.

Regularização

- O aprendizado do MaxEnt segue a mesma idéia do aprendizado de regressão logística.
  - Encontrar os pesos que melhor expliquem os dados.
  - Porém para generalizar para novos dados precisamos restringir o modelo de alguma forma.
    - Adicionamos um termo que faz com que os pesos fiquem próximos de zero = regularização.

$$\hat{w} = \underset{w}{\text{argmax}} \sum_i \log P(y^{(i)}|x^{(i)}) - \alpha \sum_{j=1}^{N} w_j^2$$

Por que máxima entropia?

- O modelo MaxEnt também pode ser formulado da seguinte maneira:
  - Encontre todas as distribuições de probabilidade que obedeçam às restrições impostas pelos atributos.
  - Entre essas distribuições escolha a de máxima entropia.
    - É o modelo que não faz nenhuma hipótese além dos atributos.

Modelos de Markov de Entropia Máxima

- Idéia: transformar MaxEnt em um classificador de seqüências.
- Quais seriam as vantagens em relação ao HMM?
  - HMM só lida com probabilidades do tipo P(rótulo|rótulo) ou P(palavra|rótulo).
  - MaxEnt pode lidar com outros tipos de informação através dos atributos P(maiúscula|rótulo), P(sufixo|rótulo), P(hífen|rótulo)...

Como usar MaxEnt para seqüências

- Idéia simples: aplicar o classificador para cada elemento da seqüência, indo da esquerda para direita.
  - Problema: não saberemos o rótulo do elemento seguinte, o que poderia ser útil para classificação.
- Ao invés disso, podemos usar uma combinação do algoritmo de Viterbi com o MaxEnt.

HMM vs. MEMM

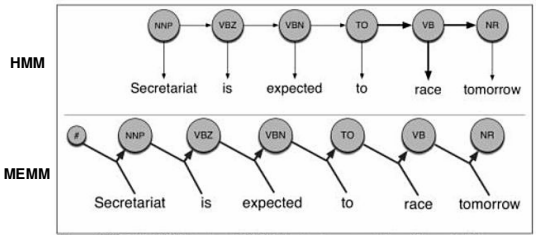
- HMM – Calculamos P(W|T)P(T) para obter P(T|W).

$$\begin{aligned}\hat{T} &= \underset{T}{\operatorname{argmax}} P(T|W) \\ &= \underset{T}{\operatorname{argmax}} P(W|T)P(T) \\ &= \underset{T}{\operatorname{argmax}} \prod_i P(word_i|tag_i) \prod_i P(tag_i|tag_{i-1})\end{aligned}$$

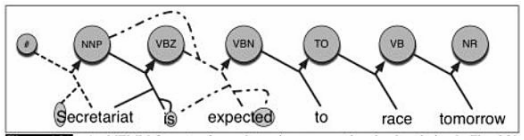
- MEMM – Calculamos P(T|W) diretamente.

$$\begin{aligned}\hat{T} &= \underset{T}{\operatorname{argmax}} P(T|W) \\ &= \underset{T}{\operatorname{argmax}} \prod_i P(tag_i|word_i, tag_{i-1})\end{aligned}$$

HMM vs. MEMM



MEMM permite o uso de atributos



HMM vs. MEMM

- HMM

$$P(Q|O) = \prod_{i=1}^{n} P(o_i|q_i) \times \prod_{i=1}^{n} P(q_i|q_{i-1})$$

- MEMM

$$P(Q|O) = \prod_{i=1}^{n} P(q_i|q_{i-1}, o_i)$$

$$P(q|q', o) = \frac{1}{Z(o, q')} \exp\left(\sum_i w_i f_i(o, q)\right)$$

Viterbi para o MEMM

- O algoritmo Viterbi é usado para "decodificar", isto é, calcular a seqüência de rótulos de maior probabilidade.
- Para o HMM

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, P(s_j|s_i)\, P(o_t|s_j) \quad 1 \le j \le N, 1 < t \le T$$

- Para o MEMM

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, P(s_j|s_i, o_t) \quad 1 \le j \le N, 1 < t \le T$$

Treliça do Viterbi para o MEMM