

Processamento Estatístico da Linguagem Natural

Aula 13

Professora Bianca
(Sala 302 – Bloco E)
bianca@ic.uff.br

<http://www.ic.uff.br/~bianca/peln/>

Aula 13 - 16/10/2008

1

Aula de Hoje

- Cap. 6 – Jurafsky & Martin – Hidden Markov and Maximum Entropy Models
 - Seções 6.1, 6.2 e 6.3

Aula 13 - 16/10/2008

2

HMMs e MEMMs

- O capítulo apresenta dois modelos estatísticos importantes para processar linguagem:
 - HMM = **H**idden **M**arkov **M**odel = Modelo de Markov Oculto
 - MEMM = **M**aximum **E**ntropy **M**arkov **M**odel = Modelo de Markov de Entropia Máxima
- Ambos são classificadores seqüenciais probabilísticos.
 - O transdutor de estado finito (cap. 3) é um classificador seqüencial determinístico.

Aula 13 - 16/10/2008

3

Classificação Seqüencial

- Entrada: seqüência de elementos (palavras, letras, morfemas, frases, etc.)
- Saída: seqüência de rótulos para cada elemento da entrada.
- Exemplos de aplicação:
 - Rotulação POS
 - Reconhecimento de voz
 - Segmentação de frases
 - Extração de informação

Aula 13 - 16/10/2008

4

Cadeias de Markov

- São um tipo de autômato finito com pesos.
 - Cada aresta é associada com a probabilidade.
 - A soma das probabilidades das arestas saindo de um nó deve ser 1.
 - A seqüência de entrada determina unicamente a saída.
 - Não há ambigüidade.

Aula 13 - 16/10/2008

5

Especificação Formal

- Uma cadeia de markov é formada pelos seguintes componentes:

$Q = q_1 q_2 \dots q_N$ a set of **states**
 $A = a_{01} a_{02} \dots a_{n1} \dots a_{nn}$ a **transition probability matrix** A , each a_{ij} representing the probability of moving from state i to state j , s.t. $\sum_{j=1}^n a_{ij} = 1 \quad \forall i$
 q_0, q_{end} a special **start and end state** which are not associated with observations.

- Cada $a_{ij} = P(q_j | q_i)$

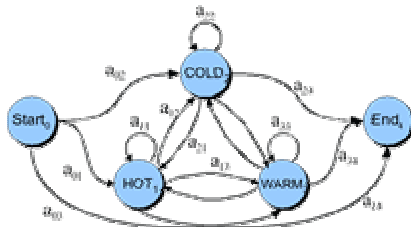
Markov Assumption: $P(q_i | q_1 \dots q_{i-1}) = P(q_i | q_{i-1})$

Aula 13 - 16/10/2008

6

Exemplo de Cadeia de Markov

- Cadeia de Markov que atribui probabilidades a seqüências de eventos meteorológicos.

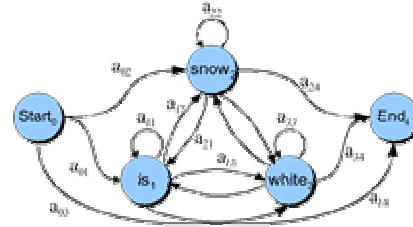


Aula 13 - 16/10/2008

7

Outra cadeia de Markov

- Cadeia de Markov representando um modelo de bigramas.



Aula 13 - 16/10/2008

8

Outra representação para cadeias de Markov

- Ao invés de ter estados iniciais e finais únicos, podemos ter:

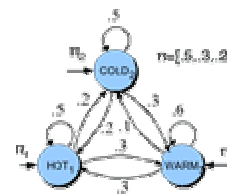
$\pi = \pi_1, \pi_2, \dots, \pi_N$ an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i . Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^N \pi_i = 1$

$QA = \{q_1, q_2, \dots\}$ a set $QA \subset Q$ of legal **accepting states**

Aula 13 - 16/10/2008

9

Um exemplo com números



- Qual é a probabilidade de:
 - Hot hot hot hot
 - Cold hot cold hot

Aula 13 - 16/10/2008

10

Modelos de Markov Ocultos

- Uma cadeia de Markov é útil quando queremos calcular a probabilidade de uma seqüência de eventos conhecidos.
- O que fazer quando queremos calcular a probabilidade de eventos não-observáveis?
 - Por exemplo, rótulos POS.
- Modelos de Markov Ocultos (HMMs) modelam conjuntamente eventos observáveis e eventos ocultos.

Aula 13 - 16/10/2008

11

Modelos de Markov Ocultos

- $Q = q_1 q_2 \dots q_N$ a set of **states**
- $A = a_{01} a_{02} \dots a_{n1} \dots a_{nm}$ a **transition probability matrix** A , each a_{ij} representing the probability of moving from state i to state j , s.t. $\sum_{j=1}^n a_{ij} = 1 \quad \forall i$
- $O = o_1 o_2 \dots o_N$ a set of **observations**, each one drawn from a vocabulary $V = v_1, v_2, \dots, v_V$.
- $B = b_1(o_1)$ A set of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_i being generated from a state i .
- q_0, q_{end} a special **start and end state** which are not associated with observation.

Aula 13 - 16/10/2008

12

Modelos de Markov Ocultos

- Também é possível usar a representação alternativa com distribuição inicial de estados e um conjunto de estados finais.
- O modelo de Markov faz duas hipóteses simplificadoras:

Markov Assumption: $P(q_i | q_1 \dots q_{i-1}) = P(q_i | q_{i-1})$

Output Independence Assumption: $P(o_1 | q_1 \dots q_T, o_2, \dots, o_n, o_1, \dots, o_T) = P(o_i | q_i)$

Aula 13 - 16/10/2008

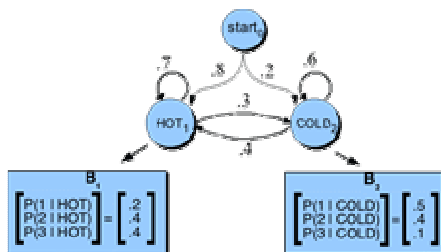
13

Exemplo: A tarefa de Jason Eisner

- Um climatologista em 2799 estudando a história do aquecimento global
- Não há registros de temperatura em Baltimore no verão de 2008.
- Mas existe o diário de Jason Eisner marcando quantos sorvetes ele tomou em cada dia.
- O climatologista pode usar essa informação para estimar a temperatura?
- Dada uma sequência de observações O ,
 - Cada observação é um inteiro = número de sorvetes tomados
 - Descobrir a sequência oculta correta Q de estados de temperatura (H or C) que levaram Jason a tomar os sorvetes.

Aula 13 - 16/10/2008

14



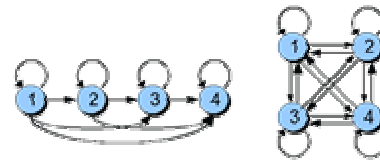
Aula 13 - 16/10/2008

15

Modelos de Markov Ocultos

Rede de Bakis
(esquerda-para-direita)

Rede Ergódica
(totalmente conectada)



Arcos que não aparecem têm probabilidade zero.

Aula 13 - 16/10/2008

16

Três Problemas Fundamentais de HMMs

- Avaliação:** Dado um HMM determinar a probabilidade de uma dada sequência de observações.
- Decodificação:** Dada uma sequência de observações e um HMM, descobrir a melhor (mais provável) sequência de estados ocultos.
- Aprendizado:** Dada uma sequência de observações e o conjunto de estados do HMM, aprender os parâmetros do HMM (matrizes A e B.)

Aula 13 - 16/10/2008

17

Três Problemas Fundamentais de HMMs – Mais formalmente

- Avaliação:** Dada a sequência de observações $O=(o_1, o_2, \dots, o_T)$, e um HMM $\Phi = (A, B)$, **como podemos calcular eficientemente** $P(O | \Phi)$, a probabilidade da sequência de observações dado o modelo.
- Decodificação:** Dada a sequência de observações $O=(o_1, o_2, \dots, o_T)$, e um HMM $\Phi = (A, B)$, **como escolhemos uma sequência de estados correspondentes** $Q=(q_1, q_2, \dots, q_T)$ que seja ótima em algum sentido (i.e., que melhor explica as observações)
- Aprendizado:** **Como ajustamos os parâmetros do modelo** $\Phi = (A, B)$ para maximizar $P(O | \Phi)$?

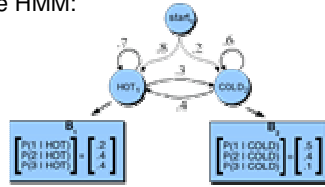
Aula 13 - 16/10/2008

18

Problema 1: calcular a probabilidade de uma observação

Computing Likelihood: Given an HMM $\lambda = (A, B)$ and an observation sequence O , determine the likelihood $P(O|\lambda)$.

- Dado seguinte HMM:



- Qual é a probabilidade da sequência 3 1 3?

Aula 13 - 16/10/2008

19

Como calcular a probabilidade

- Para uma cadeia de Markov, apenas seguimos os estados correspondentes a 3 1 3 e multiplicamos as probabilidades.
- Mas para um HMM, não sabemos quais são os estados!
- Então começaremos com uma situação mais simples.
- Calcular a probabilidade de observação **dada** uma sequência de estados.
 - Supondo que sabemos a temperatura e queremos prever quantos sorvetes Jason vai tomar.
 - I.e. $P(3 \ 1 \ 3 \mid H \ H \ C)$

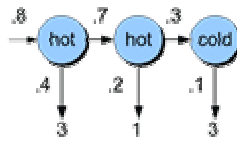
Aula 13 - 16/10/2008

20

Calculando a probabilidade dada uma sequência de estados

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

$$P(3 \ 1 \ 3 \mid \text{hot hot cold}) = P(\text{hot}|\text{start}) \times P(\text{hot}|\text{hot}) \times P(\text{cold}|\text{hot}) \times P(3|\text{hot}) \times P(1|\text{hot}) \times P(3|\text{cold})$$



Aula 13 - 16/10/2008

21

Calculando a probabilidade total

- Teríamos que somar as probabilidades da observação dada cada sequência de estados $P(O) = \sum_Q P(O|Q)P(Q)$
 - Hot hot cold
 - Hot hot hot
 - Hot cold hot
 -
- Quantas possíveis sequências de estados existem?
- Para um HMM com N estados e uma sequência com T observações, existem N^T sequências possíveis.
- Então seria muito ineficiente fazer o cálculo separadamente para cada sequência.

Aula 13 - 16/10/2008

22

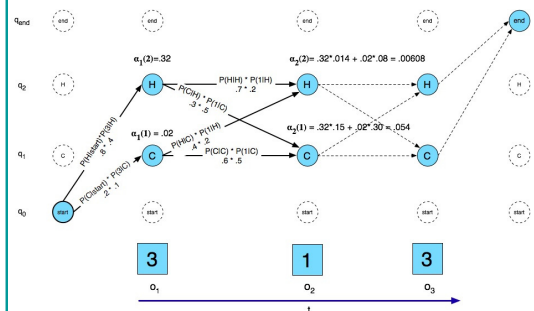
O Algoritmo Forward (Pra frente)

- Um algoritmo que usa programação dinâmica.
 - Usa uma tabela pra guardar valores intermediários.
- Idéia:
 - Calcular a probabilidade da sequência observada.
 - Somando sobre todas as sequências de estados.
 - Fazendo isso de forma eficiente.
 - Colocando todas as sequências em uma mesma **treliça**.

Aula 13 - 16/10/2008

23

A Treliça do Algoritmo Forward



Aula 13 - 16/10/2008

24

O Algoritmo Forward

- Cada célula da treliça do algoritmo forward $\alpha_t(j)$
 - Representa a probabilidade de estar no estado j depois de t observações, dado o autômato.
- Cada célula expressa a seguinte probabilidade:

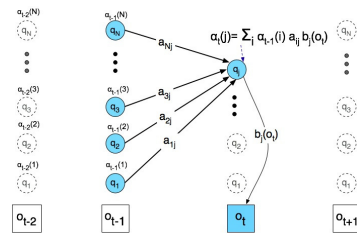
$$\alpha_t(j) = P(o_1, o_2 \dots o_t, q_t = j | \lambda)$$

Aula 13 - 16/10/2008

25

Cada célula é atualizada

$\alpha_{t-1}(i)$ the previous forward path probability from the previous time step
 a_{ij} the transition probability from previous state q_i to current state q_j
 $b_j(o_t)$ the state observation likelihood of the observation symbol o_t , given the current state j



Aula 13 - 16/10/2008

26

A recursão pra frente

- Initialization:

$$\alpha_1(j) = a_{0j} b_j(o_1) \quad 1 \leq j \leq N$$

- Recursion (since states 0 and N are non-emitting):

$$\alpha_t(j) = \left[\sum_{i=1}^{N-1} \alpha_{t-1}(i) a_{ij} \right] b_j(o_t); \quad 1 < j < N, 1 < t < T$$

- Termination:

$$P(O|\lambda) = \alpha_T(N) = \sum_{i=2}^{N-1} \alpha_T(i) a_{iN}$$

Aula 13 - 16/10/2008

27

O Algoritmo Forward

function FORWARD(observations of len T , state-graph) **returns** forward-probability

$num_states \leftarrow \text{NUM-OF-STATES}(\text{state-graph})$

Create a probability matrix $forward[num_states+2, T+2]$

$forward[0,0] \leftarrow 1.0$

for each time step t **from** 1 **to** T **do**

for each state s **from** 1 **to** num_states **do**

$forward[s,t] \leftarrow \sum_{1 \leq i \leq num_states} [forward[i',t-1] * a_{i',s}] * b_s(o_t)$

return the sum of the probabilities in the final column of $forward$

Aula 13 - 16/10/2008

28