MAC 499 – TRABALHO DE FORMATURA SUPERVISIONADO

Antônio Galves
Arnaldo Mandel
Denis Antônio Lacerda

Florestas Probabilísticas para Discriminar Português Brasileiro e Europeu

- **Seqüências genéticas, cadeias de aminoácidos, seqüências rítmicas na fala, seqüências de dados econômicos, parecem ter em comum:**
  - Um comportamento que, apesar de não ser determinístico, contém informações precisas a respeito do sistema que as produziu
  - No caso de cadeias lingüísticas uma dessas características parece estar codificada no *ritmo*

A conjectura das classes rítmicas

- Lloyd James (anos 40) e Abercrombie (anos 50) conjecturam que as línguas se agrupam em classes rítmicas.
- Classes rítmicas conjecturadas:
  - línguas acentuais: Holandês, Inglês, Polonês, Português Europeu,...
  - línguas silábicas: Catalão, Espanhol, Francês, Italiano,Português Brasileiro, ...
  - línguas moraicas: Japonês, ...

Como extrair padrões rítmicos de textos escritos?

- Uma tentativa foi marcar os elementos pertinentes do ritmo indicando para cada sílaba:
  - se ela carrega ou não o acento principal da palavra
  - se ela é ou não começo de palavra

A Codificação

- Os textos foram codificados sob o alfabeto A={0,1,2,3,4} de acordo com a localização da sílaba tônica de cada palavra como descrito abaixo
  - 0 - Sílaba não tônica
  - 1 - Sílaba tônica
  - 2 - Sílaba não tônica no início de palavra prosódica
  - 3 - Sílaba tônica no início de palavra prosódica
  - 4 – Início de sentença

A Codificação

Palavra prosódica é uma palavra lexical(com seu acento principal) e todas as palavras funcionais que a precedem

Exemplo:

| O Menino | já | comeu | a bala |

4 | 2 0 1 0 | 3 | 2 1 | 2 1 0 |

Essa codificação é feita automaticamente através do software escrito em Perl que pode ser baixado no site http://www.ime.usp.br/~tycho/prosody/vlmc/tools/silaba.pl

VLMC

Uma VLMC é uma cadeia de Markov de ordem finita cujas probabilidades de transição tem a seguinte propriedade

$$\mathbb{P}(X_0 = x_0 \mid X_{-K}^{-1} = x_{-K}^{-1}) = \mathbb{P}(X_0 = x_0 \mid X_{-K}^{-1} = x_{-\ell(x_{-K}^{-1})}^{-1})$$

Onde $\ell : \mathcal{A}^K \to \{1, \ldots, K\}$ **é uma função do passado que indica o número de passos que devemos olhar para trás para escolher o próximo estado da cadeia**

Uma VLMC é convenientemente representada por uma árvore probabilística

Misturas de VLMC

- **Uma mistura de VLMCs é o conjunto de modelos VLMC, isto é, um conjunto de árvores e suas respectivas Probabilidades de Transição, com uma distribuição de probabilidades nesse conjunto.**
- **Uma mistura de modelos pode ser interpretada como se em cada tempo t, procurássemos o modelo que melhor se ajusta ao texto até o presente momento, e gerássemos o próximo símbolo de acordo com o modelo escolhido.**

A Floresta pode esconder árvores significativas?

Como ponderar as árvores?

O algoritmo para ponderar as árvores é o seguinte:

1. **Primeiro, inicializamos os pesos da mistura com as probabilidades *a priori* das árvores ($\omega^0(\tau)$).**
2. **Então, atualizamos os pesos para cada símbolo do texto de entrada.**

$$\omega^{n+1}(\tau) = \omega^n(\tau) P_\tau^n(x_n | c_\tau(x_0, \ldots, x_{n-1}))$$

Onde $P_\tau^n(x_n | c_\tau(x_0, \ldots, x_{n-1}))$ **é o estimador de máxima Verossimilhança com a amostra até o n-ésimo símbolo**

O Peso *a priori*

- A Verossimilhança aumenta quando o tamanho da árvore aumenta, portanto, as árvores maiores recebem pesos menores.

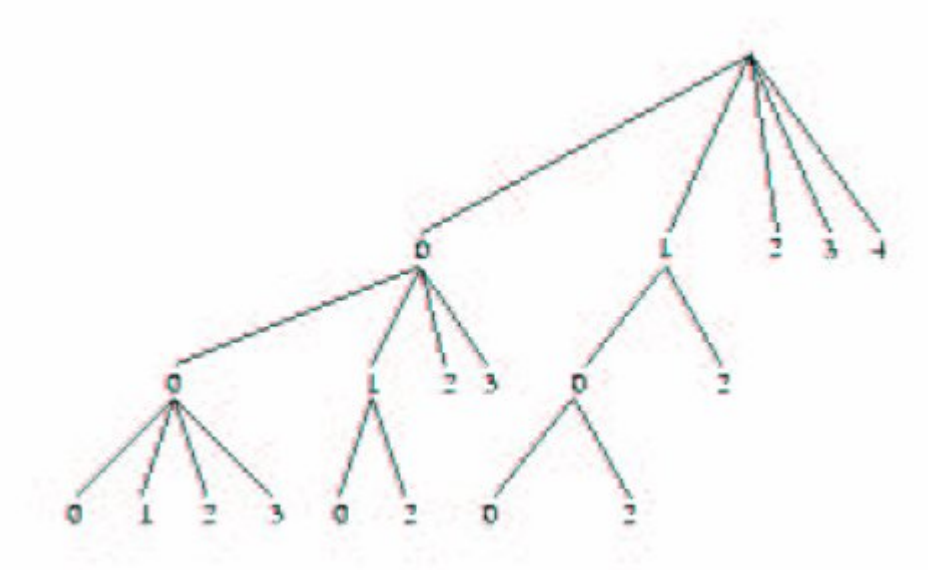
- O peso a priori também precisa considerar o tamanho da amostra. Quanto maior a amostra, menor o peso.

Ex: w(T) = (ct)-n , onde ‘n’ is o número de nós Terminais da árvore ‘T’ e ‘t’ é o amanho do texto de entrada.

Como implementar a proposta?

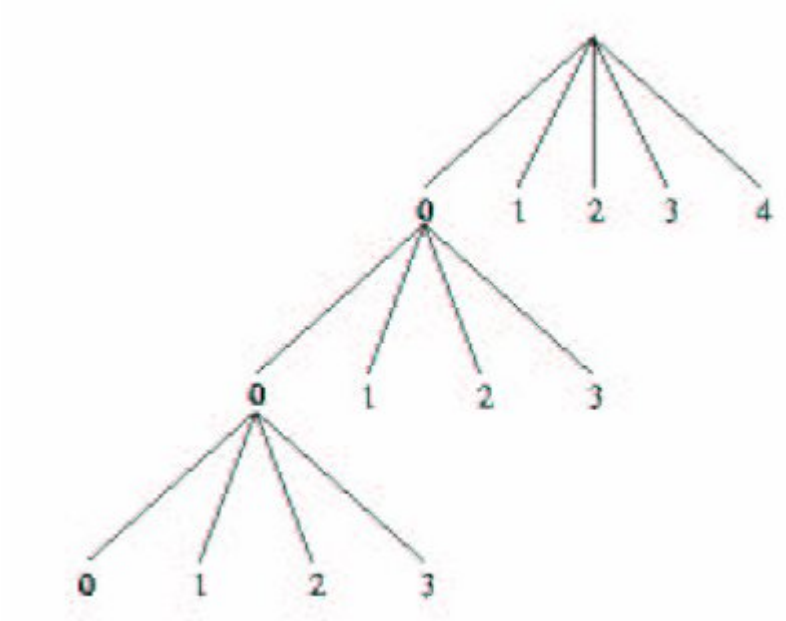
- Gerar todas as árvores.
- Resolver o problema do aumento da precisão.
  - Guardar apenas os dígitos mais significativos
- Calcular os pesos e as probabilidades de transição deforma eficiente.
  - Uso de uma árvore molde
- Aplicar restrições
  - Expressões Regulares

Resultado para PB



Árvore com probabilidade 0.8, para um tamanho de amostra de 619282 símbolos. O peso a priori foi (500t)−N.

Resultado para PE



Árvore com probabilidade 0.99 para um tamanho de amostra de 148887 símbolos. O peso a priori foi (500t)−N.

“Uma árvore resume a floresta.”