MRP y HMM

Andrea Fernández, Liliana Millán

27/05/2015

Aplicación 1: Modelo de reconocimiento de vocales

Problema

Supongamos que somos alienígenas de Las Pléyades y que no tenemos ni idea de cómo se 'lee' un lenguaje de la tierra, no sabemos de los idiomas pero como somos seres superiores sabemos de Hidden Markov Models!

- Objetivo:

Queremos establecer ciertas propiedades de este lenguaje que no conocemos, veremos que al identificar estas propiedades, de manera *natural* identificaremos las vocales de las consonantes.

Especificación del modelo

- Utilizamos HMM con el algoritmo Baum-Welch para estimar los parámetros:

1. las probabilidades inciales de los estados
2. las probabilidades de transición entre estados
3. las probabilidades de cada símbolo de pertenecer a uno de los estados

- Únicamente con la evidencia que tienen los datos (nuestras observaciones)

Baum-Welch

- Este algoritmo es una variante del EM visto en clase. Iniciamos con un modelo sin 'conocimiento'

π = probabilidades de inciar en cada estado

A= matriz de transición de estados

B= matriz de emisiones

$\lambda = (A, B, \pi)$

- En cada iteración los valores de π, A y B se van actualizando hasta convergencia
- El algoritmo ocupa el forward procedure —probabilidad de ver esta secuencia parcial y terminar en el estado i en el tiempo t— y el backward procedure —probabilidad de terminar en la esta secuencia parcial dado que empezamos en el estado i en el tiempo t—

Datos

- Tomamos el corpus de noticias de un periódico español
- 309,918 noticias

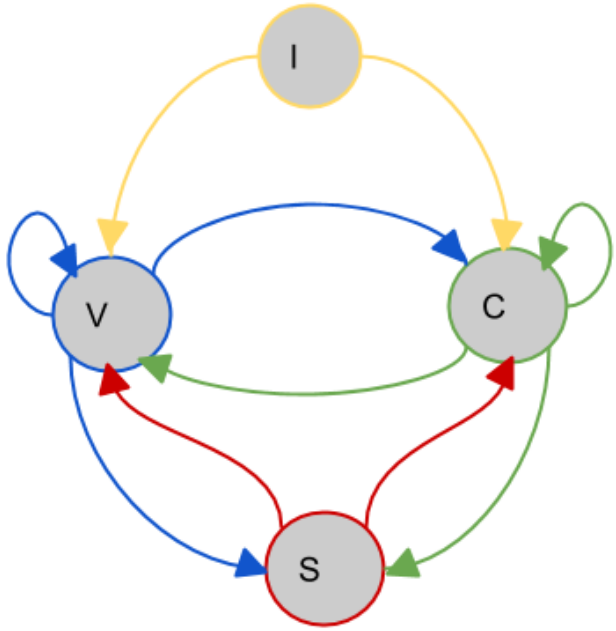
Limpieza de datos

- Eliminación de signos de puntuación
- Eliminación de dígitos
- Eliminación de tabuladores
- Todas las letras a minúsculas
- Cada palabra es separada en sus letras respetando los espacios

Suposiciones iniciales del modelo

- Nuestra base será suponer que existen 2 estados: **Consonante** y **Vocal**
- No conocemos con qué probabilidad de inicio estamos en Constante o en Vocal
- No conocemos las probabilidades de transición entre estados
- No conocemos las probabilidades de que cada símbolo del lenguaje pertenezca a uno de los estados

Modelo



I: Inicio
V: Vocal
C: Consonante
S: Espacio

Paquetes utlizadas

- Paquete HMM de R
- Algoritmo de Baum-Welch para estimación de parámetros de una HMM

Resultados

Inicial sin conocimiento:

V	C
0.5337	0.4662

Inicial después de Baum-Welch

V	C
0.5337	0.4662

Resultados

Transiciones sin conocimiento

	V	C
V	0.3099	0.6900
C	0.5200	0.4799

Transiciones después de Baum-Welch

	V	C
V	0.3045	0.6954
C	0.993	0.006

Resultados

```
$hmm$emissionProbs
      symbols
states           a            b            c           d         e            f           g
     v 0.118369315 1.018630e-65 3.344456e-76 0.008194266 0.19487864 1.743170e-66 0.008411338
     c 0.005519715 1.172973e-02 7.037839e-02 0.140688529 0.04784931 1.172973e-02 0.011348951
      symbols
states           h            i j k           l           m           n           o           p
     v 4.732532e-83 5.702808e-02 0 0 4.622025e-80 1.983201e-92 1.489029e-19 1.466436e-01 8.415060e-72
     c 1.172973e-02 7.955236e-62 0 0 1.290270e-01 3.518919e-02 1.524865e-01 7.102166e-28 4.691892e-02
      symbols
states           q           r           s           t           u           v w x y           z
     v 2.196755e-72 0.008237816 0.01800018 9.776189e-02 6.517495e-02 1.212351e-93 0 0 0 6.455586e-96
     c 3.518919e-02 0.117166365 0.11484041 7.666003e-07 1.653220e-87 2.345946e-02 0 0 0 1.172973e-02
      symbols
states           á é í ó           ú           ñ
     v 0.26915304 0 0 0 0  8.146868e-03 2.268708e-92
     c 0.01128861 0 0 0 0 3.850728e-194 1.172973e-02
```