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 periódicos españoles
- 309,918 noticias

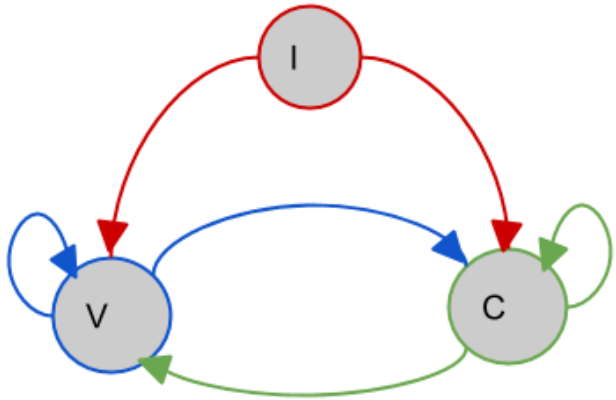
Limpieza de datos

- 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

Paquetes utlizadas

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

Resultados Español

Inicial sin conocimiento:

V	C
0.5337	0.4662

Inicial después de Baum-Welch

V	C
0.5337	0.4662

Resultados Español

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 Español

```
$hmm$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
```

Resultado Griego

Corpus:

Resultados Griego

Inicial sin conocimiento:

V	C
0.53765	0.4234

Inicial después de Baum-Welch

V	C
0.6117	0.3882

Resultados Griego

Transiciones sin conocimiento

	V	C
V	0.3558	0.6441
C	0.5161	0.4838

Transiciones después de Baum-Welch

	V	C
V	0.0093	0.9906
C	0.6178	0.3821

Resultados Griego

Vocales en griego

Griego	Vocal
A,α	a
E,ϵ	e
I,ι	i
O,o	o
Y,υ	u

Resultados Griego

```
$hmm$emissionProbs
      symbols
states           έ           ν           α           ς           χ           ρ
     v 5.895505e-02 4.252000e-81 2.181337e-01 0.004291154 0.07996329 2.843513e-16 1.162688e-25
     c 5.184611e-27 8.876421e-02 3.446021e-12 0.045388583 0.24941473 1.849254e-02 5.917614e-02
      symbols
states           ο           π           ό           τ           η           κ           β
     v 0.1766922643 2.387271e-41 7.074606e-02 1.352442e-29 9.432809e-02 2.552561e-11 1.393980e-125
     c 0.0001084652 5.177912e-02 5.927745e-37 1.405433e-01 1.107997e-26 3.698509e-02  7.397017e-03
      symbols
states           ά           λ           ε           γ           ω           ή           σ
     v 2.358202e-02 4.927002e-208 1.355966e-01 8.493358e-104 4.126854e-02 2.947753e-02 0.02748251
     c 2.758986e-49  1.109553e-02 2.827769e-31  1.849254e-02 2.026343e-41 3.752589e-44 0.03453814
      symbols
states           ώ           ύ           φ           υ           ξ           θ           ζ
     v 0.011227611 5.895505e-03 0.001745016 2.111235e-05 1.079921e-107 3.686890e-20 1.564313e-75
     c 0.004051954 3.528964e-41 0.017397818 5.546438e-02  3.698509e-03 1.479403e-02 7.397017e-03
      symbols
states           ι           ψ           ί           δ           μ
     v 2.233928e-07 5.533383e-208 0.006105212 0.014488497 1.270457e-96
     c 5.917600e-02  3.698509e-03 0.029456511 0.009403274 3.328658e-02
```

Aplicación 2: Modelos jerárquicos y postestratificación

Problema

¿Cómo realizar inferencia sobre la población objetivo con datos de encuesta recabados con un diseño no probabilístico basado en cuotas y sin marco muestral?

Objetivo

- Generar estimaciones precisas y confiables
- Controlar por sesgo de selección

Un poco de teoría de encuestas

Tipos de errores de encuestas

- Error de cobertura
- Error de muestreo
- Errores por no respuesta
- Errores de medición

Tipos de muestreo

- Probabilístico
- No probabilístico

No probabilístico por cuotas

Problema principal: sesgo de selección.

$\Rightarrow$ Para hacer inferencias acerca de la población a partir de una muestra de este tipo es necesario suponer que las personas que fueron seleccionadas son similares a las que no lo fueron...

Soluciones posibles:

- Sample matching
- Máximo entropía
- MRP

Especificación del modelo: pre MR

Se denotará al estimador por este método como $\tilde{\theta}$ y se obtiene con el siguiente proceso:

1. Identificación de una o más variables que pueden ser responsables del sesgo de selección. SPG, la cuadrícula completa de clasificación se trata como una única variable categórica G.

Limitación: Con los datos de INEGI a nivel manzana solo podemos especificar 5 modelos

- Edad x colonia
- Condición de ocupación x colonia
- Escolaridad x colonia
- Género x edad x colonia
- Condición de ocupación x género x colonia

¿Cómo se ven los datos?

Datos del censo

- Cuadricula: colonia x genero x edad
- Dimensiones: 63 x 2 x 4 = 504

idcolonia	genero	edad	value
46241	mujer	e12a17	38
46284	mujer	e12a17	0
46385	mujer	e12a17	171
46388	mujer	e12a17	124
46408	mujer	e12a17	74
46409	mujer	e12a17	108

¿Cómo se ven los datos?

Datos de encuestas

- Datos individuales con las respuestas a los cuestionarios.
- Las variables elegidas para G se recodifican *igualito* al censo.
- La variable de interés en 0 y 1.
- Morelos 2013: 5862
- Morelos 2014: 10365

idcolonia	genero	edad	victimizacion
46548	mujer	e18a29	0
46548	mujer	e12a17	1
46548	mujer	e30a49	1
46284	hombre	e30a49	0

Especificación del modelo: el MR

2. Se define un nuevo estimador
   $\gamma \equiv E(Y|D = d, G = g), d = 1, ..., J, g = 1, ..., G.$
3. Se utiliza un modelo de regresión multinivel apropiadamente especificado para estimar γ.

Especificación del modelo: el P

4. El paso de postestratificación utiliza el modelo generado en el paso 3. Se computa el estimador MRP para cada elemento θ_d de θ como la suma ponderada del subconjunto apropiado de $\hat{\gamma}$.

$$\tilde{\theta_d} = \sum_{g=1}^{G} \hat{\gamma_{d,g}} w_{g|d}$$

donde $w_{g|d} = \frac{N_{g,d}}{N_d}$. El numerador es el número de miembros de la población objetivo que pertenecen simultáneamente a la categoría g y d. El denominador es el número de miembros en la población objetivo que pertenecen a la categoría d.

Resultados

Muestras no comparables, ¡ahora lo son!

	e12a17	e18a29	e30a49	e50omas
Hombres.14	17.8684843	21.9930360	23.452651	20.998060
Hombres.13	18.4638051	21.4389763	27.861405	22.353171
Diferencia	-0.5953207	0.5540598	-4.408754	-1.355112
Mujeres.14	14.9715021	18.4999490	19.796484	17.682535
Mujeres.13	17.1969776	19.9538614	26.118268	20.920638
Diferencia	-2.2254755	-1.4539125	-6.321784	-3.238104

MRP

Resuelve: small area estimation y/o selection bias

Ventajas del método

- El uso de la regresión multinivel incrementa la precisión del estimador.
- Si G se define adecuadamente, la postestratificación ayuda a decrecer el error por sesgo de selección.
- MRP es un estimador relativamente preciso para θ.

Desventajas del método

- Se necesitan datos poblacionales para toda la clasificación DxG lo cuál limita la definición de G.
- Para obtener buenos estimadores de γ, el modelo de regresión multinivel debe ser especificado con mucho cuidado. Sin embargo, esta limitación aplica para cualquier modelo.