Tarea 1 - Aprendizaje automatizado

Estimación de máxima verosimilitud

La estimación de parámetros sólo se aplica a valores numéricos, así que se transformarán los nombres a números. Para transformarlos, se aplicará la siguiente función hash:
Sea f la función que a cada letra del alfabeto le asigna un número del 0 al 25 (no se cuenta la ñ): $f(a)=0, f(b)=1, \ldots, f(y)=24, f(z)=25$. El código hash (ch) de una cadena será el siguiente:
$ch = (\Sigma f(\text{nombre}[i]) * 26^i) \bmod 1000$
Nótese que el código hash siempre será un número menor a 10,000.
Denis: 5071
Guadalupe: 4854
Alex: 7238
Cris: 2220
Juan: 9017
Rene: 9213

Las tablas, cambiando el nombre por un número, serán las siguientes:

Género masculino		
Nombre	Estatura	Peso
5071	1.72	75.3
4854	1.82	87.6
7238	1.80	86.1
7238	1.70	77.1
2220	1.73	78.2
9017	1.80	74.8
9017	1.80	74.3

Género femenino		
Nombre	Estatura	Peso
5071	1.50	50.5
7238	1.52	45.3
2220	1.62	61.2
9213	1.67	68.0
4854	1.65	58.9
4854	1.75	68.0

En las siguientes tablas se muestran los parámetros obtenidos:

EMV (masculino)			
	μ	σ^2	σ
nombre	6379.28	5246802.77	2290.59
estatura	1.7671	0.0020204	0.044948
peso	78.2	15.7657	3.9706

EMV (femenino)			
~~nombre~~	μ	σ^2	σ
nombre	5575	4758389.33	2181.37
estatura	1.618	0.007447	0.08629
peso	58.65	71.0091	8.4266

Una vez que se encontraron los valores, se calcula $P(x_1, x_2, \ldots, x_d | y=c)\ P(y=c) = P(x_1|y)\,P(x_2|y) \ldots P(x_d|y)\,P(y)$, donde y puede tomar los valores de M o F.

En este caso:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

porque se asume distribución normal.

Se usará el logaritmo natural de $P(x_i|y)$ porque se está operando con números muy pequeños.

Los resultados se muestran en las siguientes tablas.

	Género masculino				
	ln(P(nombre | y=M))	ln(P(estatura | y=M))	ln(P(peso | y=M))	ln(P(y=M))	suma
x_1	-9.4207	0.3057	-7.823	-0.619	-17.557
x_2	-8.8772	2.110	-2.4006	-0.619	-9.7859
x_3	-8.8186	1.9154	-2.318	-0.619	-9.840
x_4	-8.7257	-2.187	-3.764	-0.619	-15.297
x_5	-10.304	-1.210	-4.430	-0.619	-16.563

Género femenino					
	$\ln(P(\text{nombre} \mid y=F))$	$\ln(P(\text{estatura} \mid y=F))$	$\ln(P(\text{peso} \mid y=F))$	$\ln(P(y=F))$	suma
x_1	-9.9973	1.272	-3.334	-0.773	-12.831
x_2	-8.6612	0.367	-6.260	-0.773	-15.333
x_3	-8.6333	-0.693	-5.966	-0.773	-16.066
x_4	-8.897	-3.808	-7.939	-0.773	-21.418
x_5	-9.789	1.462	-3.957	-0.773	-13.057

Cada ejemplar se clasifica de acuerdo a la suma mayor (dado M o dado F).

x_1: F, x_2: M, x_3: M, x_4: M, x_5: F

Estimación de máximo a posteriori

Primero se van a tomar los nombres como categorías y se ordenan alfabéticamente.

Alex: 1, Cris: 2, Denis: 3, Guadalupe: 4, Juan: 5, Rene: 6.

Se obtienen los vectores c de la distribución categórica.

$c_M = [2, 1, 1, 1, 2, 0]$

$c_F = [1, 1, 1, 2, 0, 1]$

donde $c[i] = m$ significa que hay m elementos en la categoría i.

Luego, dadas las α's ($\alpha_k = 2$ para toda k), se calculan los vectores q para cada género, que son las q_k de máximo a posteriori.

$q_M = [0.1578, 0.1052, 0.1052, 0.1578, 0.0526]$

$q_F = [0.1052, 0.1052, 0.1052, 0.1578, 0.0526, 0.1052]$

Ahora se calcula μ para estatura y peso

Género masculino

$\mu_{estatura} = 1.767$

$\mu_{peso} = 78.995$

Género femenino

$\mu_{estatura} = 1.616$

$\mu_{peso} = 60.073$

Cuando se tienen esos datos, se puede calcular la máxima verosimilitud. Una vez más, se trabajará con el logaritmo de cada uno y los resultados se muestran en las siguientes tablas.

Género masculino					
	ln(P(nombre|y=M))	ln(P(estatura|y=M))	ln(P(peso|y=M))	ln(P(y=M))	suma
x_1	-2.945	0.296	-8.511	-0.619	-11.779
x_2	-2.251	2.116	-2.329	-0.619	-3.084
x_3	-2.251	1.916	-2.297	-0.619	-3.252
x_4	-1.846	-2.233	-3.441	-0.619	-8.141
x_5	-2.251	-1.233	-4.864	-0.619	-8.969

Género femenino					
	ln(P(nombre|y=F))	ln(P(estatura|y=F))	ln(P(peso|y=F))	ln(P(y=F))	suma
x_1	-2.251	1.257	-3.221	-0.773	-4.988
x_2	-1.846	0.320	-5.846	-0.773	-8.145
x_3	-2.251	-0.754	-5.573	-0.773	-9.352
x_4	-2.251	-3.918	-7.426	-0.773	-14.369
x_5	-2.251	1.456	-3.744	-0.773	-5.313

Con esa información se tiene la siguiente clasificación

x_1: F

x_2: M

x_3: M

x_4: M

x_5: F