Proyecto

Héctor González Alvarado
Abraham Quiles Sanchez

Noviembre 2019

1. Introducción.

En el siguiente se presentan tres conjuntos de datos, el primero tiene la actividad enzimática en la sangre, el segundo de la acidez en un grupo de sustancias y el tercero la velocidad de desplazamiento de las ciertas galaxias observadas.

El objetivo del proyecto es desarrollar un modelo estadístico de una muestra con varios grupos mezclados, el cual, nos ayudará a estimar la cantidad de grupos y, para cada grupo, los parámetros de su distribución.

Para esto, tendremos los pesos de la distribución que serán $\vec{w} = (w_1, w_2, ..., w_k)$, con k la cantidad de grupos y $\sum_{i=1}^{k} w_i = 1$. Suponiendo que la población $j \in \{1, .., k\}$ tiene una distribución $N(x|\mu_j, \sigma_j^2)$. Luego,

$$p(x|\vec{w}, \vec{\mu}, \vec{\sigma}) = \sum_{j=1}^{k} w_j N(x|\mu_j, \sigma_j^2).$$

Y suponiendo una muestra $X_1, ..., X_n$, $Z_1, ..., Z_n$, donde las Xs registran la observación de la población y las Zs el grupo al que pertenece.

Usando el algoritmo EM tenemos las siguientes expresiones:

$$w_j^{t+1} = \mathbb{E}(n_j)/n.$$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^{n} x_i \mathbb{P}(Z_i = j)}{\mathbb{E}(n_j)}.$$

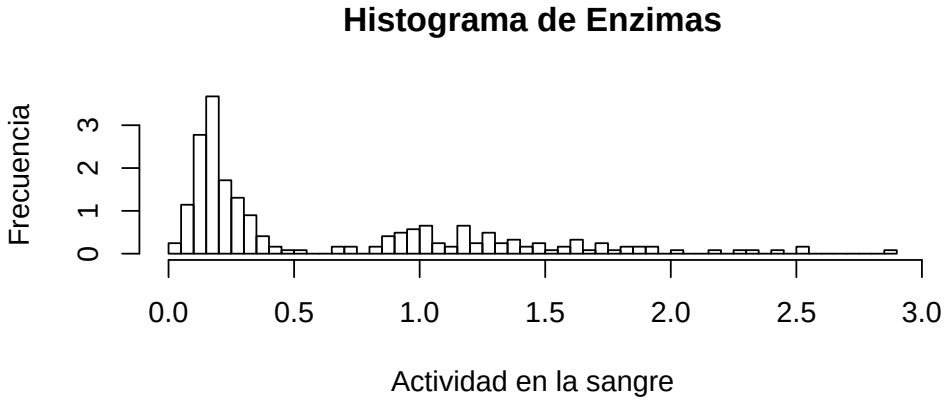
$$(\sigma_j^2)^{t+1} = \frac{\sum_{i=1}^{n} (x_i - \mu_j)^2 \mathbb{P}(Z_i = j)}{\mathbb{E}(n_j)}.$$

Donde $n_j = \sum_{i=1}^{n} \chi_{(z_i=j)}$.

Presentamos los valores AIC, AIC_c como los siguientes cálculos, $AIC = 2R - log(\mathbf{L}(\hat{\mu}, \hat{\sigma^2}, \hat{\mathbf{w}}|\mathbf{x}))$ y $AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$, con $\hat{\mu}, \hat{\sigma^2}, \hat{w}$, los estimadores máximos verosímiles obtenidos por el EM y k el modelo(que dependerá del numero de grupos en este caso). Y el vector de diferencias se obtiene restando cada AIC por el mínimo, de estos AIC.

2. Datos de enzimas.

Primero tenemos el análisis general de la base: El mínimo es 0.0210, la mediana es 0.2640, el promedio 0.6223 y el valor máximo es 2.8800. A continuación un histograma de los datos



Hicimos la estimación de los pesos, varianzas y medias, desde 1 hasta 6 grupos,es decir, $k \in \{1, ..., 6\}$. Resumimos estos datos en una matriz, donde el renglón k representa la cantidad de grupos, y la columna el parámetro correspondiente a ese modelo. Primero la matriz de pesos es

k	w_1	w_2	w_3	w_4	w_5	w_6
1	1	0	0	0	0	0
2	0.5920	0.4079	0	0	0	0
3	0.6087	0.1672	0.2240	0	0	0
4	0.6126	0.0593	0.2525	0.07542	0	0
5	0.6124	0.05944	0.2332	0.0560	0.0387	0
6	0.6119	0.0602	0.2249	0.04958	0.0297	0.0234

Ahora la de medias

k	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
1	0.6222	0	0	0	0	0
2	0.1876	1.2530	0	0	0	0
3	0.1908	1.0660	1.4633	0	0	0
4	0.1913	0.9597	1.1972	1.9320	0	0
5	0.1912	0.9595	1.1960	1.5984	2.0482	0
6	0.1912	0.9594	1.2094	1.4655	1.6104	2.3316

Por ultimo, la de varianzas

k	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_5^2	σ_6^2
1	0.38515	0	0	0	0	0
2	0.00582	0.2636	0	0	0	0
3	0.00635	0.0351	0.3155	0	0	0
4	0.00645	0.0033	0.1166	0.2400	0	0
5	0.00645	0.0033	0.1111	0.3095	0.2101	0
6	0.00643	0.0034	0.1146	0.2685	0.2740	0.1125

Con lo obtenido tendremos un vector con los AIC y AIC_c que para cada k, este es

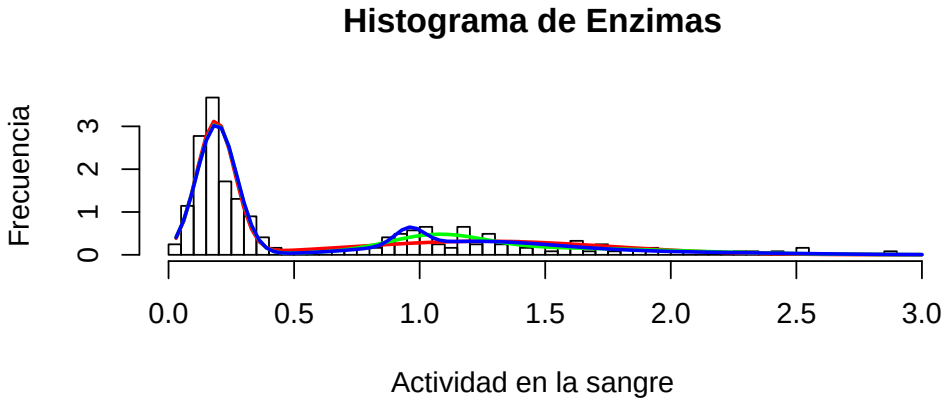
k	1	2	3	4	5	6
AIC	463.5212	113.2800	101.6536	98.5528	100.5481	102.2199
AIC_c	463.5377	113.3296	101.753	98.719	100.799	102.572

El mínimo de los AIC, es $AIC_{min} = 98.5528$.

Por lo tanto, el vector de diferencias es y pesos de los datos es

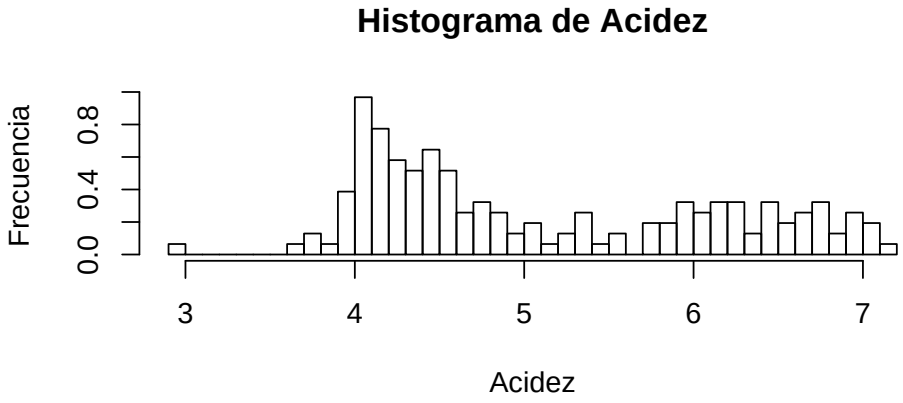
k	1	2	3	4	5	6
Δ	364.9684	14.7272	3.1007	0	1.995281	3.6671
W_d	$3.2\text{x}10^{-80}$	0.00036	0.1218	0.5742	0.2117	0.0917

Tomamos las tres inferencias con pesos de datos más grande, y anexamos al histograma la gráfica de éstos.



3. Datos de acidez.

El análisis general de la base: El mínimo es 2.929, la mediana es 4.727, el promedio 5.105 y el valor máximo es 7.105. A continuación un histograma de los datos



La matriz de pesos es

k	w_1	w_2	w_3	w_4	w_5	w_6
1	1	0	0	0	0	0
2	0.4791	0.5208	0	0	0	0
3	0.0798	0.4277	0.4923	0	0	0
4	0.0806	0.3611	0.2594	0.2987	0	0
5	0.0702	0.3115	0.1251	0.1743	0.3187	0
6	0.0723	0.3760	0.1905	0.0200	0.1722	0.1686

La de medias

k	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
1	5.1050	0	0	0	0	0
2	4.2505	5.8912	0	0	0	0
3	4.0718	4.3110	5.9625	0	0	0
4	4.0714	4.3063	5.0088	6.4332	0	0
5	4.0696	4.2405	4.6824	5.0111	6.3955	0
6	4.0701	4.2669	4.9632	3.8923	6.0705	6.736

La de varianzas

k	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_5^2	σ_6^2
1	1.0384	0	0	0	0	0
2	0.2603	0.8474	0	0	0	0
3	0.0480	0.2930	0.8118	0	0	0
4	0.0483	0.2729	0.7798	0.3989	0	0
5	0.0454	0.2408	0.3499	0.8501	0.4192	0
6	0.0459	0.2611	0.4510	0.7996	0.2211	0.21837

Con lo obtenido tendremos un vector con los AIC y AIC_c que para cada k, este es

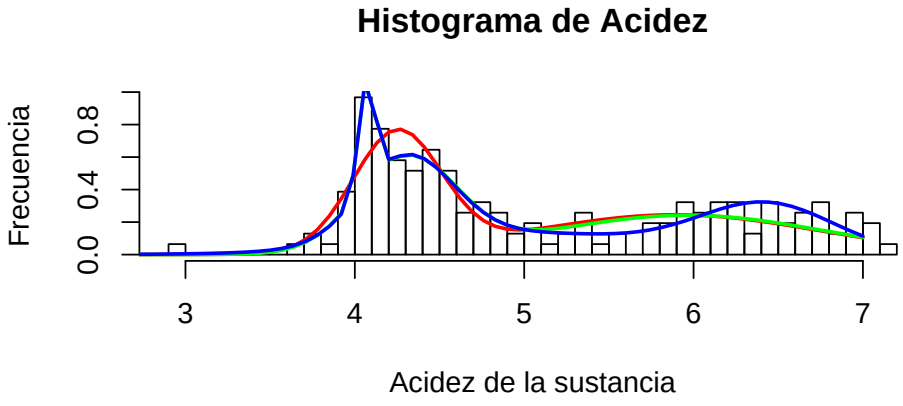
k	1	2	3	4	5	6
AIC	453.5707	378.4690	374.7178	360.3664	361.7671	357.5202
AIC_c	453.5969	378.5480	374.8768	360.6331	362.1697	358.0878

El mínimo de los AIC, es $AIC_{min} = 357.5202$.

Por lo tanto, el vector de diferencias es y pesos de los datos es

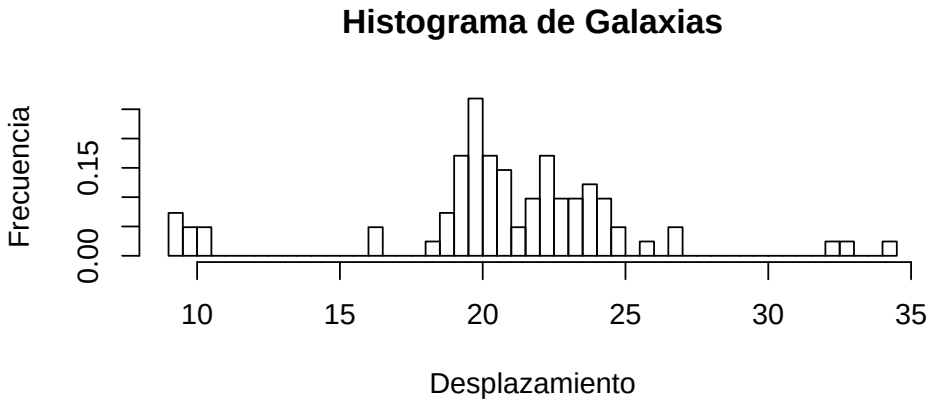
k	1	2	3	4	5	6
Δ	96.0505	20.9488	17.1976	2.8462	4.2468	0
W_d	$1.02\text{x}10^{-21}$	2.0710^{-05}	1.3510^{-04}	0.177	0.0879	0.7348

Mostramos el histograma para 2(rojo),3(verde),4(azul) grupos.



4. Datos de galaxias.

El análisis general de la base: El mínimo es 9.172, la mediana es 20.834, el promedio 20.831 y el valor máximo es 34.279. El histograma de los datos es



La matriz de pesos es

k	w_1	w_2	w_3	w_4	w_5	w_6
1	1	0	0	0	0	0
2	0.0851	0.9142	0	0	0	0
3	0.0853	0.8780	0.0365	0	0	0
4	0.0853	0.2085	0.6694	0.0365	0	0
5	0.0853	0.3387	0.2630	0.2763	0.0364	0
6	0.0853	0.3441	0.2638	0.2068	0.0635	0.0363

La de medias

k	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
1	20.8314	0	0	0	0	0
2	9.7093	21.8671	0	0	0	0
3	9.7101	21.4038	33.0443	0	0	0
4	9.7101	19.7477	21.9199	33.0445	0	0
5	9.7100	19.8168	21.8548	22.9259	33.0477	0
6	9.7100	19.8283	21.8886	22.6189	23.9978	33.0480

La de varianzas

k	σ_1^2	σ_2^2	σ_3^2	σ_4^2	σ_5^2	σ_6^2
1	4.5401	0	0	0	0	0
2	0.4221	3.1503	0	0	0	0
3	0.4225	2.2038	0.9217	0	0	0
4	0.4225	0.4362	2.2783	0.9217	0	0
5	0.4224	0.6221	2.9774	1.0028	0.9217	0
6	0.4224	0.6293	2.9904	0.7912	0.5140	0.9217

Los AIC y AIC_c que para cada k, son

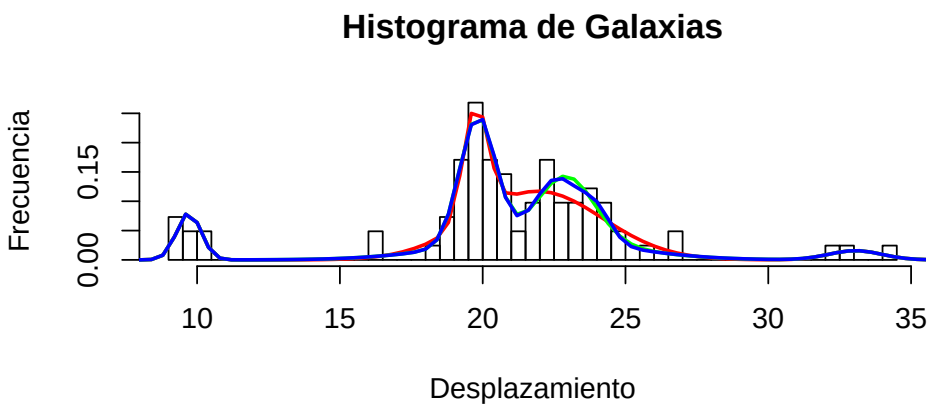
k	1	2	3	4	5	6
AIC	482.8330	444.3863	412.9640	403.4206	402.2613	403.8990
AIC_c	482.8830	444.5382	413.2717	403.9401	403.0508	405.0190

El mínimo de los AIC, es $AIC_{min} = 402.2612$.

Por lo tanto, el vector de diferencias es y pesos de los datos es

k	1	2	3	4	5	6
Δ	80.5717	42.1250	10.7026	1.159343	0	1.6376
W_d	$1.59\text{x}10^{-18}$	$3.55\text{x}10^{-10}$	0.0023	0.279	0.498	0.219

Mostramos el histograma para 4(rojo),5(verde),6(azul) grupos.



5. Conclusiones.

De lo observado en el los histogramas en general, tenemos que las aproximaciones hechas por el algoritmo EM para este tipo de datos son bastante buenas. Se puede notar en el código, que al cambiar el máximo de iteraciones, las estimaciones convergen con no tantas iteraciones.

Del AIC, podemos notar que en el caso de las enzimas tenemos aproximadamente 2 ó 3 grupos, la acidez de 3 a 4 grupos, y en el caso de las galaxias entre 5 ó 6 grupos. En el caso de la galaxia, notamos que hay datos alejados de un grupo mayor, con varianzas pequeñas. En ese caso, hay que modificar las medias y las varianzas iniciales, para que estén mas dispersos y atrapen por medio del EM los valores por estimar.