BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

SISTEMA PARA LA OBTENCIÓN DE UNA MEZCLA DE NORMALES MEDIANTE EL ALGORITMO MAXIMIZACIÓN DE LA ESPERANZA (EM)

TESIS PARA OBTENER EL GRADO DE:

LICENCIADO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA: OMAR HERNÁNDEZ MALAGÓN

ASESOR: DR. GERARDO MARTÍNEZ GUZMÁN

CO-ASESOR: DR. JOSÉ ALEJANDRO RANGEL HUERTA

MARZO 2020

Dedicatoria

Dedicado a todos mis conocidos que me ayudaron, para seguir adelante.

Agradecimientos

Agradezco a mi mamá que me dio de su tiempo para continuar mis estudios cuando lo necesitaba, a mi hermano que me motivo a continuar mis estudios, a mi demás familia por darme la motivación a concluir este trabajo final de mi carrera y a todos mis conocidos que me brindaron de su apoyo para perseverar.

Agradezco a mi asesor de tesis Dr. Gerardo Martínez Guzmán, por el tiempo brindado para enseñarme nuevos temas, despejar mis dudas que se fueron presentando y sobre todo por sus consejos. También agradezco a mi Co-Asesor de tesis el Dr. José Alejandro Rangel Huerta por el apoyo otorgado para la realización de este trabajo.

ÍNDICE

SISTEMA PARA LA OBTENCIÓN DE UNA MEZCLA DE NORMALES MEDIANTE EL ALGORITMO MAXIMIZACIÓN DE LA ESPERANZA (EM)

RESUMEN

El cáncer de seno ocurre como resultado del crecimiento anormal de células en el tejido del seno, comúnmente conocido como un tumor, los tumores pueden ser benignos (no cancerosos), o malignos (cancerosos). Las pruebas como la resonancia magnética, la mamografía, la ecografía y la biopsia se utilizan comúnmente para diagnosticar el cáncer de mama.

La mayoría de los bultos en los senos son benignos, los tumores no cancerosos en los senos son crecimientos anormales, pero no se propagan fuera de ellos. Estos tumores no representan un peligro para la vida, aunque algunos tipos de bultos benignos pueden aumentar el riesgo en una mujer de padecer cáncer de seno.

La mayoría de los cánceres de seno comienza en los conductos que llevan la leche hacia el pezón, aunque algunos cánceres se originan en las glándulas que producen leche. También hay otros tipos de cáncer de seno que son menos comunes como el tumor filodes y el angiosarcoma.

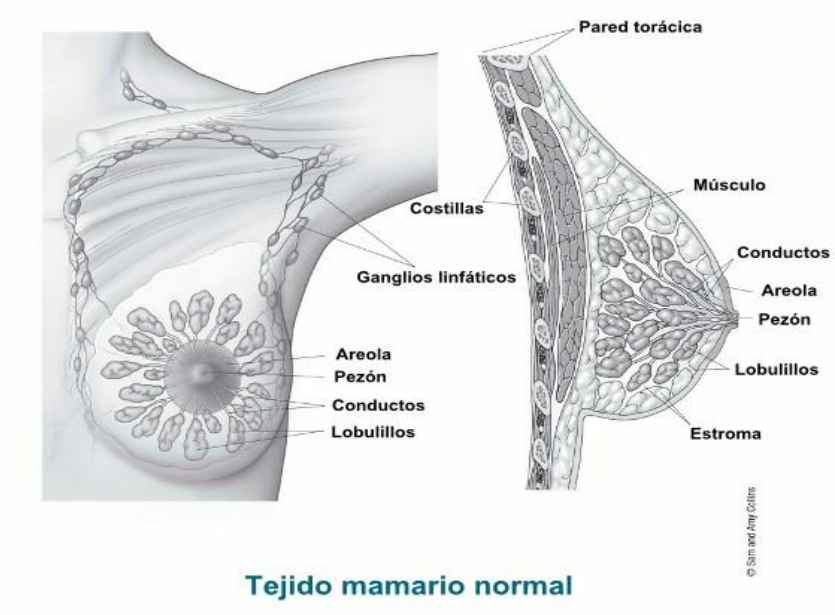

Imagen 1 Tejido mamario normal.

El cáncer de seno se puede propagar cuando las células cancerosas alcanzan la sangre o el sistema linfático y llegan a otras partes del cuerpo. El líquido transparente dentro de los vasos linfáticos, llamado linfa, contiene productos derivados de los tejidos y materia de desecho, así como células del sistema inmunitario. En el caso de cáncer de seno, las células cancerosas pueden ingresar en los vasos linfáticos y comenzar a crecer en los ganglios linfáticos.

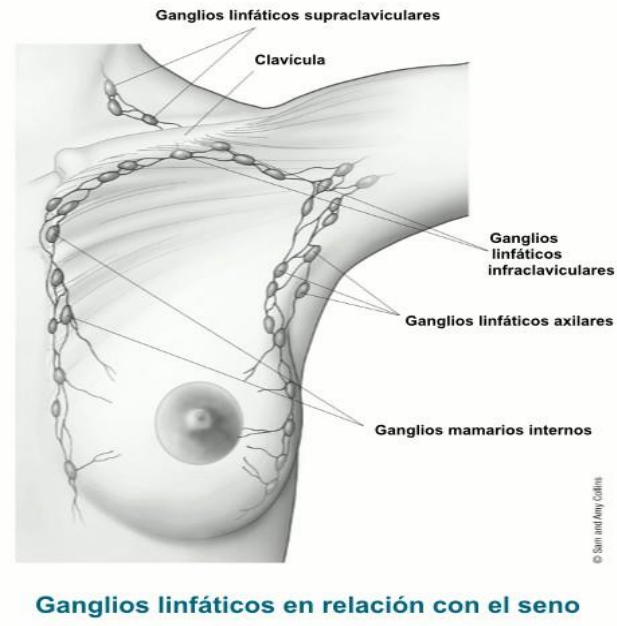

Imagen 2 Ganglios linfáticos en relación con el seno.

Si las células cancerosas se han propagado a sus ganglios linfáticos, hay una mayor probabilidad de que las células se hayan desplazado por el sistema linfático y se hayan propagado (metástasis) a otras partes de su cuerpo. Sin embargo, no todas las mujeres con células cancerosas en sus ganglios linfáticos presentan metástasis, y es posible que algunas mujeres sin células cancerosas en sus ganglios linfáticos desarrollen metástasis más adelante.

En México el cáncer de mama ocupa el primer lugar en incidencia de las neoplasias malignas en las mujeres, donde el grupo de edad más afectado se encuentra entre los 40 y los 59 años. De acuerdo con cifras del INEGI en el año 2016 en México se observaron 16 defunciones por cada 100 000 mujeres de 20 años y más.

En este trabajo se aplica el algoritmo EM (Maximización de la Esperanza) a una de las variables (radius_mean) creadas por el Dr. William H. Wolberg, médico del Hospital de la Universidad de Wisconsin en Madison, Wisconsin, EE. UU. El Dr. Wolberg creo un conjunto de datos, usando muestras de fluidos, tomadas de pacientes con masas mamarias sólidas y un programa informático gráfico fácil de usar llamado Xcyt, que es capaz de realizar el análisis de características citológicas basadas en un escáner digital. El análisis de estas variables desde el punto de vista frecuencial tiene un comportamiento de tipo mezcla de normales, por lo que llamo nuestro interés en trabajar con una de estas variables (*radius_mean*) para encontrar la forma de la mezcla de normales mediante un algoritmo de aprendizaje no supervisado llamado maximización de la esperanza.

El análisis tiene como objetivo observar qué características son más útiles para predecir el cáncer maligno o benigno y ver las tendencias generales que pueden ayudarnos en la selección del modelo. El objetivo es clasificar si el cáncer de seno es benigno o maligno. Para lograr esto, se ha usado métodos de clasificación de aprendizaje automático para ajustar una función que puede predecir la clase discreta de entrada nueva.

INTRODUCCIÓN

El algoritmo de maximización de la esperanza o algoritmo EM se usa en estadística para calcular estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables. El algoritmo EM computa la esperanza de la verosimilitud mediante la inclusión de variables latentes como si fueran observables, y computa estimadores de máxima verosimilitud de los parámetros. El algoritmo EM fue expuesto por Arthur Dempster, Nan Laird y Donald Rubin de la Royal Statistical Society en una publicación de 1977. Los autores señalan que el método ya había sido "propuesto muchas veces en situaciones especiales" por otros autores, pero la publicación de 1977 generaliza el método.

El algoritmo EM se utiliza frecuentemente para algoritmos de agrupamiento en aprendizaje automático y visión artificial y Mixturas de Gaussianas, utilizadas en procesos de clasificación o reconocimiento. De esta forma, por su capacidad para manejar información faltante y observar variables ocultas, se está convirtiendo en una herramienta importante en muchos procesos de aprendizaje automático.

Las distribuciones de mixturas finitas se han empleado para la modelización de datos heterogéneos ya que, con frecuencia, no es suficiente explicar la distribución de unos datos mediante una única distribución estadística, es necesaria la utilización de una combinación de distribuciones. Estas combinaciones comúnmente son descritas mediante los modelos de mixturas, los cuales se definen por los parámetros de cada componente y las proporciones en las que cada una de ellas contribuye a la distribución general. Este concepto conduce al agrupamiento del conjunto de observaciones en grupos con algunas características comunes. Las

distribuciones mixtas pueden ser estimadas mediante muchas técnicas, tales como métodos gráficos, el método de los momentos, de máxima verosimilitud, aproximaciones bayesianas y análisis de componentes principales por mencionar algunas.

El algoritmo EM (Maximización de la Esperanza) es una herramienta habitual iterativa para la estimación de máxima verosimilitud de las distribuciones mixtas. La idea es introducir una variable indicadora multinomial que identifica la pertenencia a un grupo de cada observación del conjunto de datos. Esto representa una aproximación conveniente para la obtención de los parámetros de las mixturas cuando no existe una solución analítica.

CAPITULO 1. ANALISIS DE LOS DATOS

Tabla 1.1 *Datos de los estudios de cáncer de seno.*

	ID	M = maligno, B = benigno	radio		ID	M = maligno, B = benigno	radio
1	862722	B	6.981	47	872608	B	9.904
2	921362	B	7.691	48	907367	B	10.030
3	921092	B	7.729	49	897880	B	10.050
4	92751	B	7.760	50	874158	B	10.080
5	85713702	B	8.196	51	924964	B	10.160
6	871001502	B	8.219	52	858970	B	10.170
7	91805	B	8.571	53	8812844	B	10.180
8	894047	B	8.597	54	8811779	B	10.200
9	858981	B	8.598	55	894604	B	10.250
10	858477	B	8.618	56	898677	B	10.260
11	872113	B	8.671	57	90317302	B	10.260
12	864496	B	8.726	58	922840	B	10.260
13	903483	B	8.734	59	924934	B	10.290
14	9010333	B	8.878	60	922577	B	10.320
15	859711	B	8.888	61	88147101	B	10.440
16	864726	B	8.950	62	884437	B	10.480
17	89346	B	9.000	63	907409	B	10.480
18	859471	B	9.029	64	862989	B	10.490
19	894329	B	9.042	65	892657	B	10.490
20	859196	B	9.173	66	864292	B	10.510
21	915186	B	9.268	67	892399	B	10.510
22	917092	B	9.295	68	901315	B	10.570
23	924342	B	9.333	69	909777	B	10.570
24	905539	B	9.397	70	8910251	B	10.600
25	905978	B	9.405	71	88466802	B	10.650
26	925236	B	9.423	72	871642	B	10.660
27	901034301	B	9.436	73	8910720	B	10.710
28	859464	B	9.465	74	869254	B	10.750
29	8510824	B	9.504	75	87127	B	10.800
30	882488	B	9.567	76	90745	B	10.800
31	898143	B	9.606	77	923465	B	10.820
32	9113778	B	9.667	78	923748	B	10.860
33	9113514	B	9.668	79	911391	B	10.880
34	915276	B	9.676	80	871149	B	10.900
35	923169	B	9.683	81	891936	B	10.910
36	875099	B	9.720	82	904971	B	10.940
37	8710441	B	9.731	83	919537	B	10.960
38	868999	B	9.738	84	909411	B	10.970
39	8910996	B	9.742	85	901836	B	11.040
40	907145	B	9.742	86	905520	B	11.040
41	9112712	B	9.755	87	89827	B	11.060
42	864033	B	9.777	88	904302	B	11.060
43	862261	B	9.787	89	91544002	B	11.060
44	917897	B	9.847	90	871641	B	11.080
45	862980	B	9.876	91	90769601	B	11.130
46	879804	B	9.876	92	923780	B	11.130

	ID	M = maligno, B = benigno	radio		ID	M = maligno, B = benigno	radio
93	901011	B	11.140	141	868223	B	11.710
94	87106	B	11.150	142	874373	B	11.710
95	906290	B	11.160	143	89864002	B	11.710
96	925311	B	11.200	144	8912055	B	11.740
97	88203002	B	11.220	145	91550	B	11.740
98	897132	B	11.220	146	8711561	B	11.750
99	897137	B	11.250	147	91858	B	11.750
100	8913049	B	11.260	148	857343	B	11.760
101	91789	B	11.260	149	904357	B	11.800
102	901549	B	11.270	150	874662	B	11.810
103	903011	B	11.270	151	8810528	B	11.840
104	868871	B	11.280	152	891703	B	11.850
105	8910748	B	11.290	153	9111596	B	11.870
106	883852	B	11.300	154	8811523	B	11.890
107	859465	B	11.310	155	8911164	B	11.890
108	88199202	B	11.320	156	905686	B	11.890
109	8911230	B	11.330	157	869476	B	11.900
110	864018	B	11.340	158	864685	B	11.930
111	916221	B	11.340	159	904647	B	11.930
112	905502	B	11.360	160	857374	B	11.940
113	89143601	B	11.370	161	8912909	B	11.940
114	865137	B	11.410	162	899147	B	11.950
115	873843	B	11.410	163	9110720	B	11.990
116	868682	B	11.430	164	8612080	B	12.000
117	869218	B	11.430	165	9111843	B	12.000
118	861103	B	11.450	166	895299	B	12.030
119	89296	B	11.460	167	9113816	B	12.040
120	898690	B	11.470	168	857155	B	12.050
121	911685	B	11.490	169	90250	B	12.050
122	88518501	B	11.500	170	871122	B	12.060
123	925291	B	11.510	171	898678	B	12.060
124	85759902	B	11.520	172	918465	B	12.070
125	884689	B	11.520	173	903554	B	12.100
126	893988	B	11.540	174	912193	B	12.160
127	921385	B	11.540	175	86211	B	12.180
128	906539	B	11.570	176	862965	B	12.180
129	862485	B	11.600	177	894090	B	12.180
130	893061	B	11.600	178	866714	B	12.190
131	911320501	B	11.600	179	89511501	B	12.200
132	906616	B	11.610	180	9011495	B	12.210
133	911366	B	11.620	181	902975	B	12.210
134	9112366	B	11.630	182	91544001	B	12.220
135	863031	B	11.640	183	877501	B	12.230
136	899187	B	11.660	184	8711003	B	12.250
137	91903901	B	11.670	185	91376701	B	12.250
138	913512	B	11.680	186	905501	B	12.270
139	919812	B	11.690	187	9113846	B	12.270
140	893783	B	11.700	188	874839	B	12.300

	ID	M = maligno, B = benigno	radio		ID	M = maligno, B = benigno	radio
189	897374	B	12.300	237	8913	B	12.890
190	8610175	B	12.310	238	869224	B	12.900
191	87139402	B	12.320	239	875878	B	12.910
192	8712064	B	12.340	240	9047	B	12.940
193	904969	B	12.340	241	891670	B	12.950
194	91813702	B	12.340	242	904689	B	12.960
195	861597	B	12.360	243	896864	B	12.980
196	863270	B	12.360	244	897604	B	12.990
197	90251	B	12.390	245	871001501	B	13.000
198	907915	B	12.400	246	9112594	B	13.000
199	883539	B	12.420	247	873357	B	13.010
200	894335	B	12.430	248	854941	B	13.030
201	913063	B	12.450	249	857810	B	13.050
202	892604	B	12.460	250	893548	B	13.050
203	914101	B	12.460	251	9010259	B	13.050
204	87930	B	12.470	252	8510653	B	13.080
205	914580	B	12.470	253	8810158	B	13.110
206	894089	B	12.490	254	9113455	B	13.140
207	901034302	B	12.540	255	8711002	B	13.150
208	91505	B	12.540	256	914102	B	13.160
209	9010258	B	12.560	257	911320502	B	13.170
210	8912521	B	12.580	258	884448	B	13.200
211	88147202	B	12.620	259	89344	B	13.200
212	911202	B	12.620	260	9112367	B	13.210
213	86408	B	12.630	261	922296	B	13.210
214	914366	B	12.650	262	9113239	B	13.240
215	89511502	B	12.670	263	861853	B	13.270
216	906024	B	12.700	264	8813129	B	13.270
217	891716	B	12.720	265	902727	B	13.280
218	90769602	B	12.720	266	901041	B	13.300
219	917080	B	12.750	267	8611161	B	13.340
220	89382602	B	12.760	268	865468	B	13.370
221	9010598	B	12.760	269	9112085	B	13.380
222	875093	B	12.770	270	9010877	B	13.400
223	924084	B	12.770	271	862009	B	13.450
224	859487	B	12.780	272	9013579	B	13.460
225	90401602	B	12.800	273	91813701	B	13.460
226	873586	B	12.810	274	912519	B	13.470
227	911408	B	12.830	275	857156	B	13.490
228	905190	B	12.850	276	893526	B	13.500
229	8610908	B	12.860	277	90401601	B	13.510
230	894855	B	12.860	278	8610629	B	13.530
231	8910506	B	12.870	279	8510426	B	13.540
232	908916	B	12.870	280	8812818	B	13.560
233	917062	B	12.880	281	8910499	B	13.590
234	924632	B	12.880	282	8911800	B	13.590
235	884626	B	12.890	283	922576	B	13.620
236	8912284	B	12.890	284	857373	B	13.640

	ID	M = maligno, B = benigno	radio		ID	M = maligno, B = benigno	radio
285	88350402	B	13.640	333	921644	B	14.740
286	8812816	B	13.650	334	89869	B	14.760
287	9013594	B	13.660	335	9110944	B	14.800
288	906878	B	13.660	336	915664	B	14.810
289	91903902	B	13.680	337	908469	B	14.860
290	9013005	B	13.690	338	914333	B	14.870
291	912558	B	13.700	339	911384	B	14.920
292	917896	B	13.710	340	86973701	B	14.950
293	869931	B	13.740	341	8915	B	14.960
294	88411702	B	13.750	342	8712291	B	14.970
295	891923	B	13.770	343	8712853	B	14.970
296	90944601	B	13.780	344	905557	B	14.990
297	86561	B	13.850	345	88147102	B	15.000
298	8911834	B	13.850	346	914862	B	15.040
299	909231	B	13.850	347	866458	B	15.100
300	901028	B	13.870	348	9012568	B	15.190
301	922297	B	13.870	349	8810436	B	15.270
302	902976	B	13.880	350	867387	B	15.710
303	911673	B	13.900	351	912600	B	15.730
304	91227	B	13.900	352	905189	B	16.140
305	918192	B	13.940	353	901303	B	16.170
306	909410	B	14.020	354	915452	B	16.300
307	88249602	B	14.030	355	9010872	B	16.500
308	909220	B	14.040	356	8711216	B	16.840
309	925292	B	14.050	357	91376702	B	17.850
310	903811	B	14.060	358	855563	M	10.950
311	89524	B	14.110	359	9013838	M	11.080
312	911654	B	14.200	360	84348301	M	11.420
313	883270	B	14.220	361	892189	M	11.760
314	86409	B	14.260	362	869691	M	11.800
315	892214	B	14.260	363	853612	M	11.840
316	8910721	B	14.290	364	875263	M	12.340
317	88143502	B	14.340	365	843786	M	12.450
318	9113156	B	14.400	366	84501001	M	12.460
319	89143602	B	14.410	367	85922302	M	12.680
320	89813	B	14.420	368	868202	M	12.770
321	86973702	B	14.440	369	881861	M	12.830
322	921386	B	14.470	370	844981	M	13.000
323	865432	B	14.500	371	863030	M	13.110
324	911150	B	14.530	372	85638502	M	13.170
325	911201	B	14.530	373	85715	M	13.170
326	915940	B	14.580	374	856106	M	13.280
327	925277	B	14.590	375	915691	M	13.400
328	89382601	B	14.610	376	87163	M	13.430
329	861648	B	14.620	377	855167	M	13.440
330	861598	B	14.640	378	855138	M	13.480
331	913102	B	14.640	379	862717	M	13.610
332	906564	B	14.690	380	866083	M	13.610

	ID	M = maligno, B = benigno	radio		ID	M = maligno, B = benigno	radio
381	84458202	M	13.710	429	873701	M	15.700
382	84667401	M	13.730	430	8812877	M	15.750
383	875938	M	13.770	431	899667	M	15.750
384	859983	M	13.800	432	84610002	M	15.780
385	87880	M	13.810	433	864877	M	15.780
386	91504	M	13.820	434	846381	M	15.850
387	8810987	M	13.860	435	845636	M	16.020
388	886452	M	13.960	436	896839	M	16.030
389	908489	M	13.980	437	8610404	M	16.070
390	8810955	M	14.190	438	869104	M	16.110
391	874858	M	14.220	439	84862001	M	16.130
392	854268	M	14.250	440	854039	M	16.130
393	858986	M	14.250	441	86730502	M	16.160
394	91979701	M	14.270	442	8912280	M	16.240
395	862548	M	14.420	443	911916	M	16.250
396	877500	M	14.450	444	895633	M	16.260
397	86135501	M	14.480	445	8953902	M	16.270
398	84799002	M	14.540	446	9012315	M	16.350
399	852763	M	14.580	447	87281702	M	16.460
400	90291	M	14.600	448	926954	M	16.600
401	848406	M	14.680	449	852552	M	16.650
402	857793	M	14.710	450	913535	M	16.690
403	859283	M	14.780	451	854253	M	16.740
404	87556202	M	14.860	452	8712729	M	16.780
405	864729	M	14.870	453	879830	M	17.010
406	907914	M	14.900	454	85382601	M	17.020
407	868826	M	14.950	455	881972	M	17.050
408	855133	M	14.990	456	89742801	M	17.060
409	91594602	M	15.050	457	911296201	M	17.080
410	862028	M	15.060	458	852631	M	17.140
411	9010018	M	15.080	459	889719	M	17.190
412	857438	M	15.100	460	859717	M	17.200
413	879523	M	15.120	461	909445	M	17.270
414	905680	M	15.130	462	888570	M	17.290
415	925622	M	15.220	463	8860702	M	17.300
416	873885	M	15.280	464	888264	M	17.350
417	852973	M	15.300	465	881094802	M	17.420
418	886776	M	15.320	466	88330202	M	17.460
419	8511133	M	15.340	467	8712766	M	17.470
420	861799	M	15.370	468	877989	M	17.540
421	8670	M	15.460	469	853201	M	17.570
422	87164	M	15.460	470	9113538	M	17.600
423	915460	M	15.460	471	8711202	M	17.680
424	903507	M	15.490	472	9110732	M	17.750
425	90602302	M	15.500	473	90439701	M	17.910
426	88725602	M	15.530	474	8911163	M	17.930
427	889403	M	15.610	475	865128	M	17.950
428	887181	M	15.660	476	842302	M	17.990

	ID	M = maligno, B = benigno	radio		ID	M = maligno, B = benigno	radio
477	90524101	M	17.990	525	89263202	M	20.090
478	914062	M	18.010	526	926682	M	20.130
479	9110127	M	18.030	527	894618	M	20.160
480	8610637	M	18.050	528	8610862	M	20.180
481	877159	M	18.080	529	908194	M	20.180
482	857392	M	18.220	530	9011494	M	20.200
483	894326	M	18.220	531	86208	M	20.260
484	844359	M	18.250	532	84358402	M	20.290
485	874217	M	18.310	533	887549	M	20.310
486	916799	M	18.310	534	895100	M	20.340
487	867739	M	18.450	535	901088	M	20.440
488	8612399	M	18.460	536	91930402	M	20.470
489	914769	M	18.490	537	883263	M	20.480
490	852781	M	18.610	538	88206102	M	20.510
491	853401	M	18.630	539	919555	M	20.550
492	857010	M	18.650	540	842517	M	20.570
493	86517	M	18.660	541	881046502	M	20.580
494	897630	M	18.770	542	91485	M	20.590
495	8911670	M	18.810	543	927241	M	20.600
496	908445	M	18.820	544	901288	M	20.640
497	859575	M	18.940	545	88995002	M	20.730
498	866203	M	19.000	546	926125	M	20.920
499	86135502	M	19.020	547	884948	M	20.940
500	855625	M	19.070	548	873593	M	21.090
501	8611792	M	19.100	549	911157302	M	21.100
502	8912049	M	19.160	550	851509	M	21.160
503	846226	M	19.170	551	9012795	M	21.370
504	877486	M	19.180	552	926424	M	21.560
505	8711803	M	19.190	553	903516	M	21.610
506	857637	M	19.210	554	9011971	M	21.710
507	854002	M	19.270	555	8910988	M	21.750
508	884180	M	19.400	556	9012000	M	22.010
509	89122	M	19.400	557	86355	M	22.270
510	913505	M	19.440	558	915143	M	23.090
511	886226	M	19.450	559	88299702	M	23.210
512	88119002	M	19.530	560	8712289	M	23.270
513	892438	M	19.530	561	878796	M	23.290
514	88649001	M	19.550	562	89812	M	23.510
515	90312	M	19.550	563	865423	M	24.250
516	871201	M	19.590	564	91762702	M	24.630
517	9111805	M	19.590	565	8611555	M	25.220
518	898431	M	19.680	566	899987	M	25.730
519	84300903	M	19.690	567	873592	M	27.220
520	885429	M	19.730	568	911296202	M	27.420
521	866674	M	19.790	569	8810703	M	28.110
522	8811842	M	19.800				
523	849014	M	19.810				
524	916838	M	19.890				

Los datos de la variable en estudio (*radius_mean*) se tiene para una muestra de 569 mujeres, donde se describe el ID de la persona, si el tumor es benigno o maligno y el valor de la variable.

Los datos anteriores se clasifican en una tabla de frecuencias considerando las siguientes clases:

A=(6,7], B=(7,8], C=(8,9], D=(9,10], E=(10,11], F=(11,12], G=(12,13], H=(13,14], I=(14,15], J=(15,16], K=(16,17], L=(17,18], M=(18,19], N=(19,20], O=(20,21], P=(21,22], Q=(22,23], R=(23,24], S=(24,25], T=(25,26], U=(26,27], V=(27,28], W=(28,29)

Construyendo un histograma de frecuencias relativas, se observa un comportamiento que puede aproximarse mediante una mezcla de distribuciones normales como se muestra en la gráfica. Para encontrar esta aproximación usaremos el algoritmo de EM (Maximización de la Esperanza). Para el conjunto de datos (*radius_mean*), este análisis tiene como objetivo observar qué características son más útiles para predecir el cáncer maligno o benigno y ver tendencias generales que pueden ayudarnos en la selección de modelos y la selección de parámetros.

Tabla 1.2 *Frecuencias de la variable*

A=(6,7]	1	0
B=(7,8]	3	0
C=(8,9]	13	0
D=(9,10]	30	0
E=(10,11]	37	1
F=(11,12]	81	5
G=(12,13]	81	7
H=(13,14]	59	19
I=(14,15]	40	19
J=(14,15]	6	26
K=(16,17]	5	18
L=(17,18]	1	25
M=(18,19]	0	21
N=19,20]	0	26

O=(20,21]	0	23
P=(21,22]	0	8
Q=(22,23]	0	2
R=(23,24]	0	5
S=24,25]	0	2
T=(25,26]	0	2
U=(26,27]	0	0
V=(27,28]	0	2
W=(28,29]	0	1

El histograma de frecuencias relativas nos indica que una aproximación por medio de una mezcla de dos normales es posible.

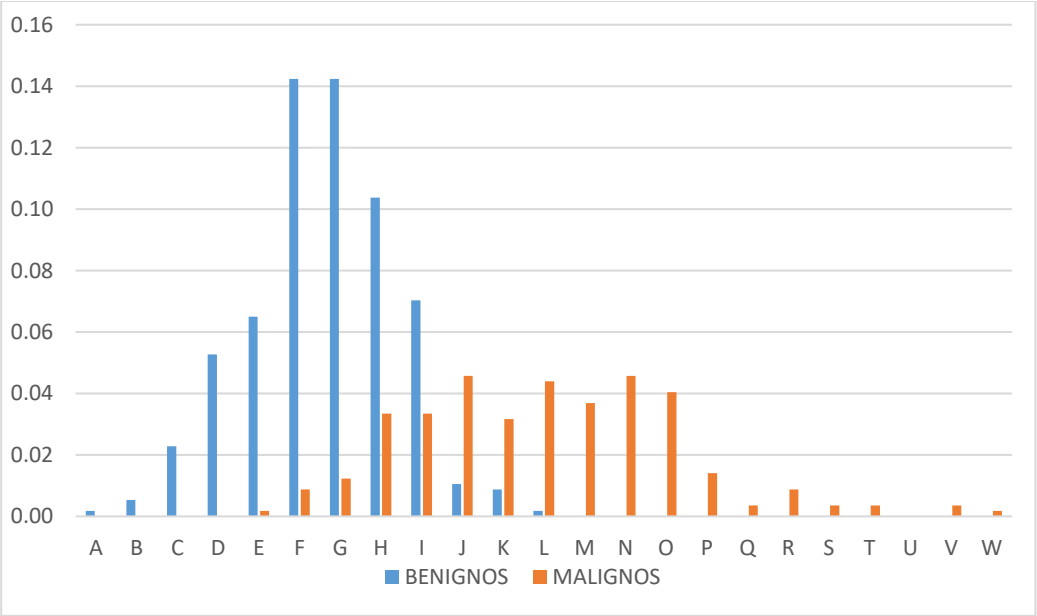

Fig. 1.1 Representación de los datos de la tabla 1.2

CAPITULO 2. DENSIDAD NORMAL

En este trabajo se consideran mixturas de funciones de densidad normal por lo que damos la siguiente definición.

Definición. Se dice que una variable aleatoria Y sigue una distribución normal o Gaussiana si su función de densidad puede escribirse como

$$\varphi(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}},$$

donde la variable y los parámetros satisfacen las condiciones:

$$-\infty < y < \infty, -\infty < \mu < \infty \ \ y \ \ \sigma > 0.$$

2.1 Simulación de la densidad Normal

Para simular la densidad normal estándar $N(0,1)$. Consideremos dos variables aleatorias independientes $u_1 \ y \ u_2$ tales que $u_1 \sim U(0,1)$ y $u_2 \sim U(0,1)$, entonces la función de densidad conjunta de $u_1 \ \ y \ \ u_2$ es

$$f(u_1,u_2) = \begin{cases} 1 & 0 \le u_1 \le 1, \ \ 0 \le u_2 \le 1 \\ 0 & en\ cualquier\ otro\ caso \end{cases}$$

Ahora, definimos

$$X = h_1(u_1,u_2) = \sqrt{2\log(1/u_1)}\ \cos(2\pi\, u_2)$$

$$Y = h_2(u_1,u_2) = \sqrt{2\log(1/u_1)}\ \operatorname{sen}(2\pi\, u_2)$$

Entonces es posible demostrar que $X \sim N(0,1)$ y $Y \sim N(0,1)$, además X e Y son independientes. También sabemos que si $X \sim N(0,1)$, entonces la

variable aleatoria definida como $Z = \sigma X + \mu$ se distribuye como una normal $N(\mu, \sigma^2)$. Utilizando esta estrategia, la distribución normal estándar y cualquier otra normal se puede simular fácilmente.

Aplicando lo anterior se ha desarrollado un programa en java que permite crear muestras de tamaño n con media μ y desviación estándar σ de una distribución normal.

```
public double[] genMuestraNormal(int n, double desvEstandar,double media){

    if(n<1 || desvEstandar<=0 || media<=0)
        return null;

    double muestra[]=new double[n];
    for(int i=0; i<n;i++)
        muestra[i]=( Math.sqrt(2*Math.log10(1/Math.random()) )
                     *Math.cos(2*Math.PI*Math.random()) )*desvEstandar
                     + media;
    return muestra;
}
```

Fig. 2.1.1 El código en java para generar una muestra de tamaño n con media μ y desviación estándar σ.

Entonces para una muestra de tamaño $n = 10$ y parámetros $\mu = 2$ y $\sigma = 1$ se puede visualizar en la Fig. (2.1.2), ahora si aumentamos $n = 100$ se puede ver que la forma de la normal empieza a aparecer, y si tomamos $n = 100{,}000$, se puede ver claramente que el comportamiento de los datos es el de una normal, Fig. (2.1.2c).

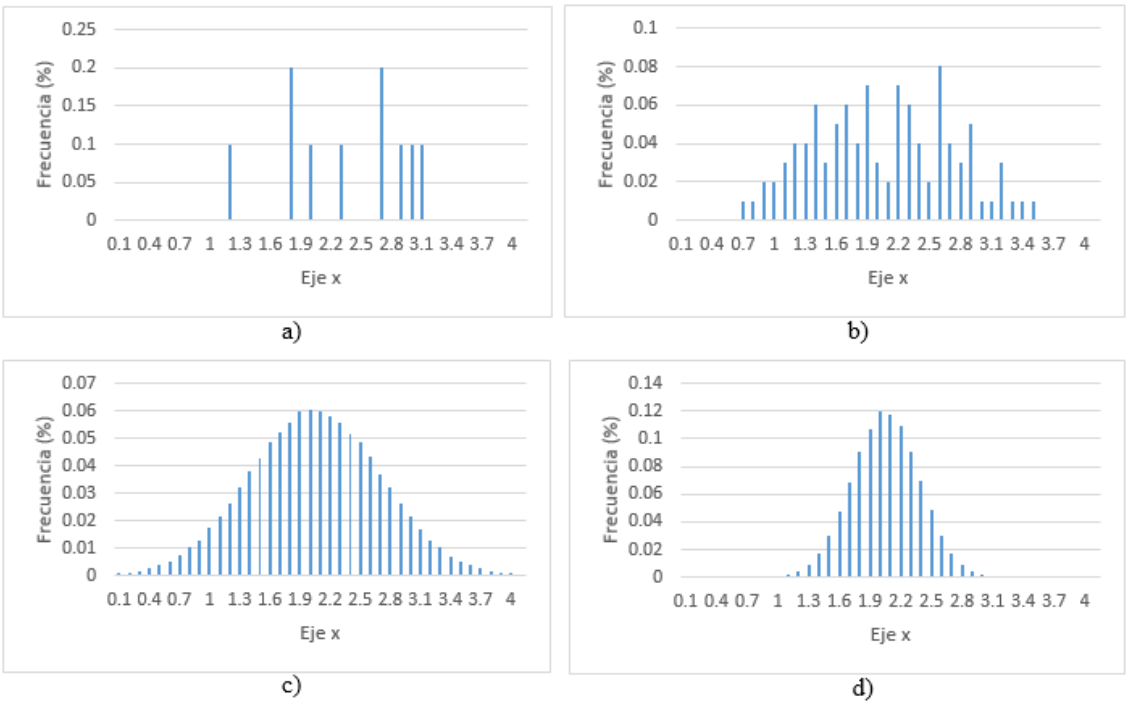

Fig. 2.1.2 Muestras de una normal con $\mu = 2, \sigma = 1$, (a) n=10, (b) n=100, (c) n=100,000. (d) Muestras de una normal con $\mu = 2, \sigma = 0.5$ y n=100,000.

2.2 Mixturas Gaussianas de dos componentes

La mixtura de dos componentes Gaussianas las podemos escribir como

$$\varphi(y) = \pi_1 \varphi(y_1|\mu_1, {\sigma_1}^2) + \pi_2 \varphi(y_2|\mu_2, {\sigma_2}^2)$$

Donde el vector de parámetros es

$$\psi = (\pi_1, \pi_2, \mu_1, \mu_2, {\sigma_1}^2, {\sigma_2}^2)$$

Se asume que las medias de las componentes se encuentran en orden ascendente,

$$\mu_1 < \mu_2$$

Si Y *es una* variable aleatoria que sigue una distribución mixta de dos componentes Gaussianas, entonces la media μ y varianza σ^2 de la mixtura son:

$$\mu = \pi_1\mu_1 + \pi_2\mu_2$$

$$\sigma^2 = \pi_1({\sigma_1}^2 + {\mu_1}^2) + \pi_2({\sigma_2}^2 + {\mu_2}^2) - \mu^2$$

CAPITULO 3. DESARROLLO

3.1 Modelo de Mixturas con g componentes

En este capítulo se dan las definiciones básicas relacionadas con los modelos de mixturas finitas y la estimación de máxima verosimilitud. Posteriormente, se presenta el algoritmo EM como un método para la obtención de soluciones para las ecuaciones de verosimilitud en el caso de que no exista una forma analítica para ellas. Se expone el problema de valores iniciales, un aspecto donde el algoritmo se muestra especialmente sensible.

Las distribuciones mixtas son utilizadas para la modelización de datos heterogéneos en multitud de situaciones experimentales, en donde aquellos pueden interpretarse como procedentes de dos o más subpoblaciones. La obtención de estas componentes conduce a la estimación de los parámetros de la mixtura. Este problema de estimación se remonta a Pearson (1894), quien trabajó con una mixtura de dos componentes con varianzas iguales.

3.2 Algoritmo EM

En el desarrollo del algoritmo EM, proporcionamos una formulación paramétrica para la representación del modelo. En lo sucesivo, mediante $Y = (Y_1, Y_2, \ldots, Y_n)$, se denotará a una muestra aleatoria de tamaño *n*, donde Y_i es un vector aleatorio *q*-dimensional con función de densidad de probabilidad $f(y_i)$ en R^q . Así, $y = (y_1, y_2, \ldots, y_n)$ representa a una muestra observada o realización de Y , donde y_i constituye un valor observado del vector aleatorio Y_i.

Definición. Si la función de densidad de una variable aleatoria Y_i es de la forma

$$f(y_i|\boldsymbol{\psi}) = \sum_{k=1}^{g} \pi_k f_k(y_i|\boldsymbol{\theta_k}) \qquad y_i \in R^q$$

se dice que posee una distribución de mixtura finita con *g* componentes, con un vector de parámetros

$$\psi = (\pi_1, \ldots, \pi_g, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_g}).$$

Aquí, $f_k(y_i|\boldsymbol{\theta_k}),\ k = 1,2, \ldots, g$, denotan las densidades de las componentes de la mixtura con parámetros, $\boldsymbol{\theta_k}$ y $\pi_1, \ldots, \pi_g$. A los parámetros $\pi_1, \ldots, \pi_g$, se les llamamos proporciones o pesos de la mixtura. Se asume también que las funciones $f_k(y_i|\boldsymbol{\theta_k})$ pueden pertenecer a diferentes familias paramétricas,

Para que la mixtura sea una función de densidad los pesos deben cumplir las condiciones

$$0 \leq \pi_k \leq 1 \quad k = 1, \ldots, g \quad y \quad \sum_{k=1}^{g} \pi_k = 1$$

Note que en la condición anterior uno de los pesos resulta redundante (uno de ellos se expresa en términos de los demás).

Para mixturas cuyas componentes pertenecen a otras familias de densidades, puede consultarse Simar (1976) y Moharir (1992) para mixturas de distribuciones Poisson; Falls (1970), para mixturas Weibull; o Blischke (1962) y Medgyessy (1961), para mixturas binomiales.

Definición. Sea $y = (y_1, y_2, \ldots, y_n)$ observaciones independientes de una variable aleatoria, cuya función de densidad $f(y|\boldsymbol{\psi})$ es una mixtura, entonces la funcion

$$L(\boldsymbol{\psi}|y) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\psi}) = \prod_{i=1}^{n} \sum_{k=1}^{g} \pi_k f_k(y_i|\boldsymbol{\theta_k})$$

recibe el nombre de función de verosimilitud de la mixtura. Tomando logaritmo natural en $L(\boldsymbol{\psi}|y)$ obtenemos su función log-verosimilitud

$$l(\boldsymbol{\psi}|y) = \log L(\boldsymbol{\psi}|y) = \log \prod_{i=1}^{n} \left\{ \sum_{k=1}^{g} \pi_k f_k(y_i|\boldsymbol{\theta_k}) \right\}$$

$$= \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{g} \pi_k f_k(y_i|\boldsymbol{\theta_k}) \right\}.$$

Para calcular el estimador de máxima verosimilitud $\hat{\psi}$ es común utilizar el logaritmo de la función de verosimilitud, pues recordemos que la función y el logaritmo de la función bajo ciertas condiciones de regularidad toman en el mismo punto su máximo. Por lo tanto, debemos resolver la ecuación de verosimilitud

$$\frac{\partial}{\partial \boldsymbol{\psi}} \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{g} \pi_k f_k(y_i|\boldsymbol{\theta_k}) \right\} = 0$$

Debido a la presencia del logaritmo de una suma, es difícil la resolución de la ecuación, por tal motivo se requiere otro tipo de procedimiento.

Definición. Sea $y = (y_1, y_2, \ldots, y_n)$ una muestra observada de tamaño n, a la que denominaremos vector de datos incompletos, correspondientes a una realización de Y, con función de densidad $f(y|\boldsymbol{\psi})$, donde $\boldsymbol{\psi}$ es el vector de parámetros a estimar. Ahora considere la variable $Z = (Z_1, Z_2, \ldots, Z_n)$ que denominaremos latente, que representa a los datos no observados y cuya realización es $z = (z_1, z_2, \ldots, z_n)$. Entonces el vector aleatorio $X = (Y, Z)$ recibe el nombre de vector de datos completos y su realización es $x_1 = (y_1, z_1), x_2 = (y_2, z_2), \ldots, x_n = (y_n, z_n)$ de tal forma que a cada realización y_i le corresponde siempre una z_i.

En este contexto podemos suponer que z_i representa una variable indicadora binaria g-dimensional cuyo elemento j-esimo z_{ij} indica la pertenencia de la observación y_i a la componente j-esima de la mixtura donde $i = 1,2 \ldots, n \;\; y \;\; j = 1,2, \ldots, g$. Así que podemos definir $z_{ij} \in \{0,1\}$ como

$$z_{ij} = \begin{cases} 1 & y_i - proviene\ de\ la\ componente\ j - esima \\ 0 & en\ otro\ caso \end{cases}$$

Podemos representar al vector de datos incompletos y como un vector transpuesto

$$y' = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

y representar en una matriz a los datos completos

$$x' = (y', z') = \begin{pmatrix} y_1 & z_1 \\ y_2 & z_2 \\ \vdots & \vdots \\ y_n & z_n \end{pmatrix} = \begin{pmatrix} y_1 & z_{11} & \cdots & z_{1g} \\ y_2 & z_{21} & \cdots & z_{2g} \\ \vdots & \vdots & \vdots & \vdots \\ y_n & z_{n1} & \cdots & z_{ng} \end{pmatrix}$$

Cada variable z_i toma (g-1) valores iguales a cero y un único valor igual a uno.

Dada la naturaleza categórica de la variable z_i al indicar la pertenencia de los puntos muestrales a una componente u otra de la mixtura, se puede suponer que z_i sigue una distribución multinomial de una realización sobre g categorías con probabilidades $\pi = (\pi_1, \pi_2, \ldots, \pi_g)$, es decir,

$$P(Z_i = z_i) = \binom{1}{z_{i1}, z_{i2}, \ldots, z_{ig}} {\pi_1}^{z_{i1}} {\pi_2}^{z_{i2}} \ldots {\pi_g}^{z_{ig}} = \prod_{k=1}^{g} {\pi_k}^{z_{ik}}$$

Donde

$$\sum_{k=1}^{g} z_{ik} = 1 \qquad \sum_{k=1}^{g} \sum_{i=1}^{n} z_{ik} = n$$

Función log verosimilitud de los datos completos

La función de densidad conjunta de una observación completa es

$$f(x_i) = f(y_i, z_i) = f(y_i|z_i)p(z_i)$$

Pero

$$f_k(y_i|z_{ik} = 1) \rightarrow f_k(y_i|\theta_k)$$

Aquí $z_{ik} = 1$ si la observación y_i proviene de la componente k.

Desarrollando la distribución de Y_i con todos los estados posibles de z_k tenemos

$$f_k(y_i, z_i) = f_k\big(Y_i = y_i, Z_{i1} = z_{i1}, \dots, Z_{ig} = z_{ig}\big) =$$

$$f_k(Y_i = y_i \big| Z_{i1} = z_{i1}, \dots, Z_{ig} = z_{ig}) P(Z_{i1} = z_{i1}, \dots, Z_{ig} = z_{ig})$$

$$= \left\{\prod_{k=1}^{g} f_k(y_i|\boldsymbol{\theta_k})^{z_{ik}}\right\}\left\{\prod_{k=1}^{g} {\pi_k}^{z_{ik}}\right\} = \prod_{k=1}^{g}[\pi_k f_k(y_i|\boldsymbol{\theta_k})]^{z_{ik}}$$

De aquí la función de verosimilitud conjunta para todos los valores observados y y todos los valores no observados z es

$$\prod_{i=1}^{n}\prod_{k=1}^{g}[\pi_k f_k(y_i|\boldsymbol{\theta_k})]^{z_{ik}}$$

Aplicando la función logaritmo tenemos

$$l(\boldsymbol{\psi}|y, z) = \log L(\boldsymbol{\psi}|y, z) = \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} log[\pi_k f_k(y_i|\boldsymbol{\theta_k})]$$

$$\sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik}[log\pi_k + log f_k(y_i|\boldsymbol{\theta_k})]$$

$$\sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} log f_k(y_i|\boldsymbol{\theta_k})$$

Aquí $\boldsymbol{\psi} = \big(\pi_1, \dots, \pi_g, \boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_g}\big)$.

3.3 Parámetros de la mixtura Gaussiana $\boldsymbol{\theta_k = (\mu_k, \sigma_k{}^2)}$

$$l(\boldsymbol{\psi}|y,z) = \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} log\varphi(y_i|\boldsymbol{\mu_k}, \boldsymbol{\sigma_k}{}^2) \; =$$

$$\sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} log\pi_k - \frac{n}{2} log2\pi - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik}\left[log\boldsymbol{\sigma_k}{}^2 + \frac{(\boldsymbol{y_i} - \boldsymbol{\mu_k})^2}{\boldsymbol{\sigma_k}{}^2}\right]$$

3.4 Agrupamiento del modelo de mixturas Gaussianas

Una vez definidas las variables z_{ik} puede introducirse ahora el concepto de agrupamiento sobre los datos observados. Uno de los propósitos de los modelos mixtos es el de proporcionar una partición de los datos en g grupos, siendo g un numero previamente establecido. La k-esima proporción de la mixtura π_k puede interpretarse como la probabilidad a priori de que una observación muestral pertenezca a la población k así;

$$P(z_{ik} = 1) = \pi_k \qquad k = 1,2,..,g$$

Recuerde que podemos obtener información acerca de los datos ausentes mediante la observación de las elecciones que se realizaron. Por ejemplo, si el ingreso de una persona no está disponible, pero observamos que la persona va comprado un automóvil costoso se puede inferir que es probable que los ingresos de esta persona están por encima de la media.

Calculamos las probabilidades de los datos ausentes condicionada a las elecciones observadas en la muestra.

$$P(z_{ik} = 1|Y_i = y_i) = \frac{P(z_{ik} = 1)P(Y_i = y_i|z_{ik} = 1)}{P(Y_i = y_i)}$$

$$= \frac{\pi_k f_k(y_i|\boldsymbol{\theta_k})}{\sum_{k=1}^{g} \pi_k f_k(y_i|\boldsymbol{\theta_k})}$$

Si definimos $h(z|y,\boldsymbol{\psi})$ como la densidad de los datos ausentes condicionada a las elecciones observadas en la muestra. Entonces

$$h(z|y,\boldsymbol{\psi}) = P(z_{ik} = 1|Y_i = y_i) = \frac{\pi_k f_k(y_i|\boldsymbol{\theta_k})}{\sum_{k=1}^{g} \pi_k f_k(y_i|\boldsymbol{\theta_k})}.$$

El procedimiento EM es iterativo y consiste en lo siguiente; comenzamos con el cálculo de la esperanza condicional dado los datos observados y y utilizando datos iniciales $\boldsymbol{\psi}^0$, es decir calculamos $h(z|y,\boldsymbol{\psi}^0)$. Una vez hecho estos cálculos, definimos una nueva función en $\boldsymbol{\psi}^t$ que se relacione con la función de verosimilitud pero que utiliza la distribución condicionada $h(z|y,\boldsymbol{\psi})$, esta nueva función es

$$\mathcal{E}(\boldsymbol{\psi}|\boldsymbol{\psi}^0) = E[l(\boldsymbol{\psi}|\boldsymbol{y},\boldsymbol{z})|\boldsymbol{Y} = \boldsymbol{y},\boldsymbol{\psi}^0)]$$

$$\mathcal{E}(\boldsymbol{\psi}|\boldsymbol{\psi}^0) = E\left[\sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} log[\pi_k f_k(y_i|\boldsymbol{\theta_k})]|\boldsymbol{y},\boldsymbol{z})|\boldsymbol{Y} = \boldsymbol{y},\boldsymbol{\psi}^0\right]$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{g} E[z_{ik}|\boldsymbol{Y_i} = \boldsymbol{y_i},\boldsymbol{\psi}^0]\,[log\pi_k + log f_k(y_i|\boldsymbol{\theta_k})]$$

Sin embargo:

$$E[z_{ik}|\boldsymbol{Y_i} = \boldsymbol{y_i},\boldsymbol{\psi}^0] = P(z_{ik} = 1|Y_i = y_i,\boldsymbol{\psi}^0)$$

$$= \frac{f_k(Y_i = y_i|z_{ik} = 1)P(z_{ik} = 1)}{P(Y_i = y_i)}\Bigg|_{\boldsymbol{\psi}^0}$$

$$= \frac{\pi_k f_k(y_i|\boldsymbol{\theta_k})}{\sum_{k=1}^{g} \pi_k f_k(y_i|\boldsymbol{\theta_k})}\Bigg|_{\boldsymbol{\psi}^0} = \hat{\tau}_{ik}{}^{(0)}$$

Por lo tanto

$$\mathcal{E}(\boldsymbol{\psi}|\boldsymbol{\psi}^0) = \sum_{i=1}^{n}\sum_{k=1}^{g} \hat{\tau}_{ik}{}^{(0)} \left[log\pi_k + log f_k(y_i|\boldsymbol{\theta_k})\right]$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{g} \hat{\tau}_{ik}{}^{(0)}\, log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{g} \hat{\tau}_{ik}{}^{(0)}\, log f_k(y_i|\boldsymbol{\theta_k})$$

Después del cálculo anterior se realiza la maximización de la función $\mathcal{E}$ respecto de $\boldsymbol{\psi}$. Esta maximización se realiza en dos partes dado que π_k aparece únicamente en el primer sumando y que $\boldsymbol{\theta_k}$ lo hace el el segundo sumando.

Comenzamos con la maximización del primer sumando: Para este caso usamos los multiplicadores de Lagrange

$$\frac{\partial}{\partial \pi_k}\left(\sum_{i=1}^{n}\sum_{k=1}^{g} \hat{\tau}_{ik}{}^{(0)}\, log\pi_k + \lambda\left[\sum_{k=1}^{g} \pi_k - 1\right]\right) = 0$$

$$\sum_{i=1}^{n} \hat{\tau}_{ik}{}^{(0)} \frac{1}{\pi_k} + \lambda = 0$$

$$\sum_{i=1}^{n} \hat{\tau}_{ik}{}^{(0)} = -\lambda\pi_k$$

Tomando la suma sobre k en ambos lados de la última igualdad obtenemos

$$n = \sum_{i=1}^{n}\sum_{k=1}^{g} \hat{\tau}_{ik}{}^{(0)} = \sum_{k=1}^{g} -\lambda\pi_k = -\lambda$$

Lo cual implica que

$$\hat{\pi}_k{}^{(1)} = \pi_k = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_{ik}{}^{(0)}$$

Para la maximización del segundo sumando respecto de $\boldsymbol{\theta_k}$ depende de la función de densidad $f_k(y_i|\boldsymbol{\theta_k})$ que son densidades Gaussianas

$$log f_k(y_i|\boldsymbol{\theta_k}) = log\varphi(y_i|\boldsymbol{\mu_k}, \boldsymbol{\sigma_k}^2) =$$

$$= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(\boldsymbol{y_i} - \boldsymbol{\mu_k})^2}{\boldsymbol{\sigma_k}^2}$$

$$= -\frac{1}{2}\log(2\pi) - log\sigma^2 - \frac{1}{2}\frac{(\boldsymbol{y_i} - \boldsymbol{\mu_k})^2}{\boldsymbol{\sigma_k}^2}$$

Empezamos derivando respecto de μ

$$\frac{\partial}{\partial \mu_k}\sum_{i=1}^{n}\sum_{k=1}^{g}\hat{\tau}_{ik}{}^{(0)}\left(-\frac{1}{2}\log(2\pi)-log\sigma_k-\frac{1}{2}\frac{(\boldsymbol{y_i}-\boldsymbol{\mu_k})^2}{\boldsymbol{\sigma_k}{}^2}\right)=0$$

$$2\sum_{i=1}^{n}\hat{\tau}_{ik}{}^{(0)}\left(\frac{\boldsymbol{y_i}-\boldsymbol{\mu_k}}{\boldsymbol{2\sigma_k}{}^2}\right)=\boldsymbol{0}$$

$$\sum_{i=1}^{n}\hat{\tau}_{ik}{}^{(0)}\,\boldsymbol{y_i}=\sum_{i=1}^{n}\hat{\tau}_{ik}{}^{(0)}\,\boldsymbol{\mu_k}$$

$$\hat{\mu}_k{}^{(1)}=\mu_k=\frac{\sum_{i=1}^{n}\hat{\tau}_{ik}{}^{(0)}\,\boldsymbol{y_i}}{\sum_{i=1}^{n}\hat{\tau}_{ik}{}^{(0)}}$$

Para obtener el estimador de $\sigma_k{}^2$ tenemos

$$\frac{\partial}{\partial \sigma_k{}^2}\sum_{i=1}^{n}\sum_{k=1}^{g}\hat{\tau}_{ik}{}^{(0)}\left(-\frac{1}{2}\log(2\pi)-\frac{log\sigma_k{}^2}{2}-\frac{1}{2}\frac{(\boldsymbol{y_i}-\boldsymbol{\mu_k})^2}{\boldsymbol{\sigma_k}{}^2}\right)=0$$

$$-\sum_{i=1}^{n}\hat{\tau}_{ik}{}^{(0)}\frac{\boldsymbol{1}}{\boldsymbol{2\sigma_k}{}^2}+\sum_{i=1}^{n}\hat{\tau}_{ik}{}^{(0)}\frac{(\boldsymbol{y_i}-\boldsymbol{\mu_k})^2}{\boldsymbol{2(\sigma_k}{}^2)^2}=\boldsymbol{0}$$

$$\sum_{i=1}^{n} \hat{\tau}_{ik}{}^{(0)} \frac{(\boldsymbol{y_i} - \boldsymbol{\mu_k})^2}{\boldsymbol{\sigma_k}^2} = \sum_{i=1}^{n} \hat{\tau}_{ik}{}^{(0)}$$

$$\boldsymbol{\sigma_k}^2 = \frac{\sum_{i=1}^{n} \hat{\tau}_{ik}{}^{(0)} (\boldsymbol{y_i} - \boldsymbol{\mu_k})^2}{\sum_{i=1}^{n} \hat{\tau}_{ik}{}^{(0)}}$$

Usando la estimación de μ_k tenemos

$$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{k}}{}^{2(1)} = \frac{\sum_{i=1}^{n} \hat{\tau}_{ik}{}^{(0)} (\boldsymbol{y_i} - \boldsymbol{\mu_k})^2}{\sum_{i=1}^{n} \hat{\tau}_{ik}{}^{(0)}}$$

3.5 Criterio de parado

Para detener la iteración puede considerarse la diferencia absoluta

$$\left| l(\boldsymbol{\psi}^{(t+1)}|y) - l(\boldsymbol{\psi}^{(t)}|y) \right|$$

O la diferencia relativa

$$\frac{\left| l(\boldsymbol{\psi}^{(t+1)}|y) - l(\boldsymbol{\psi}^{(t)}|y) \right|}{|l(\boldsymbol{\psi}^{(t)}|y)|}$$

Se utiliza más la diferencia relativa por su adimensionalidad. El valor máximo de dicha diferencia menor que 10^{-5}.

3.6 Valores iniciales

Los valores iniciales en este trabajo, para el cálculo de las medias de cada componente, se divide la muestra en g particiones y sobre cada una de ellas se calcula la media de las observaciones que contienen ver Finch(1989).

Estos valores representan $\hat{\mu}_1{}^{(0)}, \hat{\mu}_2{}^{(0)}, \ldots, \hat{\mu}_g{}^{(0)}$ es decir las medias de cada componente sobre las que el algoritmo comienza a iterar. Para las varianzas Finch(1989) obtuvo una común $\hat{\sigma}_1{}^{(0)}$ ponderada según $\hat{\mu}_1{}^{(0)}, \hat{\mu}_2{}^{(0)}, \ldots, \hat{\mu}_g{}^{(0)}$ y en cuanto a las proporciones iniciales utiliza g números aleatorios extraídos de una distribución $U(0,1)$. En este trabajo para el calculo de la desviación típica se procedió como con las medias, calculando la correspondiente para cada partición y respecto a las proporciones todas semejantes según $\pi_1{}^{(0)} = \pi_2{}^{(0)} = \cdots = \pi_g{}^{(0)} = 1/g$.

CAPITULO 4. IMPLEMENTACIÓN DEL ALGORITMO

El aprendizaje supervisado actúa como una guía para enseñar al algoritmo las conclusiones a las que debe llegar, es decir la salida del algoritmo ya es conocida. Requiere que los posibles resultados del algoritmo ya sean conocidos y que los datos utilizados para entrenar el algoritmo ya estén etiquetados con las respuestas correctas.

En el aprendizaje no supervisado no existe un conjunto de datos de entrenamiento y los resultados son desconocidos, prácticamente se entra en el problema a ciegas y se intentar encontrar algún tipo de organización que simplifique el análisis, aunque parezca increíble, el aprendizaje no supervisado es la capacidad de resolver problemas complejos utilizando solo los datos de entrada y algoritmos lógicos, y en ningún momento se tiene datos de referencias.

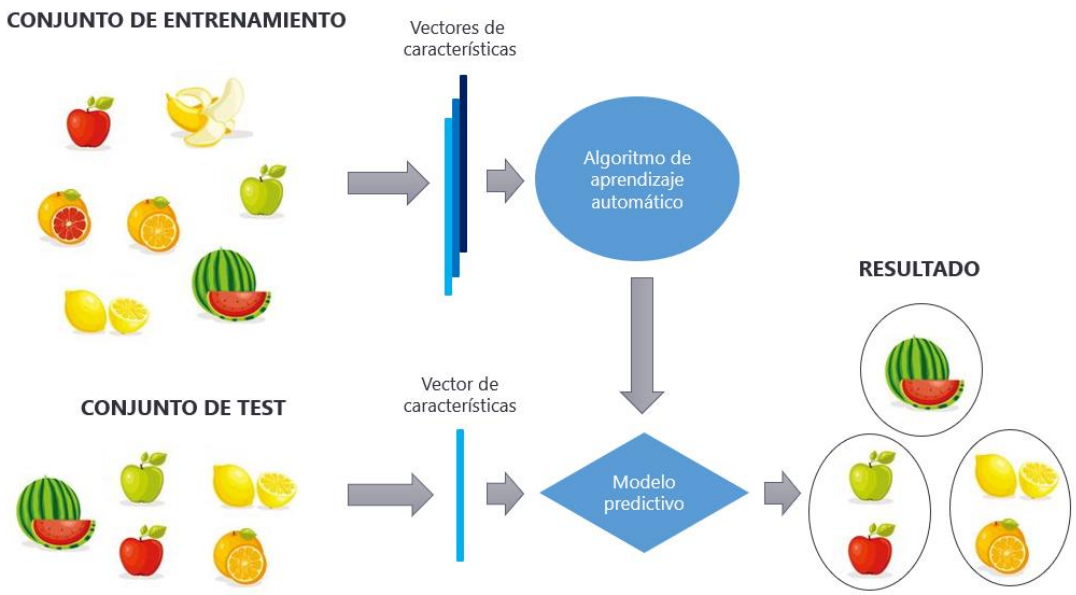

Fig. 4 Aprendizaje no supervisado.

4.1 Clustering

El clustering es una técnica de inteligencia artificial dentro del Aprendizaje No Supervisado. Esto es así debido a que de antemano no se conocen las clases o el número de clases que se tendrán como resultado de la agrupación de dichos datos.

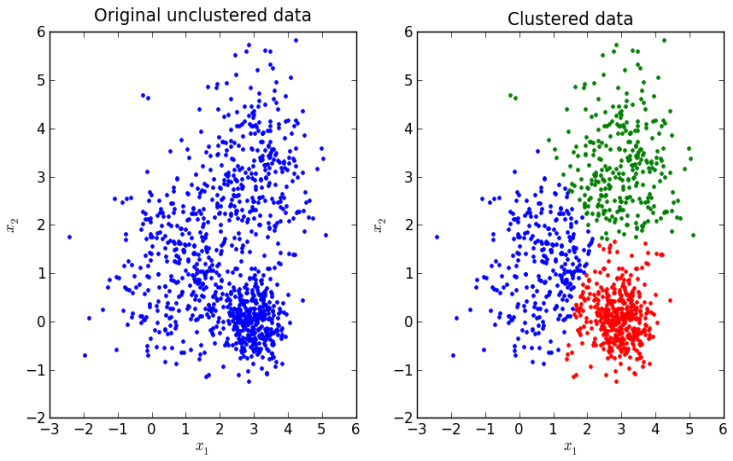

Fig. 4.1 Un cluster será cada uno de los grupos a los que pertenezcan un conjunto de datos tras la clasificación.

4.2 Tipos Clustering

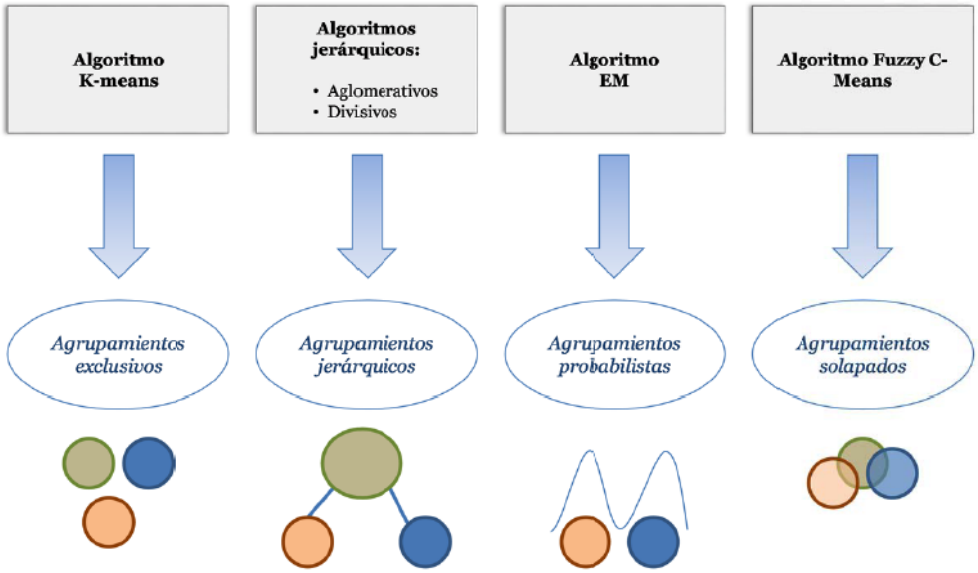

Fig. 4.2 Existen diferentes tipos de clustering en función del tipo de agrupamientos que producen.

Agrupamientos Exclusivos

Se generan por métodos que particionan los datos creando un número k de clusters, utilizando, medidas de distancia para generarlos.

Agrupamientos Jerárquicos

Algoritmos que generan una estructura jerárquica de clusters a través de creación de clusters durante iteraciones.

Agrupamientos Probabilistas

Los clusteres se generan mediante algún método probabilístico. El ejemplo más representativo sea posiblemente el algoritmo Expectación-Maximización (EM)**.**

Agrupamientos Solapados

Los objetos se agrupan en clusters difusos, donde un objeto puede pertenecer a más de un cluster al mismo tiempo o puede estar en varios cluster con diferentes grados de permanencia**.**

Por la naturaleza de los datos nosotros proponemos que estos se deben clasificar en dos clases (maligno y benignos), por tal razón procedemos ordenar nuestros datos de menor a mayor valor, luego los peticionamos en dos clases y a partir de estas clases procedemos a aplicar el algoritmo de maximización de la esperanza para que nos genere dos distribuciones gaussianas con sus respectivas medias y varianzas que esperamos sean la clase de los benignos y la clase de los malignos.

A continuación, describimos de manera completa el algoritmo.

4.3 Desarrollo del algoritmo

```
<!DOCTYPE HTML>

<html>

<head>

</head>

<body>
```

```
<script type="text/javascript">


function normalp(xy, miu, varianza) {

    var y=Math.pow(Math.E, (-Math.pow(xy-miu,2)) /(2*varianza)) /(Math.sqrt(2*Math.PI) *Math.sqrt(varianza));

    return y;

}

function opmh(){

    var phi1=0.5, phi2=0.5;

    var i,j;

var n1 =


[6.981,7.691,7.729,7.760,8.196,8.219,8.571,8.597,8.598,8.618,8.671,8.726,8.734,8.878,8.888,8.950,9.00
0,9.029,9.042,9.173,9.268,9.295,9.333,9.397,9.405,9.423,9.436,9.465,9.504,9.567,9.606,9.667,9.668,9.6
76,9.683,9.720,9.731,9.738,9.742,9.742,9.755,9.777,9.787,9.847,9.876,9.876,9.904,10.030,10.050,10.08
0,10.160,10.170,10.180,10.200,10.250,10.260,10.260,10.260,10.290,10.320,10.440,10.480,10.480,10.49
0,10.490,10.510,10.510,10.570,10.570,10.600,10.650,10.660,10.710,10.750,10.800,10.800,10.820,10.86
0,10.880,10.900,10.910,10.940,10.960,10.970,11.040,11.040,11.060,11.060,11.060,11.080,11.130,11.13
0,11.140,11.150,11.160,11.200,11.220,11.220,11.250,11.260,11.260,11.270,11.270,11.280,11.290,11.30
0,11.310,11.320,11.330,11.340,11.340,11.360,11.370,11.410,11.410,11.430,11.430,11.450,11.460,11.47
0,11.490,11.500,11.510,11.520,11.520,11.540,11.540,11.570,11.600,11.600,11.600,11.610,11.620,11.63
0,11.640,11.660,11.670,11.680,11.690,11.700,11.710,11.710,11.710,11.740,11.740,11.750,11.750,11.76
0,11.800,11.810,11.840,11.850,11.870,11.890,11.890,11.890,11.900,11.930,11.930,11.940,11.940,11.95
0,11.990,12.000,12.000,12.030,12.040,12.050,12.050,12.060,12.060,12.070,12.100,12.160,12.180,12.18
0,12.180,12.190,12.200,12.210,12.210,12.220,12.230,12.250,12.250,12.270,12.270,12.300,12.300,12.31
0,12.320,12.340,12.340,12.340,12.360,12.360,12.390,12.400,12.420,12.430,12.450,12.460,12.460,12.47
0,12.470,12.490,12.540,12.540,12.560,12.580,12.620,12.620,12.630,12.650,12.670,12.700,12.720,12.72
0,12.750,12.760,12.760,12.770,12.770,12.780,12.800,12.810,12.830,12.850,12.860,12.860,12.870,12.87
0,12.880,12.880,12.890,12.890,12.890,12.900,12.910,12.940,12.950,12.960,12.980,12.990,13.000,13.00
0,13.010,13.030,13.050,13.050,13.050,13.080,13.110,13.140,13.150,13.160,13.170,13.200,13.200,13.21
0,13.210,13.240,13.270,13.270,13.280,13.300,13.340,13.370,13.380,13.400,13.450,13.460,13.460,13.47
0,13.490,13.500,13.510,13.530,13.540,13.560,13.590,13.590,13.620,13.640,13.640];

var n2 =


[13.650,13.660,13.660,13.680,13.690,13.700,13.710,13.740,13.750,13.770,13.780,13.850,13.850,13.850,
13.870,13.870,13.880,13.900,13.900,13.940,14.020,14.030,14.040,14.050,14.060,14.110,14.200,14.220,
14.260,14.260,14.290,14.340,14.400,14.410,14.420,14.440,14.470,14.500,14.530,14.530,14.580,14.590,
14.610,14.620,14.640,14.640,14.690,14.740,14.760,14.800,14.810,14.860,14.870,14.920,14.950,14.960,
14.970,14.970,14.990,15.000,15.040,15.100,15.190,15.270,15.710,15.730,16.140,16.170,16.300,16.500,
16.840,17.850,10.950,11.080,11.420,11.760,11.800,11.840,12.340,12.450,12.460,12.680,12.770,12.830,
13.000,13.110,13.170,13.170,13.280,13.400,13.430,13.440,13.480,13.610,13.610,13.710,13.730,13.770,
13.800,13.810,13.820,13.860,13.960,13.980,14.190,14.220,14.250,14.250,14.270,14.420,14.450,14.480,
```

14.540,14.580,14.600,14.680,14.710,14.780,14.860,14.870,14.900,14.950,14.990,15.050,15.060,15.080,
15.100,15.120,15.130,15.220,15.280,15.300,15.320,15.340,15.370,15.460,15.460,15.460,15.490,15.500,
15.530,15.610,15.660,15.700,15.750,15.750,15.780,15.780,15.850,16.020,16.030,16.070,16.110,16.130,
16.130,16.160,16.240,16.250,16.260,16.270,16.350,16.460,16.600,16.650,16.690,16.740,16.780,17.010,
17.020,17.050,17.060,17.080,17.140,17.190,17.200,17.270,17.290,17.300,17.350,17.420,17.460,17.470,
17.540,17.570,17.600,17.680,17.750,17.910,17.930,17.950,17.990,17.990,18.010,18.030,18.050,18.080,
18.220,18.220,18.250,18.310,18.310,18.450,18.460,18.490,18.610,18.630,18.650,18.660,18.770,18.810,
18.820,18.940,19.000,19.020,19.070,19.100,19.160,19.170,19.180,19.190,19.210,19.270,19.400,19.400,
19.440,19.450,19.530,19.530,19.550,19.550,19.590,19.590,19.680,19.690,19.730,19.790,19.800,19.810,
19.890,20.090,20.130,20.160,20.180,20.180,20.200,20.260,20.290,20.310,20.340,20.440,20.470,20.480,
20.510,20.550,20.570,20.580,20.590,20.600,20.640,20.730,20.920,20.940,21.090,21.100,21.160,21.370,
21.560,21.610,21.710,21.750,22.010,22.270,23.090,23.210,23.270,23.290,23.510,24.250,24.630,25.220,
25.730,27.220,27.420,28.110];

var n =

[6.981,7.691,7.729,7.760,8.196,8.219,8.571,8.597,8.598,8.618,8.671,8.726,8.734,8.878,8.888,8.950,9.00
0,9.029,9.042,9.173,9.268,9.295,9.333,9.397,9.405,9.423,9.436,9.465,9.504,9.567,9.606,9.667,9.668,9.6
76,9.683,9.720,9.731,9.738,9.742,9.742,9.755,9.777,9.787,9.847,9.876,9.876,9.904,10.030,10.050,10.08
0,10.160,10.170,10.180,10.200,10.250,10.260,10.260,10.260,10.290,10.320,10.440,10.480,10.480,10.49
0,10.490,10.510,10.510,10.570,10.570,10.600,10.650,10.660,10.710,10.750,10.800,10.800,10.820,10.86
0,10.880,10.900,10.910,10.940,10.960,10.970,11.040,11.040,11.060,11.060,11.060,11.080,11.130,11.13
0,11.140,11.150,11.160,11.200,11.220,11.220,11.250,11.260,11.260,11.270,11.270,11.280,11.290,11.30
0,11.310,11.320,11.330,11.340,11.340,11.360,11.370,11.410,11.410,11.430,11.430,11.450,11.460,11.47
0,11.490,11.500,11.510,11.520,11.520,11.540,11.540,11.570,11.600,11.600,11.600,11.610,11.620,11.63
0,11.640,11.660,11.670,11.680,11.690,11.700,11.710,11.710,11.710,11.740,11.740,11.750,11.750,11.76
0,11.800,11.810,11.840,11.850,11.870,11.890,11.890,11.890,11.900,11.930,11.930,11.940,11.940,11.95
0,11.990,12.000,12.000,12.030,12.040,12.050,12.050,12.060,12.060,12.070,12.100,12.160,12.180,12.18
0,12.180,12.190,12.200,12.210,12.210,12.220,12.230,12.250,12.250,12.270,12.270,12.300,12.300,12.31
0,12.320,12.340,12.340,12.340,12.360,12.360,12.390,12.400,12.420,12.430,12.450,12.460,12.460,12.47
0,12.470,12.490,12.540,12.540,12.560,12.580,12.620,12.620,12.630,12.650,12.670,12.700,12.720,12.72
0,12.750,12.760,12.760,12.770,12.770,12.780,12.800,12.810,12.830,12.850,12.860,12.860,12.870,12.87
0,12.880,12.880,12.890,12.890,12.890,12.900,12.910,12.940,12.950,12.960,12.980,12.990,13.000,13.00
0,13.010,13.030,13.050,13.050,13.050,13.080,13.110,13.140,13.150,13.160,13.170,13.200,13.200,13.21
0,13.210,13.240,13.270,13.270,13.280,13.300,13.340,13.370,13.380,13.400,13.450,13.460,13.460,13.47
0,13.490,13.500,13.510,13.530,13.540,13.560,13.590,13.590,13.620,13.640,13.640,13.650,13.660,13.66
0,13.680,13.690,13.700,13.710,13.740,13.750,13.770,13.780,13.850,13.850,13.850,13.870,13.870,13.88
0,13.900,13.900,13.940,14.020,14.030,14.040,14.050,14.060,14.110,14.200,14.220,14.260,14.260,14.29
0,14.340,14.400,14.410,14.420,14.440,14.470,14.500,14.530,14.530,14.580,14.590,14.610,14.620,14.64
0,14.640,14.690,14.740,14.760,14.800,14.810,14.860,14.870,14.920,14.950,14.960,14.970,14.970,14.99
0,15.000,15.040,15.100,15.190,15.270,15.710,15.730,16.140,16.170,16.300,16.500,16.840,17.850,10.95
0,11.080,11.420,11.760,11.800,11.840,12.340,12.450,12.460,12.680,12.770,12.830,13.000,13.110,13.17
0,13.170,13.280,13.400,13.430,13.440,13.480,13.610,13.610,13.710,13.730,13.770,13.800,13.810,13.82
0,13.860,13.960,13.980,14.190,14.220,14.250,14.250,14.270,14.420,14.450,14.480,14.540,14.580,14.60
0,14.680,14.710,14.780,14.860,14.870,14.900,14.950,14.990,15.050,15.060,15.080,15.100,15.120,15.13
0,15.220,15.280,15.300,15.320,15.340,15.370,15.460,15.460,15.460,15.490,15.500,15.530,15.610,15.66
0,15.700,15.750,15.750,15.780,15.780,15.850,16.020,16.030,16.070,16.110,16.130,16.130,16.160,16.24
0,16.250,16.260,16.270,16.350,16.460,16.600,16.650,16.690,16.740,16.780,17.010,17.020,17.050,17.06
0,17.080,17.140,17.190,17.200,17.270,17.290,17.300,17.350,17.420,17.460,17.470,17.540,17.570,17.60
0,17.680,17.750,17.910,17.930,17.950,17.990,17.990,18.010,18.030,18.050,18.080,18.220,18.220,18.25
0,18.310,18.310,18.450,18.460,18.490,18.610,18.630,18.650,18.660,18.770,18.810,18.820,18.940,19.00
0,19.020,19.070,19.100,19.160,19.170,19.180,19.190,19.210,19.270,19.400,19.400,19.440,19.450,19.53
0,19.530,19.550,19.550,19.590,19.590,19.680,19.690,19.730,19.790,19.800,19.810,19.890,20.090,20.13
0,20.160,20.180,20.180,20.200,20.260,20.290,20.310,20.340,20.440,20.470,20.480,20.510,20.550,20.57
0,20.580,20.590,20.600,20.640,20.730,20.920,20.940,21.090,21.100,21.160,21.370,21.560,21.610,21.71
0,21.750,22.010,22.270,23.090,23.210,23.270,23.290,23.510,24.250,24.630,25.220,25.730,27.220,27.42
0,28.110

```
];
        var varianza1, miu1;
        var varianza2, miu2;
        var s1 = 0, s2 = 0;
        for (i=0; i<n1.length; i++) {
            s1 += n1[i];
        }
        miu1= s1/n1.length;
        var s3 = 0;
        for (i=0; i<n1.length; i++) {
            s3 += Math.pow(n1[i]-miu1,2);
        }
        varianza1=(s3/(n1.length-1));

        for (i=0; i<n2.length; i++) {
            s2 += n2[i];
        }
        miu2= s2/n2.length;
        var s4 = 0;
        for (i=0; i<n2.length; i++) {
            s4 += Math.pow(n2[i]-miu2,2);
        }
        varianza2=(s4/ (n2.length-1));
        var t1 = [], st1=0;
```

```
var t2 = [], st2=0;

var st3 = 0, st4=0;

var st5=0, st6=0;

var miu1k, miu2k;

var varianza1k, varianza2k;


for (i=0; i<n.length;i++) {

t1[i] = phi1*normalp(n[i], miu1, varianza1) /(phi1*normalp(n[i], miu1, varianza1) +phi2*normalp(n[i], miu2, varianza2));

}

for (i=0; i<n.length;i++) {

t2[i] = phi2*normalp(n[i], miu2, varianza2) /(phi1*normalp(n[i], miu1, varianza1) +phi2*normalp(n[i], miu2, varianza2));

}

for (j=0; j<100; j++) {

st1=0; st2=0; st3=0; st4=0; st5=0; st6=0;

for (i=0; i<n.length;i++) {

st1+=t1[i];

st3+=t1[i]*n[i];

}

for (i=0; i<n.length;i++) {

st2+=t2[i];

st4+=t2[i]*n[i];

}
```

```
                phi1=st1/n.length;
                phi2=st2/n.length;
                miu1k=st3/st1;
                miu2k=st4/st2;
                for (i=0; i<n.length-1; i++) {
                      st5+=t1[i]*Math.pow((n[i]-miu1k),2);
                }
                for (i=0; i<n.length-1; i++) {
                      st6+=t2[i]*Math.pow((n[i]-miu2k),2);
                }
                varianza1k=st5/st1;
                varianza2k=st6/st2;
                for (i=0; i<n.length;i++) {
                      t1[i] = phi1*normalp(n[i], miu1k, varianza1k)
/(phi1*normalp(n[i], miu1k, varianza1k) +phi2*normalp(n[i], miu2k,
varianza2k));
                }
                for (i=0; i<n.length;i++) {
                      t2[i] = phi2*normalp(n[i], miu2k, varianza2k)
/(phi1*normalp(n[i], miu1k, varianza1k) +phi2*normalp(n[i], miu2k,
varianza2k));
                }
          }
          alert ("phi1: "+phi1);
          alert ("phi2: "+phi2);
          alert ("miu1k: "+miu1k);
```

```
            alert ("miu2k: "+miu2k);
            alert ("varianza1k: "+varianza1k);
            alert ("varianza2k: "+varianza2k);


        }
        </script>
        <button type="button" onclick="opmh()">Obtener Esperanza-Maximización</button>
</body>
</html>
```

CAPITULO 5. INTERPRETACIÓN DE LOS DATOS

Tabla 5 *Tabla de valores obtenidos.*

Parámetro	Valor Aproximado
π_1	0.7326
π_2	0.2674
μ_1	12.48
μ_2	18.64
σ_1^2	3.68
σ_2^2	7.88

Después de realizadas 1000 iteraciones del algoritmo obtenemos los siguientes valores.

5.1 Graficación de las normales

A continuación, graficamos cada una de las normales con los parámetros de la tabla anterior.

```
> plot( 1/((1.92)·sqrt(2·π)) · e^(-((x-12.48)^2)/(2·(3.68))), x = 0..20 );
```

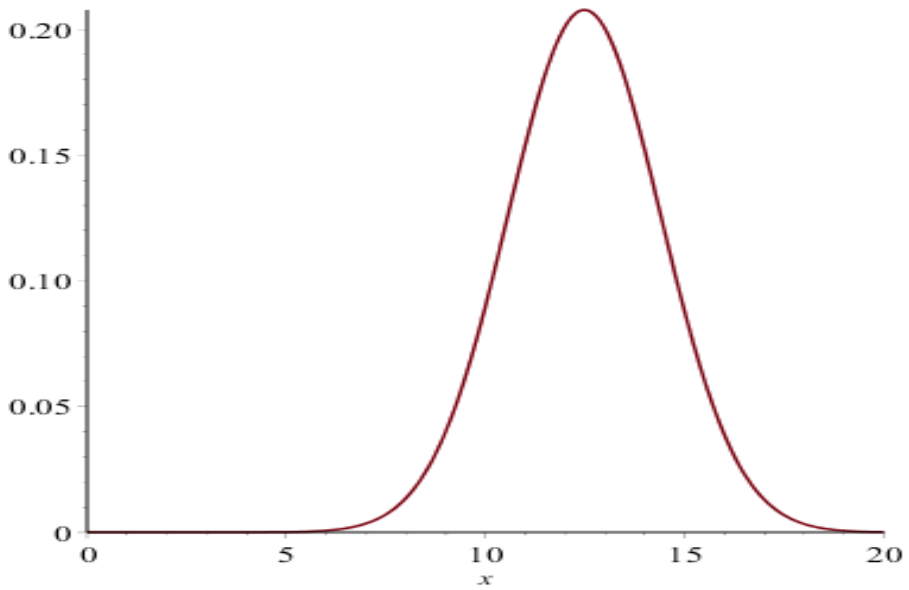

Fig. 5.1.1 Con los parámetros π_1=0.7326, μ_1=12.48, σ_1^2=3.68

```
> plot(1/((2.81)*sqrt(2*π))*e^(-((x-18.64)^2)/(2*(7.88))), x=5..40);
```

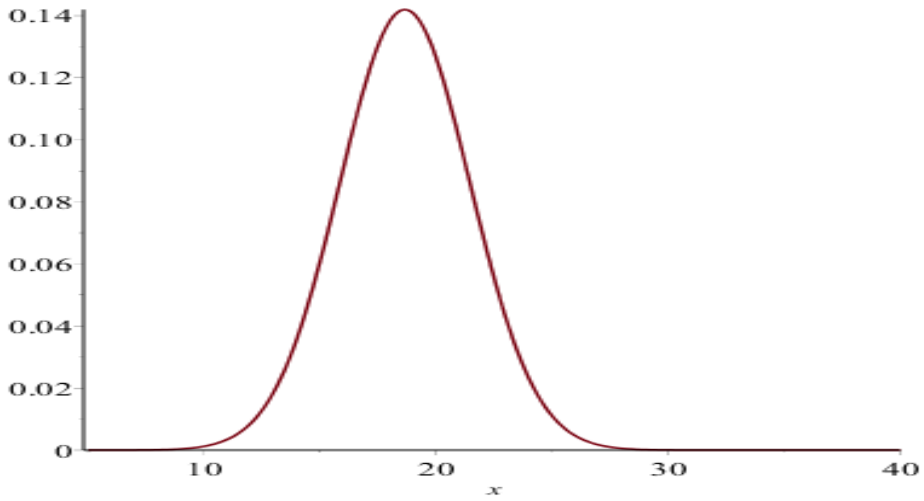

Fig. 5.1.2 Con los parámetros π_2=0.2674, μ_2=18.64,σ_2^2=7.88

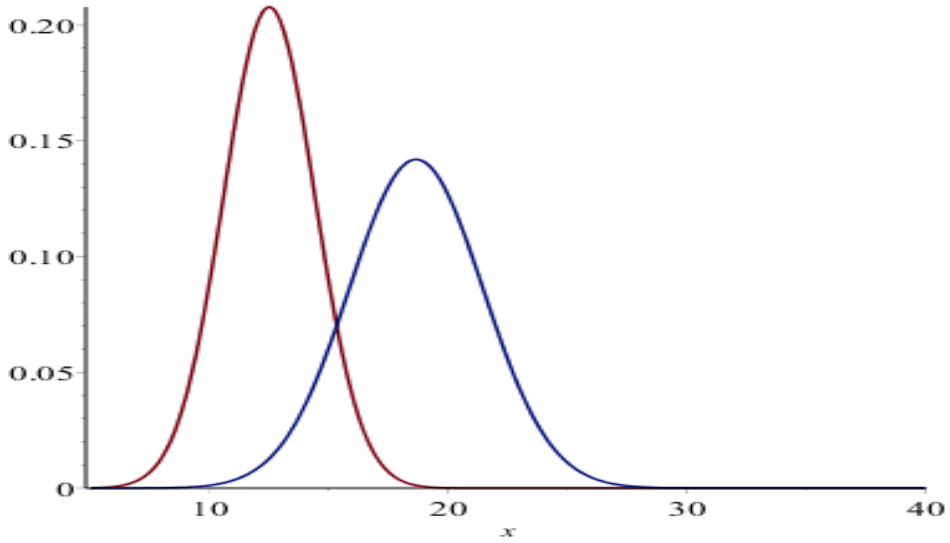

Fig. 5.1.3 Grafica de las curvas sobrepuestas.

5.2 Mezcla de las dos componentes normales

Ahora graficamos la mezcla de las dos normales usando todos los parámetros de la tabla.

$$plot\left(0.7326\cdot\left(\frac{1}{(1.92)\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(x-12.48)^2}{2\cdot(3.68)}}\right)+0.2674\cdot\left(\frac{1}{(2.81)\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(x-18.64)^2}{2\cdot(7.88)}}\right),x=5..40\right);$$

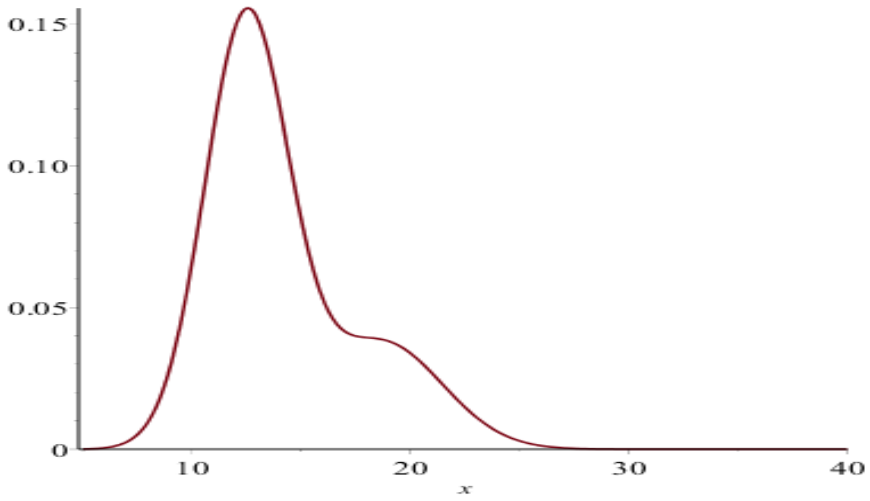

Fig. 5.2 Grafica de la mezcla de las dos normales.

5.3 Funciones de Responsabilidad

Al analizar nuevamente la tabla de datos y considerando además de la variable (*radius_mean*) el resultado de que si el tumor es maligno o benigno, encontramos que nuestra clasificación es adecuada para dividir los casos en base al tamaño del tumor. Existe algunos valores que se encuentran en el mismo rango y algunos son benignos y otros malignos, para hacer la discriminación de estos datos usamos de las funciones de densidad relativas llamadas la "responsabilidad" de cada cluster.

$$plot\left(\left[\frac{\frac{1}{(1.92)\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(x-12.48)^2}{2\cdot(3.68)}}}{\left(\frac{1}{(1.92)\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(x-12.48)^2}{2\cdot(3.68)}}\right)+\left(\frac{1}{(2.81)\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(x-18.64)^2}{2\cdot(7.88)}}\right)},\ \frac{\frac{1}{(2.81)\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(x-18.64)^2}{2\cdot(7.88)}}}{\left(\frac{1}{(1.92)\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(x-12.48)^2}{2\cdot(3.68)}}\right)+\left(\frac{1}{(2.81)\cdot\sqrt{2\cdot\pi}}\cdot e^{-\frac{(x-18.64)^2}{2\cdot(7.88)}}\right)}\right], x=10..20\right)$$

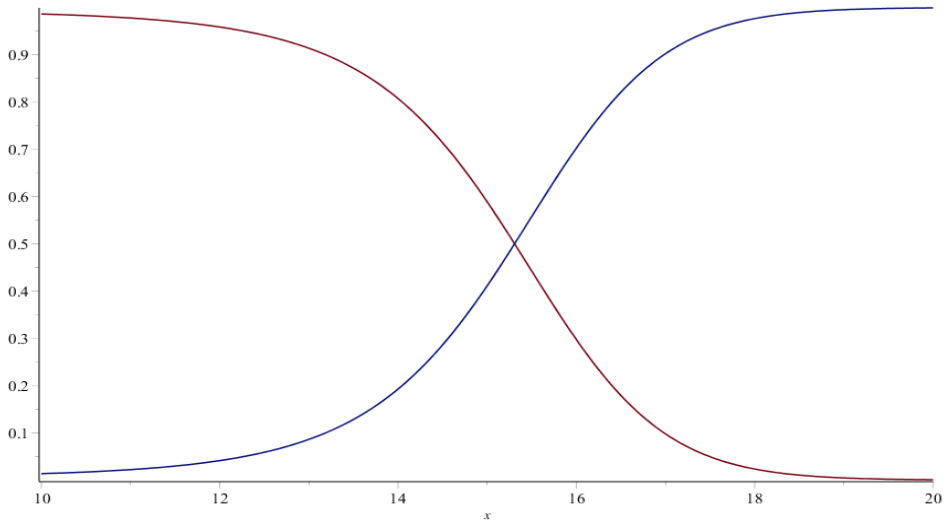

Fig. 5.3 Funciones de densidad relativas llamadas la "responsabilidad" de cada cluster.

5.4 Ubicación de los datos en la mezcla

A continuación, graficamos la mezcla de normales junto con cada punto de la base de datos donde se especifica en base al tamaño del tumor si es maligno o benigno, para darnos cuenta que el algoritmo discrimina muy bien esta relación.

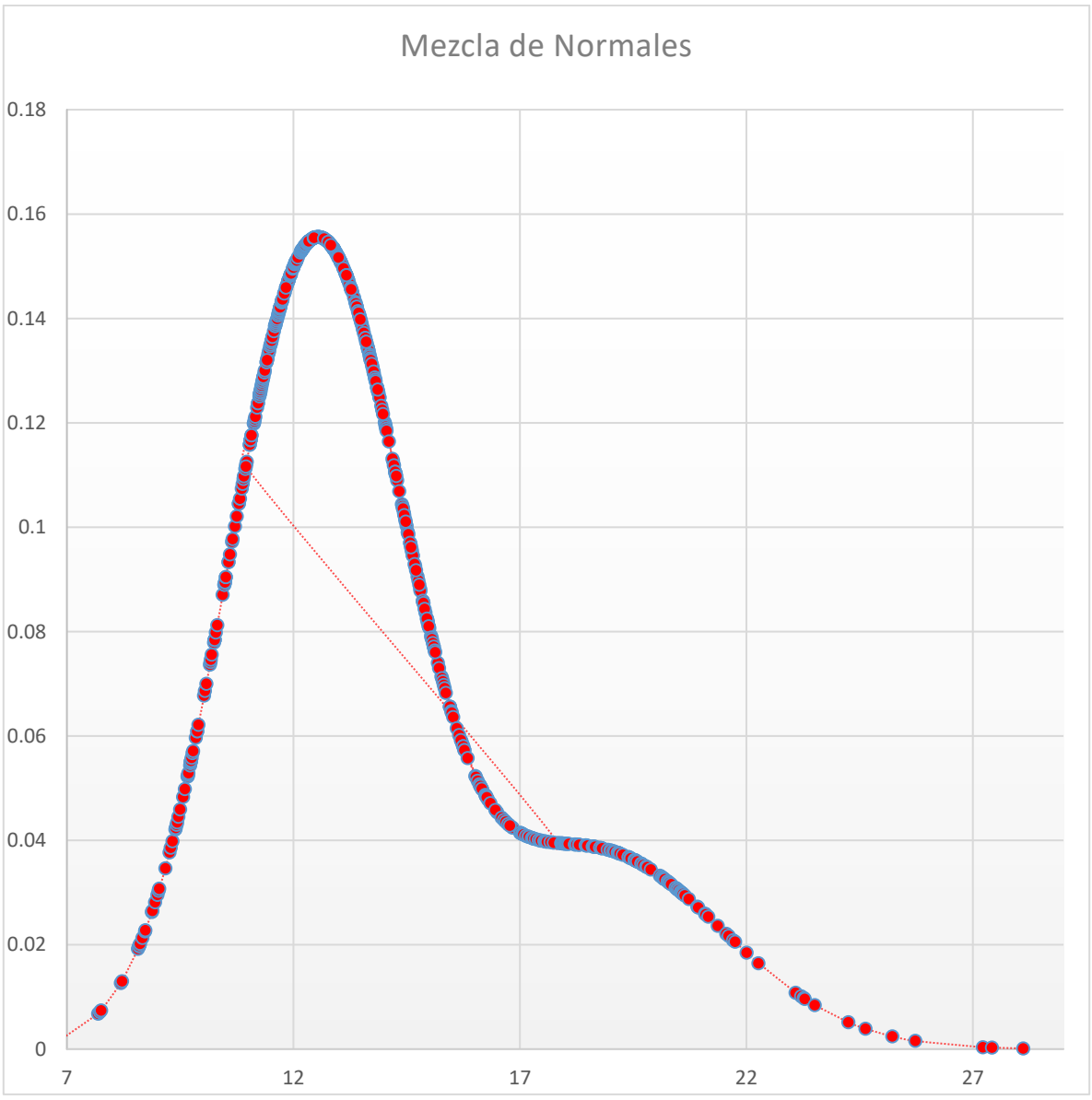

FIg. 5.4 Mezcla de normales junto con cada punto de la base de datos.

Tabla 5.4 *Valores benignos y malignos de menor a mayor*

	ID	M = maligno, B = benigno	radio	mezcla		ID	M = maligno, B = benigno	radio	mezcla
1	862722	B	6.98	0.003	47	872608	B	9.904	0.062
2	921362	B	7.69	0.007	48	907367	B	10.03	0.068
3	921092	B	7.73	0.007	49	897880	B	10.05	0.069
4	92751	B	7.76	0.007	50	874158	B	10.08	0.070
5	85713702	B	8.2	0.013	51	924964	B	10.16	0.074
6	871001502	B	8.22	0.013	52	858970	B	10.17	0.074
7	91805	B	8.57	0.019	53	8812844	B	10.18	0.075
8	894047	B	8.6	0.020	54	8811779	B	10.2	0.076
9	858981	B	8.6	0.020	55	894604	B	10.25	0.078
10	858477	B	8.62	0.020	56	898677	B	10.26	0.078
11	872113	B	8.67	0.021	57	90317302	B	10.26	0.078
12	864496	B	8.73	0.023	58	922840	B	10.26	0.078
13	903483	B	8.73	0.023	59	924934	B	10.29	0.080
14	9010333	B	8.88	0.026	60	922577	B	10.32	0.081
15	859711	B	8.89	0.026	61	88147101	B	10.44	0.087
16	864726	B	8.95	0.028	62	884437	B	10.48	0.089
17	89346	B	9	0.029	63	907409	B	10.48	0.089
18	859471	B	9.03	0.030	64	862989	B	10.49	0.090
19	894329	B	9.04	0.031	65	892657	B	10.49	0.090
20	859196	B	9.17	0.035	66	864292	B	10.51	0.090
21	915186	B	9.27	0.038	67	892399	B	10.51	0.090
22	917092	B	9.3	0.039	68	901315	B	10.57	0.093
23	924342	B	9.33	0.040	69	909777	B	10.57	0.093
24	905539	B	9.4	0.042	70	8910251	B	10.6	0.095
25	905978	B	9.41	0.042	71	88466802	B	10.65	0.097
26	925236	B	9.42	0.043	72	871642	B	10.66	0.098
27	901034301	B	9.44	0.043	73	8910720	B	10.71	0.100
28	859464	B	9.47	0.044	74	869254	B	10.75	0.102
29	8510824	B	9.5	0.046	75	87127	B	10.8	0.105
30	882488	B	9.57	0.048	76	90745	B	10.8	0.105
31	898143	B	9.61	0.050	77	923465	B	10.82	0.106
32	9113778	B	9.67	0.052	78	923748	B	10.86	0.107
33	9113514	B	9.67	0.052	79	911391	B	10.88	0.108
34	915276	B	9.68	0.053	80	871149	B	10.9	0.109
35	923169	B	9.68	0.053	81	891936	B	10.91	0.110
36	875099	B	9.72	0.054	82	904971	B	10.94	0.111
37	8710441	B	9.73	0.055	358	855563	M	10.95	0.112
38	868999	B	9.74	0.055	83	919537	B	10.96	0.112
39	8910996	B	9.74	0.055	84	909411	B	10.97	0.113
40	907145	B	9.74	0.055	85	901836	B	11.04	0.116
41	9112712	B	9.76	0.056	86	905520	B	11.04	0.116
42	864033	B	9.78	0.057	87	89827	B	11.06	0.117
43	862261	B	9.79	0.057	88	904302	B	11.06	0.117
44	917897	B	9.85	0.060	89	91544002	B	11.06	0.117
45	862980	B	9.88	0.061	90	871641	B	11.08	0.118
46	879804	B	9.88	0.061	359	9013838	M	11.08	0.118

	ID	M = maligno, B = benigno	radio	mezcla		ID	M = maligno, B = benigno	radio	mezcla
91	90769601	B	11.1	0.120	136	899187	B	11.66	0.141
92	923780	B	11.1	0.120	137	91903901	B	11.67	0.141
93	901011	B	11.1	0.120	138	913512	B	11.68	0.141
94	87106	B	11.2	0.121	139	919812	B	11.69	0.142
95	906290	B	11.2	0.121	140	893783	B	11.7	0.142
96	925311	B	11.2	0.123	141	868223	B	11.71	0.142
97	88203002	B	11.2	0.124	142	874373	B	11.71	0.142
98	897132	B	11.2	0.124	143	89864002	B	11.71	0.142
99	897137	B	11.3	0.125	144	8912055	B	11.74	0.143
100	8913049	B	11.3	0.126	145	91550	B	11.74	0.143
101	91789	B	11.3	0.126	146	8711561	B	11.75	0.144
102	901549	B	11.3	0.126	147	91858	B	11.75	0.144
103	903011	B	11.3	0.126	148	857343	B	11.76	0.144
104	868871	B	11.3	0.127	361	892189	M	11.76	0.144
105	8910748	B	11.3	0.127	149	904357	B	11.8	0.145
106	883852	B	11.3	0.127	362	869691	M	11.8	0.145
107	859465	B	11.3	0.128	150	874662	B	11.81	0.145
108	88199202	B	11.3	0.128	151	8810528	B	11.84	0.146
109	8911230	B	11.3	0.129	363	853612	M	11.84	0.146
110	864018	B	11.3	0.129	152	891703	B	11.85	0.146
111	916221	B	11.3	0.129	153	9111596	B	11.87	0.147
112	905502	B	11.4	0.130	154	8811523	B	11.89	0.147
113	89143601	B	11.4	0.130	155	8911164	B	11.89	0.147
114	865137	B	11.4	0.132	156	905686	B	11.89	0.147
115	873843	B	11.4	0.132	157	869476	B	11.9	0.148
360	84348301	M	11.4	0.132	158	864685	B	11.93	0.148
116	868682	B	11.4	0.133	159	904647	B	11.93	0.148
117	869218	B	11.4	0.133	160	857374	B	11.94	0.149
118	861103	B	11.5	0.133	161	8912909	B	11.94	0.149
119	89296	B	11.5	0.134	162	899147	B	11.95	0.149
120	898690	B	11.5	0.134	163	9110720	B	11.99	0.150
121	911685	B	11.5	0.135	164	8612080	B	12	0.150
122	88518501	B	11.5	0.135	165	9111843	B	12	0.150
123	925291	B	11.5	0.136	166	895299	B	12.03	0.151
124	85759902	B	11.5	0.136	167	9113816	B	12.04	0.151
125	884689	B	11.5	0.136	168	857155	B	12.05	0.151
126	893988	B	11.5	0.137	169	90250	B	12.05	0.151
127	921385	B	11.5	0.137	170	871122	B	12.06	0.151
128	906539	B	11.6	0.138	171	898678	B	12.06	0.151
129	862485	B	11.6	0.139	172	918465	B	12.07	0.151
130	893061	B	11.6	0.139	173	903554	B	12.1	0.152
131	911320501	B	11.6	0.139	174	912193	B	12.16	0.153
132	906616	B	11.6	0.139	175	86211	B	12.18	0.153
133	911366	B	11.6	0.139	176	862965	B	12.18	0.153
134	9112366	B	11.6	0.140	177	894090	B	12.18	0.153
135	863031	B	11.6	0.140	178	866714	B	12.19	0.153

	ID	M = maligno, B = benigno	radio	mezcla		ID	M = maligno, B = benigno	radio	mezcla
179	89511501	B	12.2	0.153	221	9010598	B	12.76	0.155
180	9011495	B	12.2	0.154	222	875093	B	12.77	0.155
181	902975	B	12.2	0.154	223	924084	B	12.77	0.155
182	91544001	B	12.2	0.154	368	868202	M	12.77	0.155
183	877501	B	12.2	0.154	224	859487	B	12.78	0.155
184	8711003	B	12.3	0.154	225	90401602	B	12.8	0.155
185	91376701	B	12.3	0.154	226	873586	B	12.81	0.155
186	905501	B	12.3	0.154	227	911408	B	12.83	0.154
187	9113846	B	12.3	0.154	369	881861	M	12.83	0.154
188	874839	B	12.3	0.155	228	905190	B	12.85	0.154
189	897374	B	12.3	0.155	229	8610908	B	12.86	0.154
190	8610175	B	12.3	0.155	230	894855	B	12.86	0.154
191	87139402	B	12.3	0.155	231	8910506	B	12.87	0.154
192	8712064	B	12.3	0.155	232	908916	B	12.87	0.154
193	904969	B	12.3	0.155	233	917062	B	12.88	0.154
194	91813702	B	12.3	0.155	234	924632	B	12.88	0.154
364	875263	M	12.3	0.155	235	884626	B	12.89	0.154
195	861597	B	12.4	0.155	236	8912284	B	12.89	0.154
196	863270	B	12.4	0.155	237	8913	B	12.89	0.154
197	90251	B	12.4	0.155	238	869224	B	12.9	0.153
198	907915	B	12.4	0.155	239	875878	B	12.91	0.153
199	883539	B	12.4	0.156	240	9047	B	12.94	0.153
200	894335	B	12.4	0.156	241	891670	B	12.95	0.153
201	913063	B	12.5	0.156	242	904689	B	12.96	0.153
365	843786	M	12.5	0.156	243	896864	B	12.98	0.152
202	892604	B	12.5	0.156	244	897604	B	12.99	0.152
203	914101	B	12.5	0.156	245	871001501	B	13	0.152
366	84501001	M	12.5	0.156	246	9112594	B	13	0.152
204	87930	B	12.5	0.156	370	844981	M	13	0.152
205	914580	B	12.5	0.156	247	873357	B	13.01	0.152
206	894089	B	12.5	0.156	248	854941	B	13.03	0.151
207	901034302	B	12.5	0.156	249	857810	B	13.05	0.151
208	91505	B	12.5	0.156	250	893548	B	13.05	0.151
209	9010258	B	12.6	0.156	251	9010259	B	13.05	0.151
210	8912521	B	12.6	0.156	252	8510653	B	13.08	0.150
211	88147202	B	12.6	0.156	253	8810158	B	13.11	0.150
212	911202	B	12.6	0.156	371	863030	M	13.11	0.150
213	86408	B	12.6	0.156	254	9113455	B	13.14	0.149
214	914366	B	12.7	0.156	255	8711002	B	13.15	0.149
215	89511502	B	12.7	0.156	256	914102	B	13.16	0.149
367	85922302	M	12.7	0.156	257	911320502	B	13.17	0.149
216	906024	B	12.7	0.155	372	85638502	M	13.17	0.149
217	891716	B	12.7	0.155	373	85715	M	13.17	0.149
218	90769602	B	12.7	0.155	258	884448	B	13.2	0.148
219	917080	B	12.8	0.155	259	89344	B	13.2	0.148
220	89382602	B	12.8	0.155	260	9112367	B	13.21	0.148

	ID	M = maligno, B = benigno	radio	mezcla		ID	M = maligno, B = benigno	radio	mezcla
261	922296	B	13.2	0.148	384	859983	M	13.8	0.129
262	9113239	B	13.2	0.147	385	87880	M	13.81	0.128
263	861853	B	13.3	0.146	386	91504	M	13.82	0.128
264	8813129	B	13.3	0.146	297	86561	B	13.85	0.127
265	902727	B	13.3	0.146	298	8911834	B	13.85	0.127
374	856106	M	13.3	0.146	299	909231	B	13.85	0.127
266	901041	B	13.3	0.145	387	8810987	M	13.86	0.127
267	8611161	B	13.3	0.144	300	901028	B	13.87	0.126
268	865468	B	13.4	0.143	301	922297	B	13.87	0.126
269	9112085	B	13.4	0.143	302	902976	B	13.88	0.126
270	9010877	B	13.4	0.142	303	911673	B	13.9	0.125
375	915691	M	13.4	0.142	304	91227	B	13.9	0.125
376	87163	M	13.4	0.142	305	918192	B	13.94	0.123
377	855167	M	13.4	0.141	388	886452	M	13.96	0.123
271	862009	B	13.5	0.141	389	908489	M	13.98	0.122
272	9013579	B	13.5	0.141	306	909410	B	14.02	0.120
273	91813701	B	13.5	0.141	307	88249602	B	14.03	0.120
274	912519	B	13.5	0.140	308	909220	B	14.04	0.119
378	855138	M	13.5	0.140	309	925292	B	14.05	0.119
275	857156	B	13.5	0.140	310	903811	B	14.06	0.119
276	893526	B	13.5	0.139	311	89524	B	14.11	0.117
277	90401601	B	13.5	0.139	390	8810955	M	14.19	0.113
278	8610629	B	13.5	0.138	312	911654	B	14.2	0.113
279	8510426	B	13.5	0.138	313	883270	B	14.22	0.112
280	8812818	B	13.6	0.137	391	874858	M	14.22	0.112
281	8910499	B	13.6	0.136	392	854268	M	14.25	0.111
282	8911800	B	13.6	0.136	393	858986	M	14.25	0.111
379	862717	M	13.6	0.136	314	86409	B	14.26	0.110
380	866083	M	13.6	0.136	315	892214	B	14.26	0.110
283	922576	B	13.6	0.135	394	91979701	M	14.27	0.110
284	857373	B	13.6	0.135	316	8910721	B	14.29	0.109
285	88350402	B	13.6	0.135	317	88143502	B	14.34	0.107
286	8812816	B	13.7	0.134	318	9113156	B	14.4	0.104
287	9013594	B	13.7	0.134	319	89143602	B	14.41	0.104
288	906878	B	13.7	0.134	320	89813	B	14.42	0.104
289	91903902	B	13.7	0.133	395	862548	M	14.42	0.104
290	9013005	B	13.7	0.133	321	86973702	B	14.44	0.103
291	912558	B	13.7	0.133	396	877500	M	14.45	0.102
292	917896	B	13.7	0.132	322	921386	B	14.47	0.102
381	84458202	M	13.7	0.132	397	86135501	M	14.48	0.101
382	84667401	M	13.7	0.131	323	865432	B	14.5	0.100
293	869931	B	13.7	0.131	324	911150	B	14.53	0.099
294	88411702	B	13.8	0.131	325	911201	B	14.53	0.099
295	891923	B	13.8	0.130	398	84799002	M	14.54	0.099
383	875938	M	13.8	0.130	326	915940	B	14.58	0.097
296	90944601	B	13.8	0.130	399	852763	M	14.58	0.097

	ID	M = maligno, B = benigno	radio	mezcla		ID	M = maligno, B = benigno	radio	mezcla
327	925277	B	14.6	0.097	423	915460	M	15.46	0.066
400	90291	M	14.6	0.096	424	903507	M	15.49	0.065
328	89382601	B	14.6	0.096	425	90602302	M	15.5	0.064
329	861648	B	14.6	0.095	426	88725602	M	15.53	0.064
330	861598	B	14.6	0.095	427	889403	M	15.61	0.061
331	913102	B	14.6	0.095	428	887181	M	15.66	0.060
401	848406	M	14.7	0.093	429	873701	M	15.7	0.059
332	906564	B	14.7	0.093	350	867387	B	15.71	0.059
402	857793	M	14.7	0.092	351	912600	B	15.73	0.058
333	921644	B	14.7	0.091	430	8812877	M	15.75	0.058
334	89869	B	14.8	0.090	431	899667	M	15.75	0.058
403	859283	M	14.8	0.089	432	84610002	M	15.78	0.057
335	9110944	B	14.8	0.088	433	864877	M	15.78	0.057
336	915664	B	14.8	0.088	434	846381	M	15.85	0.056
337	908469	B	14.9	0.086	435	845636	M	16.02	0.052
404	87556202	M	14.9	0.086	436	896839	M	16.03	0.052
338	914333	B	14.9	0.086	437	8610404	M	16.07	0.051
405	864729	M	14.9	0.086	438	869104	M	16.11	0.051
406	907914	M	14.9	0.084	439	84862001	M	16.13	0.050
339	911384	B	14.9	0.084	440	854039	M	16.13	0.050
340	86973701	B	15	0.083	352	905189	B	16.14	0.050
407	868826	M	15	0.083	441	86730502	M	16.16	0.050
341	8915	B	15	0.082	353	901303	B	16.17	0.050
342	8712291	B	15	0.082	442	8912280	M	16.24	0.049
343	8712853	B	15	0.082	443	911916	M	16.25	0.049
344	905557	B	15	0.081	444	895633	M	16.26	0.048
408	855133	M	15	0.081	445	8953902	M	16.27	0.048
345	88147102	B	15	0.081	354	915452	B	16.3	0.048
346	914862	B	15	0.079	446	9012315	M	16.35	0.047
409	91594602	M	15.1	0.079	447	87281702	M	16.46	0.046
410	862028	M	15.1	0.079	355	9010872	B	16.5	0.045
411	9010018	M	15.1	0.078	448	926954	M	16.6	0.044
347	866458	B	15.1	0.077	449	852552	M	16.65	0.044
412	857438	M	15.1	0.077	450	913535	M	16.69	0.044
413	879523	M	15.1	0.076	451	854253	M	16.74	0.043
414	905680	M	15.1	0.076	452	8712729	M	16.78	0.043
348	9012568	B	15.2	0.074	356	8711216	B	16.84	0.042
415	925622	M	15.2	0.073	453	879830	M	17.01	0.041
349	8810436	B	15.3	0.071	454	85382601	M	17.02	0.041
416	873885	M	15.3	0.071	455	881972	M	17.05	0.041
417	852973	M	15.3	0.070	456	89742801	M	17.06	0.041
418	886776	M	15.3	0.070	457	911296201	M	17.08	0.041
419	8511133	M	15.3	0.069	458	852631	M	17.14	0.041
420	861799	M	15.4	0.068	459	889719	M	17.19	0.041
421	8670	M	15.5	0.066	460	859717	M	17.2	0.041
422	87164	M	15.5	0.066	461	909445	M	17.27	0.040

	ID	M = maligno, B = benigno	radio	mezcla		ID	M = maligno, B = benigno	radio	mezcla
462	888570	M	17.3	0.040	507	854002	M	19.27	0.037
463	8860702	M	17.3	0.040	508	884180	M	19.4	0.037
464	888264	M	17.4	0.040	509	89122	M	19.4	0.037
465	881094802	M	17.4	0.040	510	913505	M	19.44	0.037
466	88330202	M	17.5	0.040	511	886226	M	19.45	0.037
467	8712766	M	17.5	0.040	512	88119002	M	19.53	0.036
468	877989	M	17.5	0.040	513	892438	M	19.53	0.036
469	853201	M	17.6	0.040	514	88649001	M	19.55	0.036
470	9113538	M	17.6	0.040	515	90312	M	19.55	0.036
471	8711202	M	17.7	0.040	516	871201	M	19.59	0.036
472	9110732	M	17.8	0.040	517	9111805	M	19.59	0.036
357	91376702	B	17.9	0.040	518	898431	M	19.68	0.036
473	90439701	M	17.9	0.040	519	84300903	M	19.69	0.036
474	8911163	M	17.9	0.039	520	885429	M	19.73	0.035
475	865128	M	18	0.039	521	866674	M	19.79	0.035
476	842302	M	18	0.039	522	8811842	M	19.8	0.035
477	90524101	M	18	0.039	523	849014	M	19.81	0.035
478	914062	M	18	0.039	524	916838	M	19.89	0.035
479	9110127	M	18	0.039	525	89263202	M	20.09	0.033
480	8610637	M	18.1	0.039	526	926682	M	20.13	0.033
481	877159	M	18.1	0.039	527	894618	M	20.16	0.033
482	857392	M	18.2	0.039	528	8610862	M	20.18	0.033
483	894326	M	18.2	0.039	529	908194	M	20.18	0.033
484	844359	M	18.3	0.039	530	9011494	M	20.2	0.033
485	874217	M	18.3	0.039	531	86208	M	20.26	0.032
486	916799	M	18.3	0.039	532	84358402	M	20.29	0.032
487	867739	M	18.5	0.039	533	887549	M	20.31	0.032
488	8612399	M	18.5	0.039	534	895100	M	20.34	0.032
489	914769	M	18.5	0.039	535	901088	M	20.44	0.031
490	852781	M	18.6	0.039	536	91930402	M	20.47	0.031
491	853401	M	18.6	0.039	537	883263	M	20.48	0.031
492	857010	M	18.7	0.039	538	88206102	M	20.51	0.030
493	86517	M	18.7	0.039	539	919555	M	20.55	0.030
494	897630	M	18.8	0.039	540	842517	M	20.57	0.030
495	8911670	M	18.8	0.039	541	881046502	M	20.58	0.030
496	908445	M	18.8	0.039	542	91485	M	20.59	0.030
497	859575	M	18.9	0.038	543	927241	M	20.6	0.030
498	866203	M	19	0.038	544	901288	M	20.64	0.030
499	86135502	M	19	0.038	545	88995002	M	20.73	0.029
500	855625	M	19.1	0.038	546	926125	M	20.92	0.027
501	8611792	M	19.1	0.038	547	884948	M	20.94	0.027
502	8912049	M	19.2	0.038	548	873593	M	21.09	0.026
503	846226	M	19.2	0.038	549	911157302	M	21.1	0.026
504	877486	M	19.2	0.038	550	851509	M	21.16	0.025
505	8711803	M	19.2	0.038	551	9012795	M	21.37	0.024
506	857637	M	19.2	0.038	552	926424	M	21.56	0.022

ID	M = maligno, B = benigno	radio	mezcla	
553	903516	M	21.6	0.022
554	9011971	M	21.7	0.021
555	8910988	M	21.8	0.021
556	9012000	M	22	0.018
557	86355	M	22.3	0.016
558	915143	M	23.1	0.011
559	88299702	M	23.2	0.010
560	8712289	M	23.3	0.010
561	878796	M	23.3	0.010
562	89812	M	23.5	0.008
563	865423	M	24.3	0.005
564	91762702	M	24.6	0.004
565	8611555	M	25.2	0.002
566	899987	M	25.7	0.002
567	873592	M	27.2	0.000
568	911296202	M	27.4	0.000
569	8810703	M	28.1	0.000

Se especifica cada valor clasificado de radio de menor a mayor para cada dato.

CONCLUSIÓN

Se puede decir que existen algoritmos que permiten obtener la graficación de ciertos problemas abordados, que son de gran utilidad ya que permiten el estudio adecuado de un gran funcionamiento de distintas aplicaciones que se utilizan en el campo médico y en otros campos. Los algoritmos que fueron desarrollados tomaron forma para la adecuada programación de este mismo entre estos fueron el de EM que permitió una adecuada obtención de valores para que se aplicasen con el algoritmo de mezcla de mixturas Gaussianas y que fue de gran desarrollo definiendo variables y matrices y fórmulas como el logaritmo de una función, además el algoritmo de la función de la densidad normal, todo esto llevo a una representación del problema planteado por el doctor William H.

Las funciones de matemáticas conllevan a un gran papel en lo académico y en el desarrollo, porque a través de estas se pueden hacer funciones de gran utilidad que permiten el estudio adecuado de sus aplicaciones. El algoritmo de esperanza maximización permitió obtener valores aproximados para calcular ya que solamente a través de sus aproximaciones es posible arrojar resultados que después pueden ser graficados y que además son valores muy puntuales.

Los lenguajes de programación conllevan a una amplia ayuda para ver en gráfico los algoritmos previamente estudiados y que son de amplia utilidad ya que soportan características únicas que forman parte de un grupo de funciones; previamente elaboradas por gente que ha dedicado su tiempo en elaborarla, sus aplicaciones fueron de gran utilidad ya que solamente se puede visualizar de esta forma los datos arrojados de un cierto algoritmo.

El propósito fue programar diferentes algoritmos para la obtención de un problema aplicado en un propósito real. Lo que se quiso mostrar es que por medio de la programación es posible abordar la graficación de datos que son recabados por científicos y colaboradores.

Las imágenes obtenidas son muestra de lo escrito previamente en esta conclusión, lo que se muestra en el desarrollo.

Apéndice A

Información del trabajo del Dr. William H.

Enfoques sobre diagnósticos y pronósticos del cáncer mamario, fue el resultado de la colaboración de la universidad de Wisconsin-Madison entre Olvi L. Mangasarian del departamento de ciencias de la computación y el Dr., William H. Wolberg del departamento de cirugía y oncología humana.

Diagnostico

Este trabajo nace por el deseo del Dr. Wolberg para puntualizar el diagnostico de las masas mamarias basado solamente en la aspiración con aguja fina (FNA). Identifico nueve características visualmente juzgadas de un ejemplo de FNA que considero relevante para diagnóstico. En colaboración con el Prof. Mangasarian y dos de sus estudiantes graduados, Rudy Setiono y Kristin Bennett, un clasificador fue construido usando el método de superficie múltiple (MSM) de una separación de patrón en las nueve características que completamente diagnosticaron 97% de casos nuevos. La información resultante es bien conocida en la Wisconsin Breast Cáncer Data. El análisis de imagen empezó in 1990 con la ayuda de Nick Street para el grupo de búsqueda. La meta era diagnosticar por medio de una imagen digital de una pequeña sección de diapositivas de FNA. Los resultados fueron consolidados en un software llamado Xcyt, que es comúnmente usado por el Dr. Wolberg en su práctica clínica. El proceso de diagnóstico es realizado así:

- Un FNA es tomado de la masa del pecho. El material es montado en un microscopio y manchado para resaltar el núcleo celular.

- El usuario separa el núcleo individual usando Xcyt. Usando el puntero del ratón, el usuario dibuja la aproximación de los límites de cada núcleo. Usando un enfoque de visión por computadora conocido como "snakes", estas aproximaciones convergen a exactos límites del núcleo. Este proceso interactivo toma entre 2 y 5 minutos por diapositiva.
- Una vez todo (o más) que el núcleo ha sido separado, el programa computa valores para cada una de las 10 características de cada núcleo, tamaño de medición, forma y textura. El significado, error estándar y valores extremos de estas características son computadas, resultando en un total de 30 características nucleares por cada ejemplo.
- Basado en 569 casos, un clasificador lineal fue construido para diferenciar ejemplos de casos benignos de malignos. Este clasificador consiste en un plano simple en el espacio de 3 características: valor extremo del área, valor extremo de lisura, y el valor principal de textura. Proyectando todos los casos sobre lo normal del plano de separación, aproximando las densidades de probabilidad de puntos de benigno y maligno son construidos. Hasta la fecha, el sistema ha correctamente diagnosticado 176 consecutivos pacientes (119 benignos, 57 malignos). En solo 8 de estos casos Xcyt retorno un diagnostico sospechoso (esto es, una probabilidad estimada de malignidad entre 0.3 y 0.7).

Apéndice B

Imágenes del trabado del Dr. William H.

Tomando en cuenta el problema de cáncer de seno se pueden observar algunas imágenes tomadas por el Dr. William H.

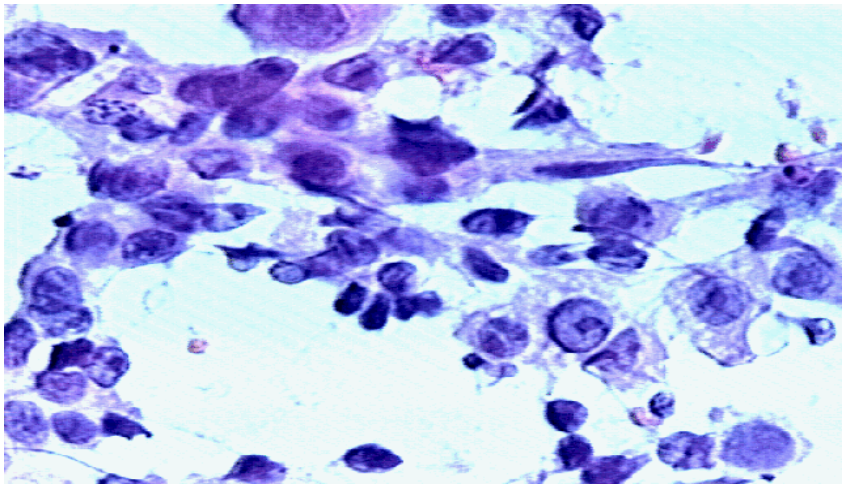

Imagen 1 Núcleos celulares

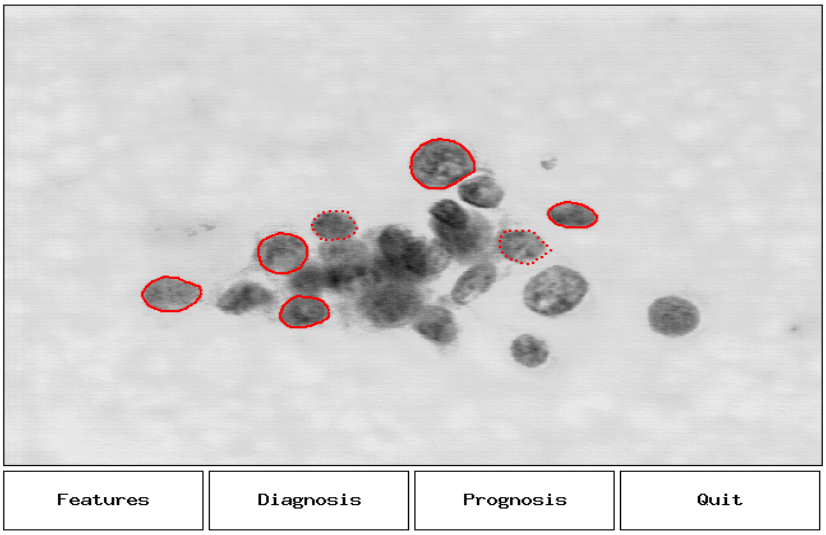

Imagen 2 Usando el software Xcyt para seleccionar células

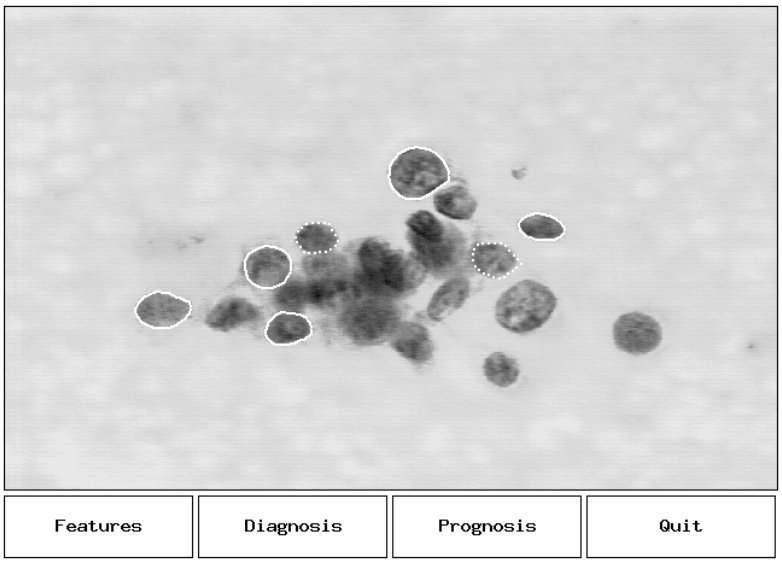

Imagen 3 Usando el software Xcyt para seleccionar células con otro color

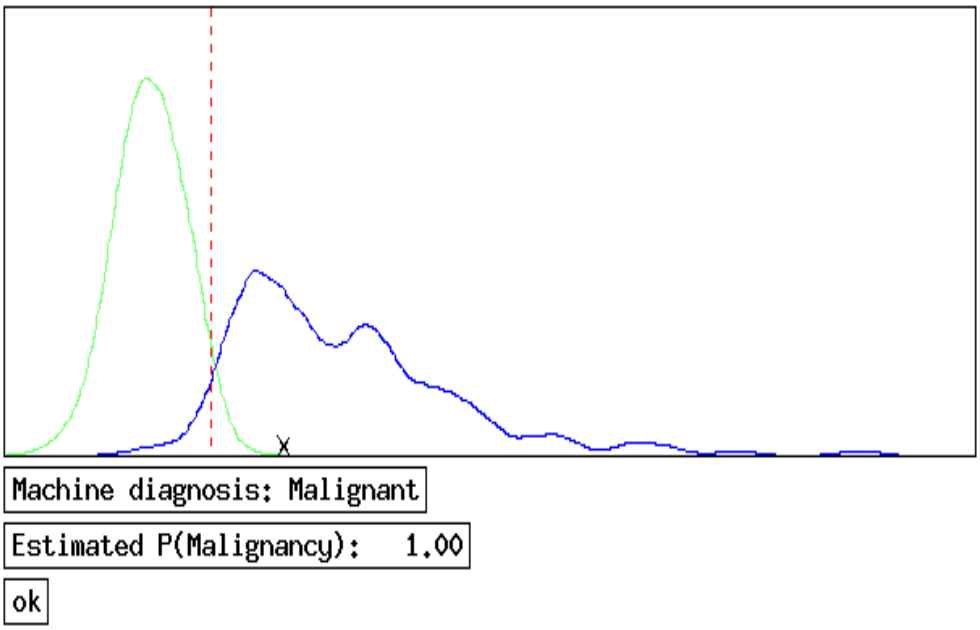

Imagen 4 Graficas de las células benignas y malignas

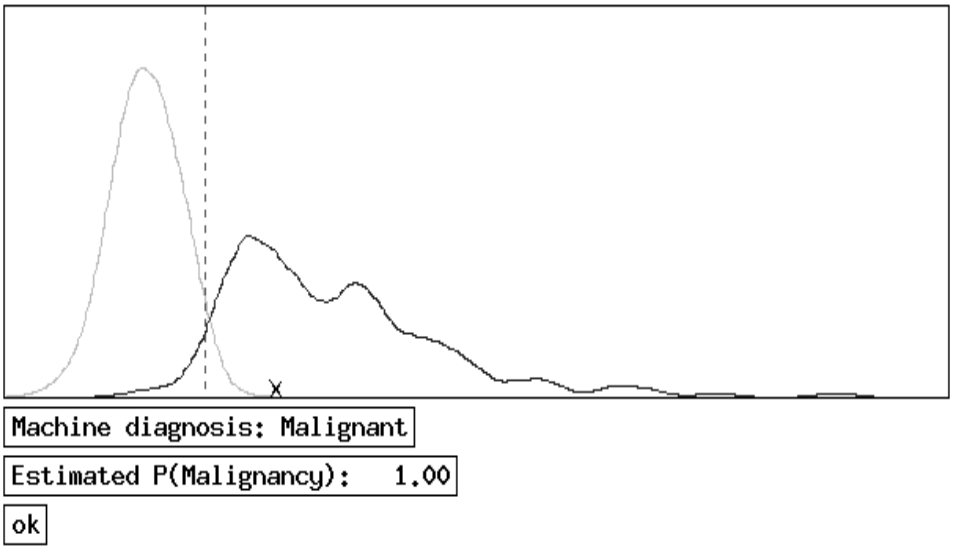

Imagen 5 Graficas de las células benignas y malignas en color gris

Apéndice C

Herramientas de programación

JavaScript

JavaScript (abreviado comúnmente JS) es un lenguaje de programación interpretado, dialecto del estándar ECMAScript. Se define como orientado a objetos, basado en prototipos, imperativo, débilmente tipado y dinámico.

Se utiliza principalmente en su forma del lado del cliente (client-side), implementado como parte de un navegador web permitiendo mejoras en la interfaz de usuario y páginas web dinámicas, aunque existe una forma de JavaScript del lado del servidor (Server-side JavaScript o SSJS). Su uso en aplicaciones externas a la web, por ejemplo en documentos PDF, aplicaciones de escritorio (mayoritariamente widgets) es también significativo.

Desde el 2012, todos los navegadores modernos soportan completamente ECMAScript 5.1, una versión de JavaScript. Los navegadores más antiguos soportan por lo menos ECMAScript 3. La sexta edición se liberó en julio del 2015.

JavaScript se diseñó con una sintaxis similar a C, aunque adopta nombres y convenciones del lenguaje de programación Java. Sin embargo, Java y JavaScript tienen semánticas y propósitos diferentes.

Todos los navegadores modernos interpretan el código JavaScript integrado en las páginas web. Para interactuar con una página web se

provee al lenguaje JavaScript de una implementación del Document Object Model (DOM).

Tradicionalmente se venía utilizando en páginas web HTML para realizar operaciones y únicamente en el marco de la aplicación cliente, sin acceso a funciones del servidor. Actualmente es ampliamente utilizado para enviar y recibir información del servidor junto con ayuda de otras tecnologías como AJAX. JavaScript se interpreta en el agente de usuario al mismo tiempo que las sentencias van descargándose junto con el código HTML.

Desde el lanzamiento en junio de 1997 del estándar ECMAScript 1, han existido las versiones 2, 3 y 5, que es la más usada actualmente (la 4 se abandonó). En junio de 2015 se cerró y publicó la versión ECMAScript 6.

Nacimiento de JavaScript

JavaScript fue desarrollado originalmente por Brendan Eich de Netscape con el nombre de Mocha, el cual fue renombrado posteriormente a LiveScript, para finalmente quedar como JavaScript. El cambio de nombre coincidió aproximadamente con el momento en que Netscape agregó compatibilidad con la tecnología Java en su navegador web Netscape Navigator en la versión 2.002 en diciembre de 1995. La denominación produjo confusión, dando la impresión de que el lenguaje es una prolongación de Java, y se ha caracterizado por muchos como una estrategia de mercadotecnia de Netscape para obtener prestigio e innovar en el ámbito de los nuevos lenguajes de programación web.

«JAVASCRIPT» es una marca registrada de Oracle Corporation. Es usada con licencia por los productos creados por Netscape Communications y entidades actuales como la Fundación Mozilla.

Características

Imperativo y estructurado

JavaScript es compatible con gran parte de la estructura de programación de C (por ejemplo, sentencias if, bucles for, sentencias switch, etc.). Con una salvedad, en parte: en C, el ámbito de las variables alcanza al bloque en el cual fueron definidas; sin embargo JavaScript no es compatible con esto, puesto que el ámbito de las variables es el de la función en la cual fueron declaradas.

Tipado dinámico

Como en la mayoría de lenguajes de scripting, el tipo está asociado al valor, no a la variable. Por ejemplo, una variable x en un momento dado puede estar ligada a un número y más adelante, religada a una cadena.

Entorno de ejecución

JavaScript normalmente depende del entorno en el que se ejecute (por ejemplo, en un navegador web) para ofrecer objetos y métodos por los que los scripts pueden interactuar con el "mundo exterior". De hecho, depende del entorno para ser capaz de proporcionar la capacidad de incluir o importar scripts (por ejemplo, en HTML por medio del tag <script>). (Esto no es una característica del lenguaje, pero es común en la mayoría de las implementaciones de JavaScript.)

Ejemplos sencillos

Las variables en JavaScript se definen usando la palabra clave var:

```
var x; // define la variable x, aunque no tiene ningún valor asignado por defecto
var y = 2; // define la variable 'y' y le asigna el valor 2 a ella
```

A considerar los comentarios en el ejemplo de arriba, los cuales van precedidos con 2 barras diagonales.

Una función recursiva:

```
function factorial(n) {
    if (n === 0) {
        return 1;
    }
    return n * factorial(n - 1);
}
```

Uso en páginas web

El uso más común de JavaScript es escribir funciones embebidas o incluidas en páginas HTML y que interactúan con el Document Object Model (DOM o Modelo de Objetos del Documento) de la página. Algunos ejemplos sencillos de este uso son:

- Animación de los elementos de página, hacerlos desaparecer, cambiar su tamaño, moverlos, etc.
- Contenido interactivo, por ejemplo, juegos y reproducción de audio y vídeo.

- Validación de los valores de entrada de un formulario web para asegurarse de que son aceptables antes de ser enviado al servidor.
- Transmisión de información sobre los hábitos de lectura de los usuarios y las actividades de navegación a varios sitios web. Las páginas Web con frecuencia lo hacen para hacer análisis web, seguimiento de anuncios, la personalización o para otros fines.

Imagen 1 Eslogan de JavaScript.

5.1 Bibliografía:

[1] Abella, R., & Medina, J. E. (2014). Segmentación lineal de texto por tópicos. Serie Gris, CENATAV.

[2] Blei, D., & Jordan, M. (2002). Modeling Annotated Data. Technical Report UCB//CSD-02-1202, U.C. Berkeley Computer Science Division.

[3] Dickey, J., Jiang, J.-M., & Kadane, J. (1987). Bayesian Methods for Censored Categorical Data. Journal of the American Statistical Association, 82, 773-781.

[4] Gebali, F. (2011). Algorithms and parallel computing. Hoboken, N.J.: Wiley.

[5] Griffiths, T., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In Proceedings of the 24th Annual Conference of Cognitive Science Society.

[6] Griffiths, T., & Steyvers, M. (2003). Prediction and Semantic Association. 11-18.

[7] Griffiths, T., & Steyvers, M. (s.f.). Finding scientific topics. Proceedings of the National Academy of Science, 101, 5228-5235.

[8] Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2005). Integrating Topics and Syntax. In Advances in Neural Information Processing Systems, 17.

[9] Heinrich, G. (2008). Parameter estimation for text analysis. Technical Report Fraunhofer.

[10] Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence.

[11] Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999. F. Jelinek. Statistical Methods for Speech Recognition. MIT Press, Cambridge, MA, 1997. .

[12] Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent. Machine Learning, 42, 177-196.

[13] Jordan, M. (1999). Learning in Graphical Models. Cambridge: MIT Press.

[14] Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. (1999). An Introduction to Variational Methods for Graphical Models (Vol. 37). Machine Learning.

[15] Landauer, T., & Dutnais, S. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological Review, 104, 2011-240.

[16] Landauer, T., Foltz, P., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.

[17] Minka, T. (2000). Estimating a Dirichlet distribution. Technical report, M.I.T.

[18] Minka, T., & Lafferty, J. (2002). Expectation-Propagation for the Generative Aspect Model. In Uncertainty in Artificial Intelligence (UAI).

[19] Pacheco, P. S. (2011). An introduction to parallel programming. Amsterdam: Morgan Kaufmann.

[20] Petersen, W., & Arbenz, P. (2004). Introduction to parallel computing. Oxford: Oxford University Press.

[21] Real Academia Española. (2017). Real Academia Española. Obtenido de www.rae.es

[22] Reese, S., Boleda, G., Cuadros, M., Padró, L., & Rigau, G. (2010). Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. La Valleta, Malta: In Proceedings of 7th Language Resources and Evaluation Conference (LREC'10).

[23] Steyvers, M., & Griffiths, T. (s.f.). Probabilistic Topic Models . In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds).

[24] Universitat Politècnica de Catalunya. Research Group on Natural Language Processing; Gemma Boleda; Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada (IULA). (2012). GrAF version of Spanish portions of Wikipedia Corpus.

[25]Wang, Y. (2008). Distributed Gibbs Sampling of Latent Topic Models: The Gritty Details.

5.2 Fuentes de internet:

[1] Barney, B. (2015). *OpenMP*. Obtenido de https://computing.llnl.gov/tutorials/openMP/

[2] Barney, B. (2017). *Introduction to Parallel Computing.* Obtenido de https://computing.llnl.gov/tutorials/parallel_comp/

[3] Barney, B. (2017). *Message Passing Interface (MPI).* Obtenido de https://computing.llnl.gov/tutorials/mpi/

[4] Bezanson, J. (2017). *Julia Benchmarks.* Obtenido de https://julialang.org/benchmarks/

[5] LNS. (2017). *Laboratorio Nacional de Supercomputo*. Obtenido de http://www.lns.buap.mx

[6] MPI Forum. (2017). *MPI Documents.* Obtenido de http://mpi-forum.org/docs/

[7] Software Intel. (2017). *OpenMP* Pragmas and Clauses Summary.* Obtenido de https://software.intel.com/es-es/node/522685

[8] Top500.org. (2017). *TOP500 Supercomputer Sites*. Obtenido de https://www.top500.org/

[9] Wikipedia. (2019). *Google Chart.* Obtenido de https://en.wikipedia.org/wiki/Google_Charts

[10] Wikipedia. (2019). *JavaScript.* Obtenido de https://es.wikipedia.org/wiki/JavaScript

[11] UNIVERSITY OF WISCONSIN-MADISON. (2019). *Machine Learnining for Cancer Diagnosis and Prognosis.* Obtenido de http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html