Aprendizaje automático

Departamento de Ingeniería en Informática
ITBA

Trabajo Práctico 4

Objetivo: Aprender a agrupar datos mediante análisis de clusters

Determinación de clases. Clustering.

1. Los datos de la tabla que fueron extraídos del anuario económico de *La Caixa* del año 2006 corresponden a características sociodemográficas de las 52 provincias españolas.
   Los datos son:
   - Nombre de la provincia
   - Población
   - Tasa varones/mujeres
   - Tasa extranjeros/españoles
   - Extensión de la provincia (en km^2)
   - Paro (porcentaje de desocupados)
   - Número de teléfonos fijos registrados
   - Número de vehículos de motor matriculados
   - Número de oficinas bancarias
   - Precio medio del m^2 de vivienda

Almería	612315	1,0599	0,1792	8742	4,2	200091	379954	593	1511,3
Cádiz	1180817	0,9909	0,0240	7441	7,8	340497	604460	645	1713
Córdoba	784376	0,9628	0,0180	13771	6	260552	416945	618	1405,9
Granada	860898	0,9737	0,0440	12647	4,7	325645	516625	749	1338,7
Huelva	483792	0,9938	0,0405	10128	5,2	144113	245393	382	1554,8
Jaén	660284	0,9874	0,0175	13489	4,6	215307	344782	595	993
Málaga	1453409	0,9752	0,1428	7306	4,6	641125	882852	1089	2184,1
Sevilla	1813908	0,9646	0,0211	14036	5,9	616896	977502	1298	1601,5
Huesca	215864	1,0365	0,0754	15626	2,4	97349	144535	353	1532,3
Teruel	141091	1,0502	0,0743	14797	2,6	64583	89203	228	966,8
Zaragoza	912072	0,9739	0,0857	17274	3	419196	469822	1087	1995,4
Asturias	1076635	0,9215	0,0255	10604	5	420156	564790	875	1612,1
Balears (Illes)	983131	1,0044	0,1890	4992	2,8	487704	787713	1134	2192,6
Palmas (Las)	1011928	1,0241	0,1205	4066	7,5	418992	636376	582	1739,1
Santa Cruz de Tenerife	956352	0,9919	0,1346	3381	6,3	385469	633599	586	1638,7
Cantabria	562309	0,9558	0,0379	5253	4,1	200844	322186	461	1907,4
Ávila	167032	1,0102	0,0377	8050	3,9	77504	100652	195	1294
Burgos	361021	1,0123	0,0505	14000	3,2	158022	209915	519	1704

León	495902	0,9512	0,0294	15542	4,7	200215	290875	461	1137,4
Palencia	173471	0,9734	0,0207	8052	4,4	69567	99042	218	1193,5
Salamanca	352414	0,9528	0,0325	12349	5,2	147118	190341	388	1466,5
Segovia	155517	1,0142	0,0831	6849	2,6	73087	98766	183	1479,6
Soria	92773	1,0133	0,0614	10303	2,5	43782	58612	153	1353,2
Valladolid	514674	0,9645	0,0364	8110	4,6	210686	279743	532	1588,9
Zamora	198045	0,9784	0,0191	10561	4,9	79052	112143	241	1135,8
Albacete	384640	1,0032	0,0564	14918	5,9	112127	219192	313	1260,3
Ciudad Real	500060	0,9823	0,0472	19813	5,4	183450	269700	431	1000,2
Cuenca	207974	1,0180	0,0736	17140	3,6	85086	129315	241	987,8
Guadalajara	203737	1,0444	0,0929	12169	3	98910	123124	221	1898,6
Toledo	598256	1,0164	0,0727	15370	4,2	245628	357285	582	1396
Barcelona	5226354	0,9686	0,1222	7728	3,8	2740093	3168920	5364	2551,7
Girona	664506	1,0241	0,1779	5910	2,7	353064	497492	773	1979,6
Lleida	399439	1,0398	0,1274	12150	2,4	182055	272189	544	1294,3
Tarragona	704907	1,0255	0,1354	6303	3,1	335505	460133	748	1726,4
Alicante	1732389	0,9996	0,2280	5817	4,4	685447	1079688	1466	1701,5
Castellón	543432	1,0074	0,1374	6632	3	205675	348533	580	1649,6
Valencia	2416628	0,9759	0,0876	12033	4,4	857686	1498361	2235	1458,5
Badajoz	671299	0,9813	0,0189	21766	7,6	235451	369348	685	946,9
Cáceres	412580	1,0044	0,0322	19868	5,6	159380	239193	471	939,5
Coruña (La)	1126707	0,9243	0,0216	7950	5,8	391210	614346	923	1397,3
Lugo	357625	0,9412	0,0198	9856	4,5	135360	215720	340	1060,4
Ourense	339555	0,9234	0,0346	7273	6	132482	227883	377	1060,7
Pontevedra	938311	0,9338	0,0299	4495	6,4	312027	570210	747	1561,6
Madrid	5964143	0,9387	0,1506	8022	3,7	2864928	3767288	4989	2902,7
Murcia	1335792	1,0278	0,1409	11313	3,1	405430	807577	1151	1441,7
Navarra	593472	0,9990	0,0918	9801	3,6	243969	370704	684	1670,8
Álava	299957	0,9953	0,0532	2932	3,5	124038	170521	288	2548
Guipúzcoa	688708	0,9672	0,0323	1909	3,3	279922	385767	530	2788
Vizcaya	1136181	0,9459	0,0329	2217	4	424039	582188	905	2772,7
Rioja (La)	301084	1,0159	0,1151	5028	3,1	121815	163592	437	1595
Ceuta	75276	1,0366	0,0420	19	8	24849	49811	22	1648,3
Melilla	65488	1,0359	0,0462	13	9	23060	42805	20	1320,5

(a) Aplicar el método de las componentes principales para encontrar el número apropiado de clusters para agrupar a las provincias españolas de acuerdo a las características sociodemográficas.
(b) Aplicando el algoritmo de k-medias agrupar a las provincias.
¿La agrupación se puede asociar a la ubicación geográfica?

2. Clasificar los “datos de los lirios Fisher” (ver ej. 5 tp2) en 3 grupos.
   (a) Usando el algoritmo de k-medias.
   (b) Usando el algoritmo de clusters jerárquicos aglomerativos

(c) Calcular las matrices de confusión para las clasificaciones obtenidas en los ítems (a) y (b) y comparar los resultados con el obtenido en el ejercicio 5 del Tp2.

3. La Universidad de Leeds realizó un estudio sobre los dialectos ingleses en el cual se eligieron 331 individuos de zonas rurales. Los individuos de distintas localidades usualmente usan distintas palabras para el mismo ítem.
En la tabla se reportan los datos de 25 localidades en la región central este (East Midland) de Inglaterra; se da como medida de similitud entre pares de localidades el porcentaje de 60 de los ítems cubiertos en el estudio para los cuales se usan las mismas palabras en los dos lugares.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
B	71																							
C	58	57																						
D	49	45	48																					
E	63	63	47	59																				
F	64	66	50	53	71																			
G	71	75	52	53	71	68																		
H	52	56	36	34	60	58	69																	
I	46	50	57	33	42	43	43	44																
J	61	49	52	40	58	61	56	61	52															
K	57	60	56	35	53	48	55	48	63	59														
L	39	46	45	30	42	40	47	44	50	53	60													
M	42	50	53	28	41	36	43	39	48	47	58	48												
N	32	34	47	20	27	29	31	23	44	39	43	39	63											
O	32	39	50	19	25	25	36	37	43	41	48	49	64	63										
P	23	27	42	14	20	22	24	28	36	27	38	35	54	62	68									
Q	41	47	56	25	38	42	46	38	48	54	54	48	72	57	61	59								
R	39	42	48	24	37	34	36	42	43	49	60	56	59	51	56	47	54							
S	32	36	43	22	22	24	34	29	47	34	45	47	38	46	51	42	44	53						
T	27	36	38	19	22	20	25	25	40	25	40	40	45	49	54	49	42	44	63					
U	28	37	37	20	25	25	31	33	42	29	41	37	46	48	49	47	43	44	58	59				
V	26	26	30	20	21	28	28	28	41	33	39	55	34	33	40	33	38	40	58	54	47			
W	30	33	32	16	25	26	33	32	41	37	37	46	47	46	49	39	46	58	42	44	42	50		
X	36	49	45	26	29	31	41	32	47	32	52	46	57	49	56	49	54	53	63	68	73	51	51	
Y	31	44	40	23	29	32	32	31	47	33	43	45	45	47	53	43	46	53	60	61	62	55	54	72

Las localidades A, B y C corresponden al condado de Nottinghamshire, D a L al de Lincolnshire, M a P al de Leicestershire, Q al de Rutland, R a U al de Northamptonshire, V al de Huntingdonshire, W al de Cambridgeshire, X al de Buckinghamshire e Y al de Bedfordshire
Se quiere ver cómo agrupar dialectos similares.
Aplicar el algoritmo de clusters jerárquicos aglomerativos con distintas funciones de unión y comparar los resultados.

4. La tabla muestra el número de veces que 15 congresistas de New Jersey votaron en forma diferente en 19 proyectos ambientales.

	Nombre	Partido	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Hunt	R	0	8	15	15	10	9	7	15	16	14	15	16	7	11	13
2	Sandman	R	8	0	17	12	13	13	12	16	17	15	16	17	13	12	16
3	Howard	D	15	17	0	9	16	12	15	5	5	6	5	4	11	10	7
4	Thompson	D	15	12	9	0	14	12	13	10	8	8	8	6	15	10	7
5	Freylinghuyseb	R	10	13	16	14	0	8	9	13	14	12	12	12	10	11	11
6	Forsythe	R	9	13	12	12	8	0	7	12	11	10	9	10	6	6	10
7	Widnall	R	7	12	15	13	9	7	0	17	16	15	14	15	10	11	13
8	Roe	D	15	16	5	10	13	12	17	0	4	5	5	3	12	7	6
9	Heltoski	D	16	17	5	8	14	11	16	4	0	3	2	1	13	7	5
10	Rodino	D	14	15	6	8	12	10	15	5	3	0	1	2	11	4	6
11	Minish	D	15	16	5	8	12	9	14	5	2	1	0	1	12	5	5
12	Rinaldo	R	16	17	4	6	12	10	15	3	1	2	1	0	12	6	4
13	Maraziti	R	7	13	11	15	10	6	10	12	13	11	12	12	0	9	13
14	Daniels	D	11	12	10	10	11	6	11	7	7	4	5	6	9	0	9
15	Patten	D	13	16	7	7	11	10	13	6	5	6	5	4	13	9	0

No se han registrado las abstenciones, pero hay dos congresistas que se abstienen más frecuentemente que otros y son Sandman (9 abstenciones) y Thompson (6 abstenciones) ¿Pueden detectarse las afiliaciones partidarias? (R = republicano, D = demócrata).
Aplicar el algoritmo de clusters jerárquicos aglomerativos con distintas funciones de unión y comparar los resultados.