Trabajo de Clustering

Fecha de entrega: 25 de marzo de 2020

28 de enero de 2020

1. Introducción

El objetivo de este trabajo es que los alumnos realicen un estudio exploratorio de los datos usando análisis cluster. Para ello habrá que seleccionar las variables que sean relevantes, definir cuántos grupos significativos se pueden encontrar en los datos, analizar la importancia de cada grupo, analizar la importancia de las variables en su definición, descripción estadística de los grupos y finalmente sintetizar la anterior información en una descripción semántica de los patrones de clientes.

1.1. Datos: *Wholesale customers Data Set*

Los datos han sido tomados del UCI repositorio, Center for Machine Learning and Intelligent Systems, University of California. Esta base de datos representa las ventas anuales de un distribuidor mayoristas a sus 440 clientes minoristas. El objetivo que se persigue es descubrir los patrones de clientes existentes, con el objetivo de poder orientar adecuadamente las estrategias de marketing a realizar. Una visualización de los 10 primeros clientes se muestran en la figura 1.

	A	B	C	D	E	F	G	H
1	2	3	12669	9656	7561	214	2674	1338
2	2	3	7057	9810	9568	1762	3293	1776
3	2	3	6353	8808	7684	2405	3516	7844
4	1	3	13265	1196	4221	6404	507	1788
5	2	3	22615	5410	7198	3915	1777	5185
6	2	3	9413	8259	5126	666	1795	1451
7	2	3	12126	3199	6975	480	3140	545
8	2	3	7579	4956	9426	1669	3321	2566
9	1	3	5963	3648	6192	425	1716	750
10	2	3	6006	11093	18881	1159	7425	2098

Figura 1: Datos: *Wholesale customers Data Set*

.

Esta base de datos tiene $N = 440$ registros (clientes) y un total de $Q = 8$ atributos, definidos por

A: Tipo cliente (1, Distribuidor minorista, 2 Hotel,cafeterías,...)
B: Región (1, zona Lisboa, 2 zona Oporto, 3 Otros)
C: Gasto anual en productos frescos.
D: Gasto anual en productos lácteos.
E: Gasto anual en productos de ultramarinos.
F: Gasto anual en productos congelados.
G: Gasto anual en detergentes y productos de papelería.
H: Gasto anual en productos delicatessen.

1.2. Métodos

Todos los alumnos realizarán el mismo trabajo pero con diferentes métodos. Los métodos que vamos a considerar son:

- K-means
- Fuzzy c-means
- Algoritmo jerárquico acumulativo (HAC)
- Basado en densidad (DBSCAN)
- Algoritmo Expectation-Maximization
- Mapas Autoorganizados
- Spectral Clustering

El listado con los métodos asignados a cada alumno se publicará en Campus Virtual.

1.3. ENTREGABLES

- Código fuente utilizado para la solución del problema (bien comentado). Se entregará mediante el enlace al repositorio github (o bitbucket) correspondiente.
- Memoria explicativa del trabajo utilizando la plantilla de ACM con una limitación de 4 páginas. En la memoria tienen que explicitarse los enlaces a los componentes anteriores. https://www.acm.org/publications/proceedings-template

HITO 1: Determinación de clusters

El objetivo de este hito es descubrir qué patrones de clientes existe en el conjunto de datos. Para ello se realizará un análisis cluster para obtener un conjunto de grupos de clientes con unas características similares dentro de cada grupo.

COSAS QUE EL ALUMNO TIENE QUE HACER EN EL HITO 1:

1. Análisis de que variables deben intervenir en el estudio.
2. Determinación de outliers. Hay que identificar los mejores clientes respecto al volumen de ventas. Estos clientes deben ser identificados para dispensarles un trato diferenciado.
3. Core del negocio. El *Principio de Pareto* o *Regla del 80/20* (ó 20/80), establece que, de forma general y para un amplio número de fenómenos, aproximadamente el $80\,\%$ de las consecuencias proviene del $20\,\%$ de las causas. En este contexto, ¿se cumple?, ¿el $20\,\%$ de los clientes generan el $80\,\%$ de los ingresos? Este colectivo, el que genera el $80\,\%$ de los ingresos representa el nucleo del negocio de la empresa y conviene ser identificado para concentrar los esfuerzos a ellos.
4. Realización de un análisis cluster sobre el conjunto de clientes que se considere oportuno. Determinar el número óptimo de grupos a efectuar.
5. Estudiar la parametrización óptima del algoritmo asignado.

HITO 2 : Semántica de los clusters

La descripción semántica de los grupo facilita la comprensión del fenómeno y la comunicación de resultados. Este es el objetivo del hito. Una vez realizado el Hito 1 y obtenido los grupos hay que calcular el representante de cada grupo y analizar cuales son las características esenciales.

EL ALUMNO TIENE QUE HACER EN EL HITO 2:

1. Cálculo de los representantes de los grupos.
2. Determinación de la importancia de cada grupo para la empresa.
3. Hay que estudiar estadísticamente cuales son las variables que son significativamente diferentes entre los grupos. Por ejemplo dos grupos pueden presentar valores iguales en todas las variables excepto en la variable F: *Gasto anual en productos congelados* en la que toman valores diferentes, siendo el factor diferenciador entre grupos.Para ello se introduce en la sección 2 los test estadísticos no paramétricos que son una herramienta de gran utilidad para esta tarea.
4. Una vez analizada las diferencias estadísticamente significativas entre grupos y conociendo sus representantes, hay que realizar una descripción semántica de los grupos obtenidos.

2. Herramientas estadísticas

2.1. Introducción

Los tests de contraste de hipótesis son una herramienta estadística que nos permite dilucidar si ciertos hechos observados son productos del azar o no. Por ejemplo, nos podemos plantear que si dos grupos de clientes obtenidos tras realizar el análisis cluster poseen la misma media en una determinada variable (por ejemplo la anterior F: *Gasto anual en productos congelados*) o por el contrario cada uno de ellos posee una media diferente. Lo que queremos testar es que si las diferencias observadas en las ventas medias de ese producto en los diferentes grupos (muestras) son debidas al azar o por el contrario esta diferencia es sistemáticamente no nula.

El anterior ejemplo se formaliza matemáticamente planteándose dos hipótesis excluyentes: la nula (denominada H_0) y la hipótesis alternativa (denominada H_1), que en este ejemplo tomarían la expresión:

$\mathbf{H_0}$: Las ventas media del producto tipo x en los grupos cumple $\mu_i = \mu$ para todo grupo i

$\mathbf{H_1}$: Existen un par de grupos u y v en los que su media $\mu_u \neq \mu_v$.

Los test de hipótesis asumen que los datos siguen una determinada distribución de probabilidad. Si la distribución de probabilidad se basa en la normal se llaman test paramétricos y si no se asume ninguna distribución específica se denomina test no paramétricos. El test de hipótesis paramétrico más conocido es el *test t* que permite determinar si una variable posee la misma media en dos muestras (dos grupos en el análisis cluster). El *análisis de la varianza de un factor* (ANOVA) extiende el test t al caso de tener varias muestras (varios grupos en el análisis cluster) y permite contestar si la media de una variable (aleatoria normal) tiene la misma media en todas las muestras. La aplicación de estos métodos requiere la comprobación de que los datos cumplen las hipótesis de: i) se distribuyen según una distribución normal, ii) poseen la misma varianza dentro de cada muestra y iii) son observaciones independientes. Los métodos no paramétricos permiten aplicar estos procedimientos sin necesidad de verificar ninguna hipótesis adicional. El inconveniente de emplear los métodos no paramétricos es que

requiere un número mayor de observaciones para detectar que existen diferencias entre las medias. Si el número de observaciones es mayor que 30 en cada submuestra ambos tipos de métodos obtienen las mismas conclusiones.

Los test de hipótesis se basan en el llamado *estadístico de contraste* (una función de los datos) que tienen una distribución muestral conocida cuando la hipótesis nula es verdadera. Esto permite calcular la probabilidad que este estadístico tome valores en un determinado intervalo. Asumiendo que la hipótesis H_0 es verdadera se puede calcular la llamada *región de confianza* a un nivel de confianza determinado $1-\alpha$ (α recibe el nombre de nivel de confianza). Esto significa que si la hipótesis H_0 fuese verdadera, para el $1-\alpha$ muestras que obtuviésemos el estadístico de contraste calculado en dicha muestra debería de estar dentro de la región de confianza calculada. La forma de proceder es la siguiente, se calcula el estadístico de contraste para nuestra muestra y si este valor cae dentro de la región de confianza se acepta H_0 y en caso contrario se aceptaría la hipótesis H_1.

La mayoría de los programas estadísticos calculan el denominado $p-$*value*. Este valor indica que la hipótesis nula será rechazada para cualquier nivel de confianza $\alpha \leq p$. Por ejemplo, si el $p-value = 0{,}023$ y estamos trabajando con un nivel de confianza de $\alpha = 5\,\% = 0{,}05$ se rechazaría la hipótesis a un nivel de confianza $\alpha = 5\,\%$.

2.2. Test no paramétricos

Los tests no paramétricos se clasifican de acuerdo a los siguientes tres criterios:

Número de muestras. Se distingue entre *dos muestras* y *más de dos muestras* (k muestras). Esto significa que podemos tener una medida (variable aleatoria) en dos grupos o en otro caso, esa misma medida se obtiene en tres o más grupos. Por ejemplo, podemos considerar las ventas de una determinada categoría de productos entre dos grupos determinados (aplicaríamos un test para dos muestras) o podemos considerarla entre los k grupos obtenidos en el análisis cluster (aplicaríamos un test para k muestras). En este caso si se aceptara la hipótesis nula H_0 ese tipo de producto no sería relevante para la conformación de grupos, esto es, las ventas de ese producto en todos los grupos es similar.

Tipo de muestras. Se distingue entre *muestras independientes* y *muestras relacionadas* o también llamadas *dependientes*. Esto significa que en el primer caso tenemos las medidas sobre diferentes sujetos experimentales mientras que en el segundo caso se obtienen todas las medidas sobre el mismo sujeto experimental. Por ejemplo, como evaluamos diferentes clientes tenemos muestras independientes. Si consideramos el problema de evaluar dos algoritmos sobre n problemas y midiéramos el tiempo de ejecución se trataría de un problema con muestras apareadas ya que se evalúan los dos (o k algoritmos si fuese el caso) sobre los mismos sujetos experimentales (los problemas).

Tipo de datos. Se distinguen entre datos que tiene una naturaleza cuantitativa o al menos medidos en una escala ordinal (primero, segundo, tercero, ...) frente a los datos cualitativos (hombre o mujer, blanco, negra o azul, etc.)

Atendiendo a los anteriores criterios la tabla 1 recoge el nombre del procedimiento a emplear en cada situación para variables cuantitativas u ordinales.

Cuadro 1: Test no paramétricos para datos cuantitativos u ordinales

Tipo \ número	2 **muestras**	k **muestras**
muestras independientes	U de Mann-Whitney	test Kruskal-Wallis
muestras relacionadas o dependientes	test rangos con signos de Wilconxon	test de Friedman

2.3. Aplicaciones

Vamos a discutir dos aplicaciones de los test no paramétricos. La primera nos ayuda a establecer **la semántica de los grupos** (Hito 2 del trabajo de clustering) y la segunda la **comparación de métodos** (trabajo final).

Semántica de grupos. Si consideramos que dos grupos son diferentes debemos saber respecto a que variables se puede considerar que son estadísticamente diferentes. El test de Kruskal-Wallis se puede aplicar a este problema y testar que una variable cuantitativa determinada (por ejemplo la F) toma un valor medio diferente en cada grupo. Si se rechaza la hipótesis nula se tiene evidencia estadística que esta variable es importante para discernir entre grupos. Si no se rechaza la hipótesis nula significaría que esta variable no es significativa en la conformación de los grupos y por tanto no interviene en la caracterización de los mismos.

Si se rechaza la hipótesis nula cabría plantearse una nueva cuestión. ¿Cuales son los pares de grupos que difieren entre si respecto a dicha variable? En este caso aplicaríamos el test U Mann-Whitney para cada par posible de grupos.

Los resultados de todos los test son pistas que hay que combinar para formar una descripción semántica de los clusters.

Comparación de métodos. Una segunda aplicación es la comparación de métodos respecto a una cierta característica (variable cuantitativa). Por ejemplo nos podemos plantear si un conjunto de clasificadores tienen o no la misma precisión. En este caso, como todos se evalúan sobre el mismo conjunto de subconjuntos tests, se trata de muestras relacionadas. Podemos aplicar el test de Friedman para determinar si son todos iguales o no y el de Wilcoxon para determinar par a par cuales son estadísticamente diferentes.