Evaluación 3a - Minería de Datos.

Herrera, Sánchez.

2022-08-10

Contexto.

En el siguiente informe se detalla el proceso de obtención de clústeres sobre una base de datos de un banco alemán. Esta base contiene información de distintos ámbitos para 1.000 personas (incluyendo una evaluación de riesgo crediticio). Para la obtención de los clústeres se utiliza el algoritmo de K medias (o *K means* en inglés).

Elección de variables y transformaciones realizadas.

Para la generación de clusters se elegirán las siguientes 11 variables que engloban información bancaria, patrimonial, personal y laboral (ver Anexo 1 para detalle de las variables):

- *amount*
- *age*
- *persons*
- *property*
- *savings*
- *employed*
- *credits*
- *history*
- *status*
- *personal*
- *housing*

Posteriormente, algunas de estas variables fueron transformadas antes de ser utilizadas en el trabajo con K medias.

Para la transformación es relevante mencionar que, dentro de las variables escogidas, algunas son de tipo continuas (*amount, age*), una es binaria (*persons*), otras son ordinales (*property, savings, employed, credits*) y otras, categóricas (*history, status, personal, housing*). Para el caso de las 4 variables categóricas, cada una de estas fue transformada a $\boldsymbol{n}$ variables binarias, donde n representa la cantidad de valores distintos que puede tomar cada variable categórica.

Cabe mencionar que para cada una de las variables categóricas, se excluye el primer valor, el cual será tomado como valor de referencia. Por ejemplo, para el caso de la variable *housing*,

esta puede tomar 3 valores indicando la situación de uso de la vivienda de la persona: 1 gratuita, 2 arrendada y 3 propia. Así, se mantendrán $n-1$ binarias creadas (*housing_2* y *housing_3*), y cuando ambas tomen valor 0, entonces será el caso de una persona que tiene usa una vivienda de forma gratuita (o *housing* en nivel 1).

Luego, todas las variables resultantes son normalizadas y por último, las variables binarias creadas a partir de las categóricas son divididas por el n mencionado con el objetivo de no alterar los resultados de la métrica euclidiana en función de la cantidad de categorías.

Selección del valor k.

Recordamos que el método de *K means* consiste en asignar k observaciones de manera pseudo-aleatoria y asignarlas como centros iniciales para el algoritmo. Luego, las observaciones más cercanas (bajo norma euclidiana en este trabajo) a cada centro se asocian al cluster indicado por el centro. Iterativamente, se actualizan los centros de cada cluster por el centro geométrico de sus observaciones para eventualmente converger.

Como hemos notado, la elección del K es fundamental para una buena segmentación de nuestros datos. Por esto, en la siguiente sección de realiza un análisis para ayudarnos con la elección.

Criterios para la elección del valor K.

Con el objetivo escoger el valor ideal de k para el modelo, se utilizarán dos criterios:

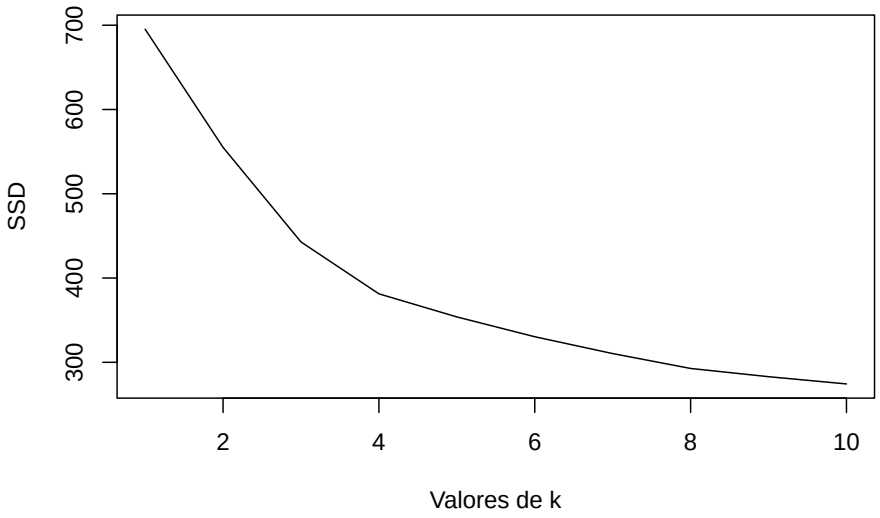
- Método del Codo. Análisis de la suma cuadrática de las distancias.
- Método de la Silueta. Cohesión y separación de cada clúster.

Codo.

Para este método utilizaremos como métrica de dispersión de los clusters, la suma cuadrática de las distancias de cada observación a su centroide. Tomando en consideración que esta suma **debe** disminuir a medida que aumenta el K, buscaremos un punto de inflexión K^* tal que la SSD se mantenga "estable" desde K^*+1.

$$\text{SSD}_i = \sum_{j \in \mathcal{C}_i} (x_j - \overline{x}_i)^2$$

Figura 1. La suma cuadrática de distancias decrece con el valor de k



Observamos como posibles candidatos $K = 3$ y $K = 4$. Dado que no encontramos un punto de inflexión claro, complementamos nuestra decisión con el método de la silueta.

Silueta.

El valor de la silueta cuantifica la diferencia entre la distancia promedio intra-cluster y la mínima distancia promedio inter-cluster, dividido entre el máximo de estas cantidades. Es decir,

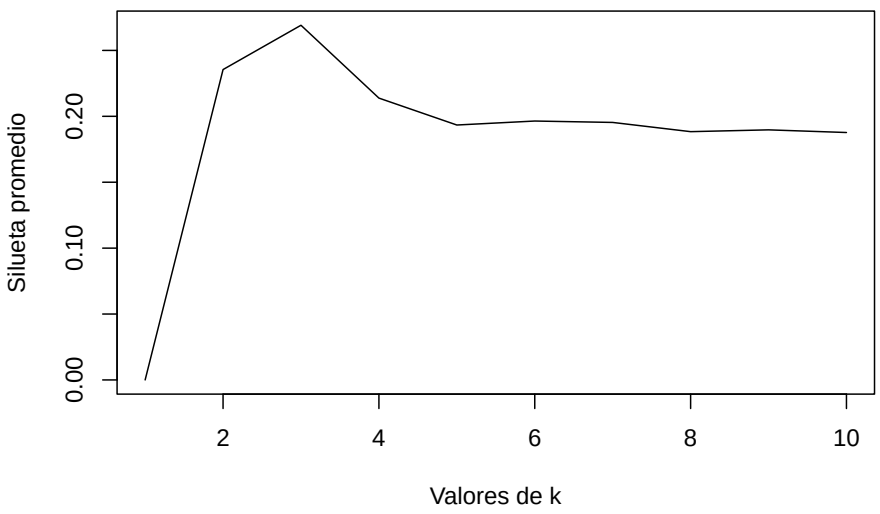
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \ , \ \forall i$$

donde,

- a(i) : distancia promedio intra-cluster, $a(i) = \dfrac{1}{|\mathcal{C}_i| - 1} \cdot \sum_{j \in \mathcal{C}_i} d(i, j)$.
- b(i) : menor distancia promedio inter-cluster, $\min_k \{ \dfrac{1}{|\mathcal{C}_i|} \cdot \sum_{j \in \mathcal{C}_k} d(i, j)$.

Con esta lógica, buscando idealmente valores más cercanos a 1 (por construcción) promediamos lo valores de la silueta para distintos K.

Figura 2. Silueta para distintos valores de k.



Notando que para K = 4 la distancia promedio intra-cluster, a(i), aumenta en general con respecto a K = 3. Así, para k = 3, las observaciones se emparejan mejor con su cluster.

Estimación y análisis de los resultados.

Luego de la elección del k=3, se estima la selección de clústeres. Con esto se obtienen tres clústeres que agrupan a las 1.000 observaciones analizadas. Los resultados se estudian a continuación.

Tamaño de los clústeres.

Con tres cluster y las variables escogidas, el cluster número 2 será el de mayor tamaño acumulando el 60% de las observaciones. El 40% restante es repartido entre el cluster número 3 (el segundo en tamaño) y el cluster 1 como el de menor tamaño con 155 observaciones.

Table 1: Tamaño de cada cluster.

Cluster	Tamaño
1	155
2	606
3	239

Cohesión y separación de los clústeres.

El valor total de las sumas cuadráticas de las distancias con tres clústeres es de **442.83**. A modo de referencia, la suma cuadrática de distancias inicial era de **695.2** (es decir la suma cuadrática de las distancias si se tuviera un sólo gran cluster). A nivel de cada cluster, las

sumas cuadráticas se presentan a continuación, siendo el cluster de mayor tamaño el que acumula el mayor valor.

Esta disminución de un **36** se puede observar en la Figura 1 incluída en el análisis de criterios para la elección de K.

Table 2: Suma cuadrática de las distancias, por cluster.

Cluster	Distancias
1	88.788
2	249.477
3	104.568

Caracterización de cada cluster.

Table 3: Centros de cada cluster.

	1	2	3
amount	3384.18	3220.89	3325.70
age	38.68	34.09	37.19
persons	1.00	0.00	0.00
property	2.39	2.36	2.34
savings	2.21	1.16	4.43
employed	3.66	3.20	3.68
credits	1.55	1.39	1.35
history_1	0.08	0.05	0.03
history_2	0.44	0.54	0.57
history_3	0.12	0.09	0.06
history_4	0.32	0.28	0.32
status_2	0.20	0.30	0.24
status_3	0.06	0.07	0.06
status_4	0.41	0.33	0.55
personal_2	0.09	0.36	0.32
personal_3	0.88	0.47	0.54
personal_4	0.01	0.11	0.09
housing_2	0.68	0.71	0.74
housing_3	0.19	0.09	0.09

- **Cluster 1 (Padres de familia)**: En su totalidad tienen tres o más cargas, y en un 90% de los casos son hombres casados, viudos o divorciados. Además, en este grupo registramos la mayor cantidad de personas que poseen casa propia, duplicando a los otros 2.
- **Cluster 2 (Sin ahorros)**: Clasificación de las personas más jóvenes de los registros. Tienen pocas o inexistentes cargas familiares. Es el grupo con mayor presencia de

viviendas gratuitas y con bajos ahorros (<= 100DM). Es el grupo más heterogeneo en cuanto al estado civil y sexo.

- **Cluster 3 (Con ahorros)**: Se distingue por los altos ahorros (tanto por lo indicado en las variables *savings* como *status*). Es el grupo con mayor presencia de arrendatarios.

Anexos.

Anexo 1.

Descripción de las variables consideradas para la estimación.

- **amount**: Variable continua. Monto del crédito.
- **age**: Variable continua. Edad de la persona.
- **persons**: Variable binaria. Identifica si la persona tiene 3 o más cargas, o no.
- **property**: Variable ordinal. Bien más valioso.
  - (1) nada, (2) auto, (3) seguro de vida, (4) propiedad.
- **savings**: Variable ordinal. Historial de ahorros.
  - (1) sin ahorros, (2) menos de 100 DM, (3) entre 100 y 500 DM, (4) entre 500 y 1.000 DM, (5) más de 1.000 DM.
- **employed**: Variable ordinal. Tiempo empleado.
  - (1) desempleado, (2) menos de 1 año, (3) entre 1 y 4 años, (4) entre 4 y 7 años, (5) más de 7 años.
- **credits**: Variable ordinal. Créditos vigentes.
  - (1) 1, (2) 2 o 3, (3) 4 o 5, (4) 6 o más.
- **history**: Variable categórica. Historial de créditos.
  - (0) retrasos en el pasado, (1) cuenta crítica o créditos en otros bancos, (2) sin créditos anteriores, (3) créditos vigentes al día, (4) todos los créditos pagados
- **status**: Variable categórica. Historial de cuenta corriente.
  - (1) sin cuenta, (2) cuenta con deuda, (3) ahorros hasta 200 DM, (4) ahorros desde 200 DM.
- **personal**: Variable categórica. Estado civil y sexo.
  - (1) hombre divorciado, (2) hombre soltero o mujer no soltera, (3) hombre casado o viudo, (4) mujer soltera.
- **housing**: Variable categórica. Tipo de vivienda.
  - (1) gratuita, (2) arrendada, (3) propia.