MBD - Estadística - Práctica II (Clustering)

Víctor Juez Cañellas

12/19/2020

Contents

Introducción

Actualmente, los teléfonos móviles almacenan una gran cantidad de datos en tiempo real sobre nuestras actividades rutinarias. Entre otros parámetros recogen información de nuestra movilidad gracias a sensores integrados dentro del mismo dispositivo. La compañía de teléfonos móviles SAMSAPPLE quiere clasificar la actividad de los usuarios de sus dispositivos en 6 niveles en base a la información recibida en tiempo real. Aunque es un problema que tienen bastante resuelto, han realizado un experimento con 21 voluntarios, los cuales reportaban en cada instante su estado real de actividad categorizado en 6 niveles: tumbado (laying); sentado (sitting); de pie (standing); caminando (walk); bajando escaleras (walkdown); o subiendo (walkup).

Objetivo

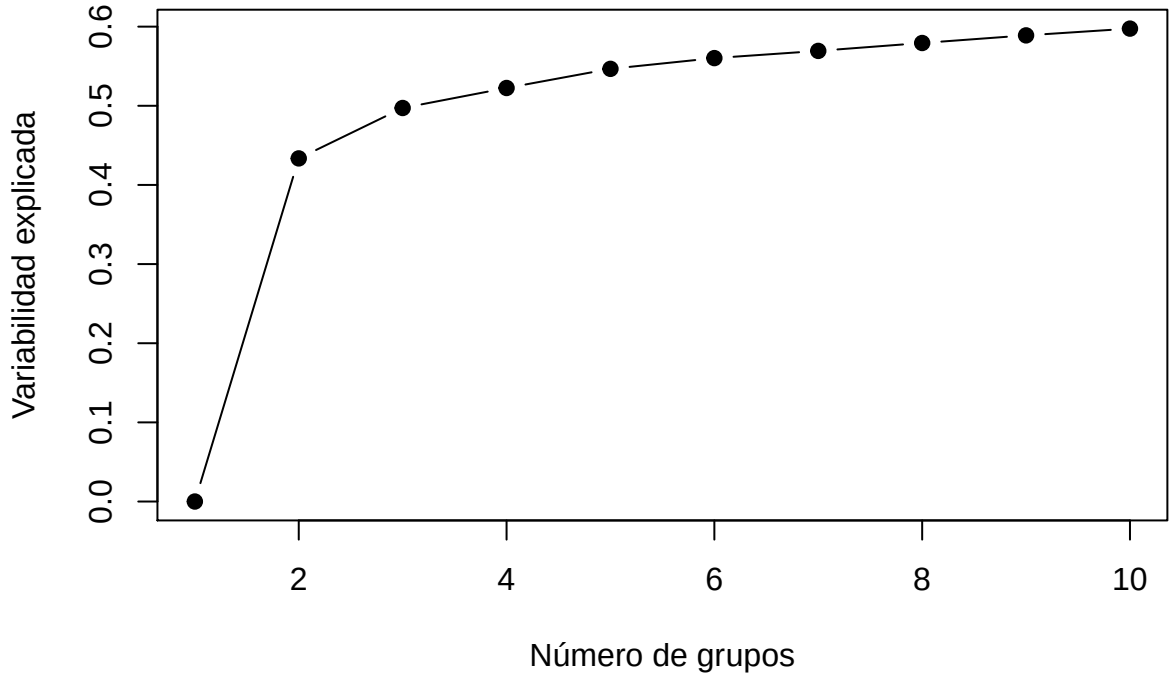
1. Agrupar las distintas respuestas de los sensores procurando mantener las máximas similitudes entre clústeres y la máxima heterogeneidad entre los mismos para discernir el número de tipos de actividades posibles (=número de grupos). Para llevarlo a cabo, se usará la técnica de clustering no supervisado del k-means. Sólo usar los datos de entrenamiento sin usar variable respuesta para hallar el número de grupos.
2. Construir un modelo predictivo que sea capaz de clasificar los individuos en cada instante en una de las 6 categorías de actividad. Para realizar la predicción sobre los datos de test, se deberán usar algunos (o preferentemente todos) los algoritmos vistos en las sesiones: KNN, Naive Bayes, Conditional Trees, Random Forests y SVM. Usar datos de entrenamiento para construir el modelo y hacer las predicciones sobre el conjunto de test

Parte I. Clustering no supervisado

Para identificar el número de clústers hemos utilizado el K-means. Sabemos que en el conjunto de datos final hay 6 categorías de actividad diferentes, así que, lo ideal sería identificar 6 clústers en el conjunto de datos. Para ello hemos utilizado por un lado el método de la regla del codo y por otro, la librería NbClust que utilizando diferentes índices determina cual es el número de clústeres ideal.

Regla del codo

Ejecutamos el k-means 10 veces utilizando de 1 a 10 clústeres respectivamente, y por cada resultado comparamos la variabilidad explicada.



Como podemos observar, según la regla del codo nos quedamos con dos clústers, ya que, es en este punto donde se forma el codo y, a partir de dos para arriba, el incremento de variabilidad explicada es muy pequeño. A continuación podemos observar la variabilidad explicada por cada número de clústers.

Num. Clusters	Variabilidad explicada
1	0.00
2	0.43
3	0.50
4	0.52
5	0.55
6	0.56
7	0.57
8	0.58
9	0.59
10	0.60

Aunque la variabilidad explicada sí que va aumentando a medida que se incrementan el número de clústers, vemos que el incremento más grande se produce cuando se pasa de un clúster a dos.

NbClust

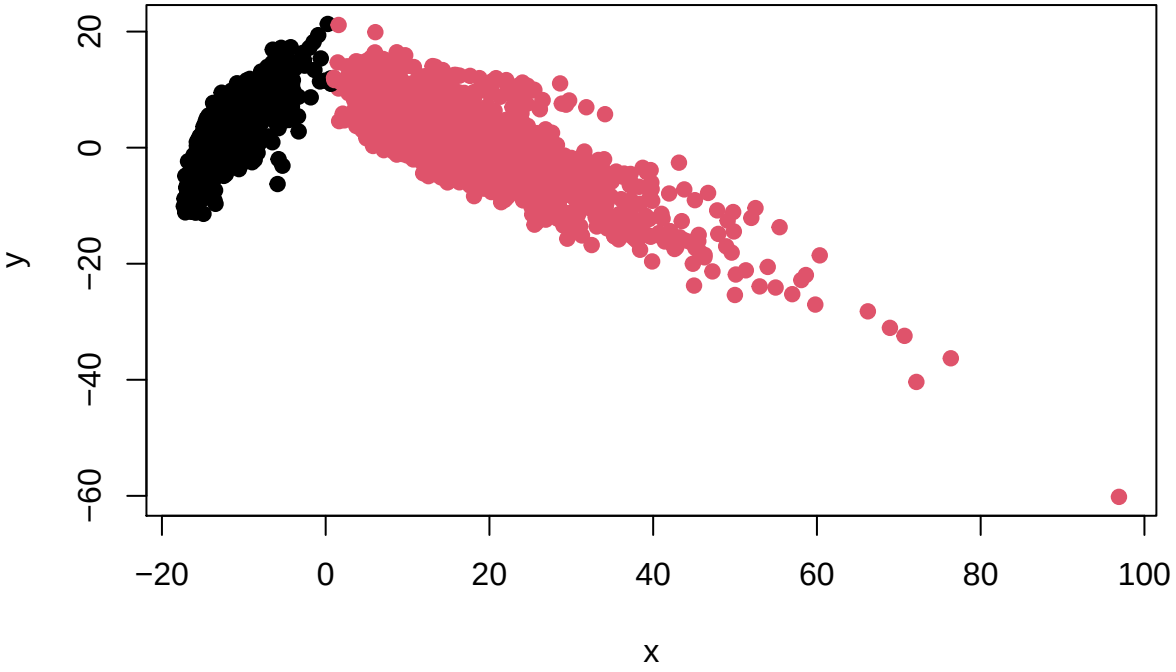
La librería NbClust utiliza 30 índices diferentes variando por cada uno todas las combinaciones de número de clústers, tipo de distancias y métodos de clusterización para determinar el número de clústers que más encaja al conjunto de datos. El resultado obtenido tras la ejecución es el siguiente:

- 14 índices proponen 2 como el mejor número de clústers.
- 2 índices proponen 3 como el mejor número de clústers.
- 5 índices proponen 4 como el mejor número de clústers.
- 1 índice propone 8 como el mejor número de clústers.

Por mayoría, el mejor número de clústers es 2.

Análisis del resultado

Para representar gráficamente los clústers, hemos hecho un análisis de componentes principales (PCA), con el que hemos extraído los dos componentes más significativos. Éstos van a formar el eje X e Y respectivamente, y sobre ellos mostramos el conjunto de datos marcados con colores distintos para representar el clúster al que pertenecen.



Vemos que la partición en dos clústers es coherente y que a simple vista no se identifican más que éstos dos. Esto nos indica que utilizando el conjunto de datos de muestra, no podemos discernir con presición entre las 6 actividades diferentes. Por éste mismo motivo, observamos la variabilidad explicada utilizando únicamente dos clústers es de 0.43, un valor bajo, y es que estamos dejando de identificar 4 categorías.

Parte II. Clustering supervisado