Tarea 2

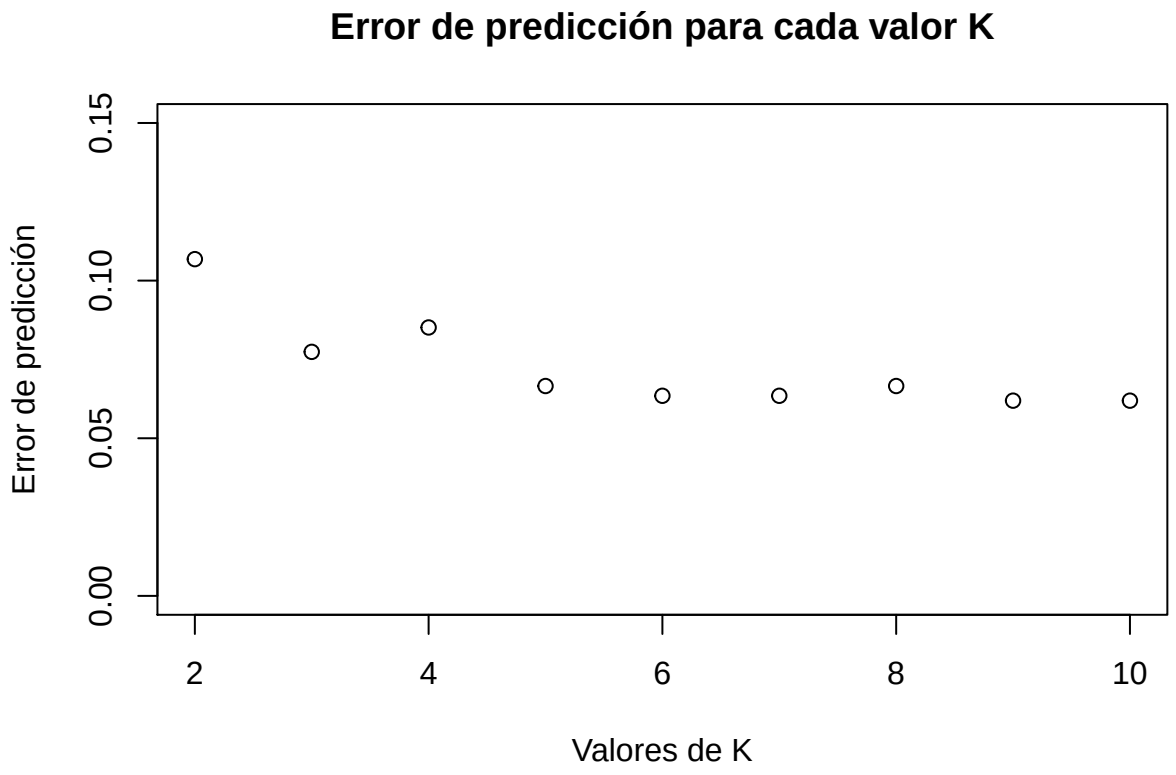
Francisco Alonso

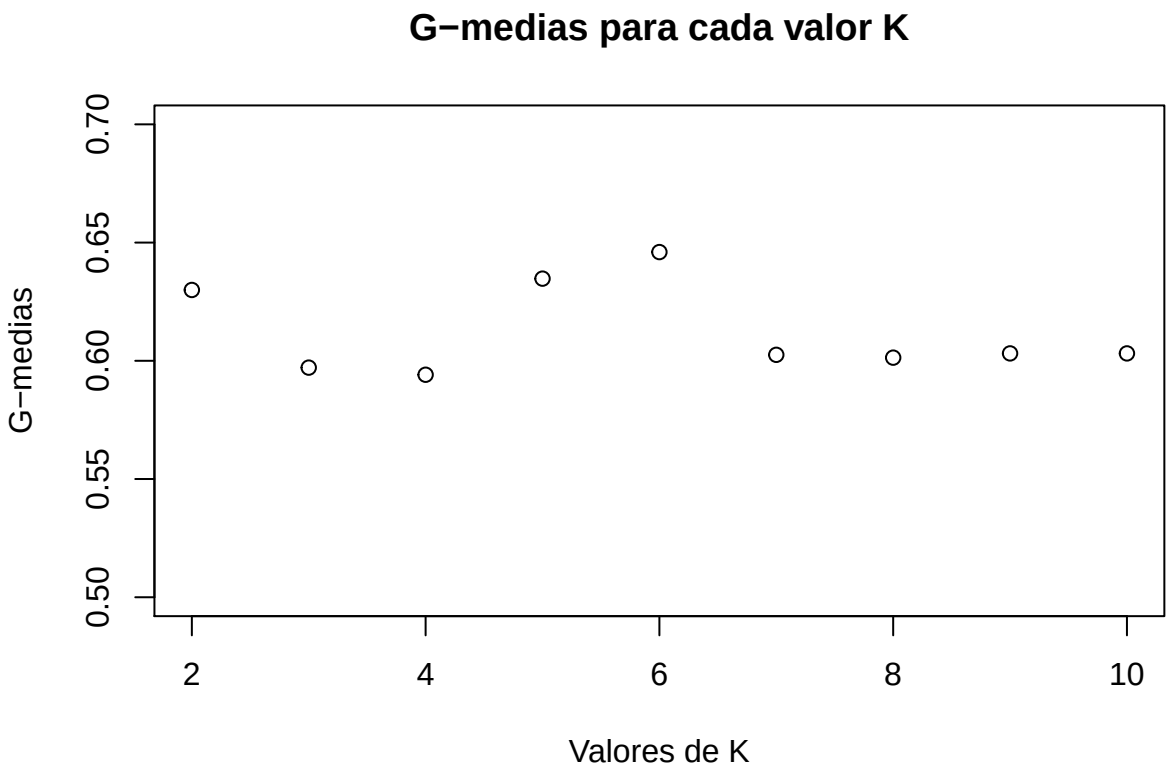
April 27, 2016

Ejercicio 1

Se aplica el proceso de minería de de datos sobre el conjunto de datos "SEISMIC-BUMPS", en este están clasificados eventos en minas según si en el siguente estado ocurrió o no un movimiento sísmico con mucha energía, este se clasifica como "estado peligroso" (1) o "estado no peligroso" (0) según un conjunto de 13 variables correspondientes a mediciones de energía en movimientos sísmicos y otras 5 que indican, en general, el nivel de peligrosidad de los movimientos ya ocurridos.

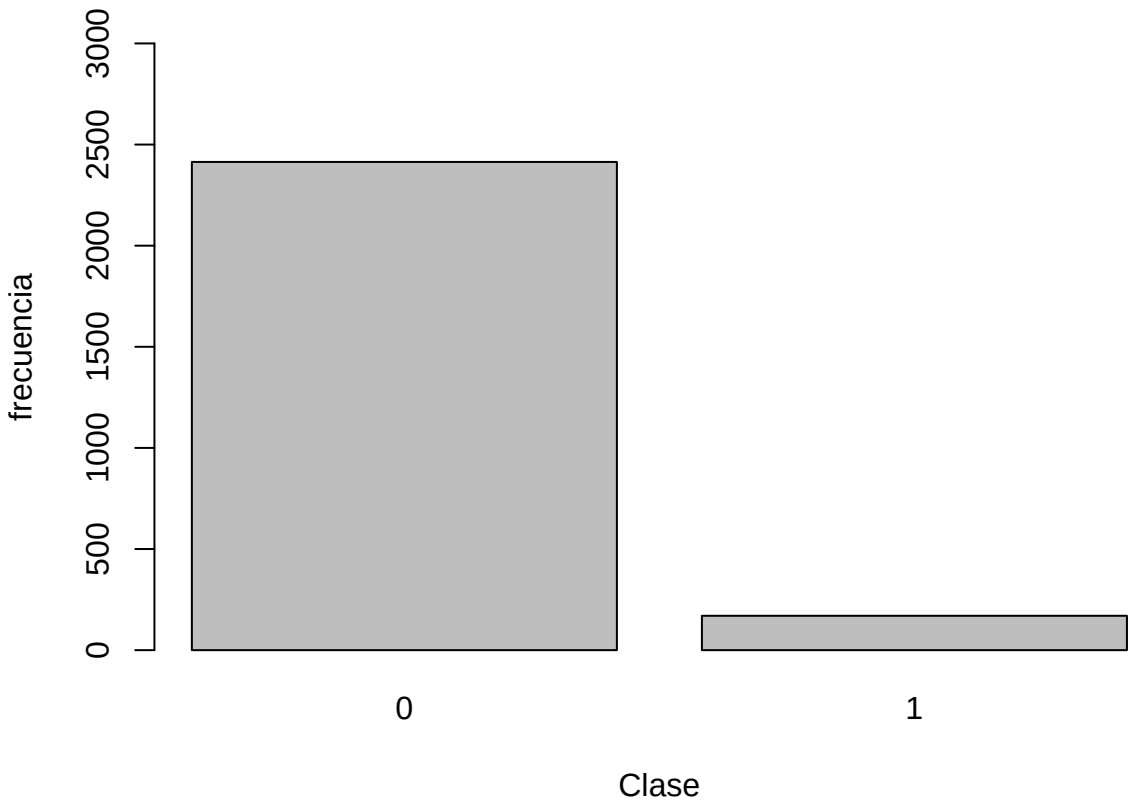
Se prepara la data para la creación del modelo, es necesario que las variables nominales sean convertidas a valores numéricos, luego se divide el conjunto de datos en un conjunto de entrenamiento y uno de prueba para construir nueve modelos usando el algoritmo k-vecinos más cercanos (k nearest neighborgs - knn) variando el número de grupos de dos a diez. Por último se toman medidas del error de predicción y g-medias para comparar cada modelo construido.





Se observa que el error de predicción se reduce al usar k = 9, sin embargo dado que esta métrica es susceptible al sesgo entre clases y observamos que las clases no están bien representadas en el conjunsto de datos, nos referimos al valor de g-medias para determinar que el valor de k apropiado es cuatro.

```
##    0    1
## 2414  170
```

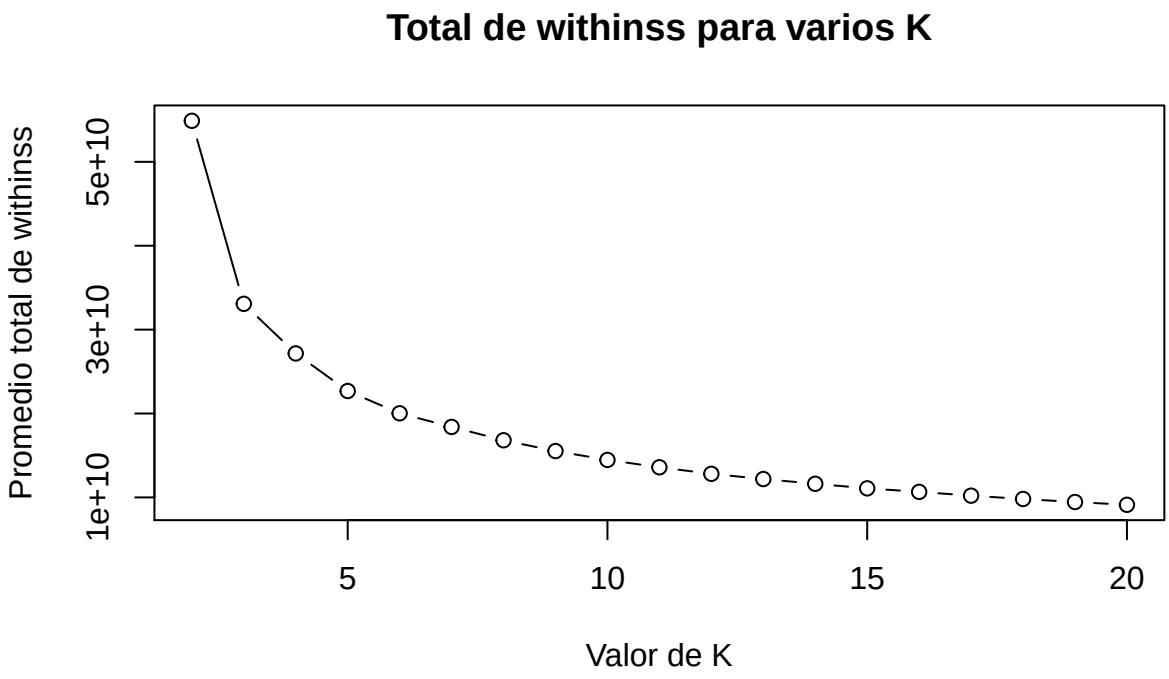


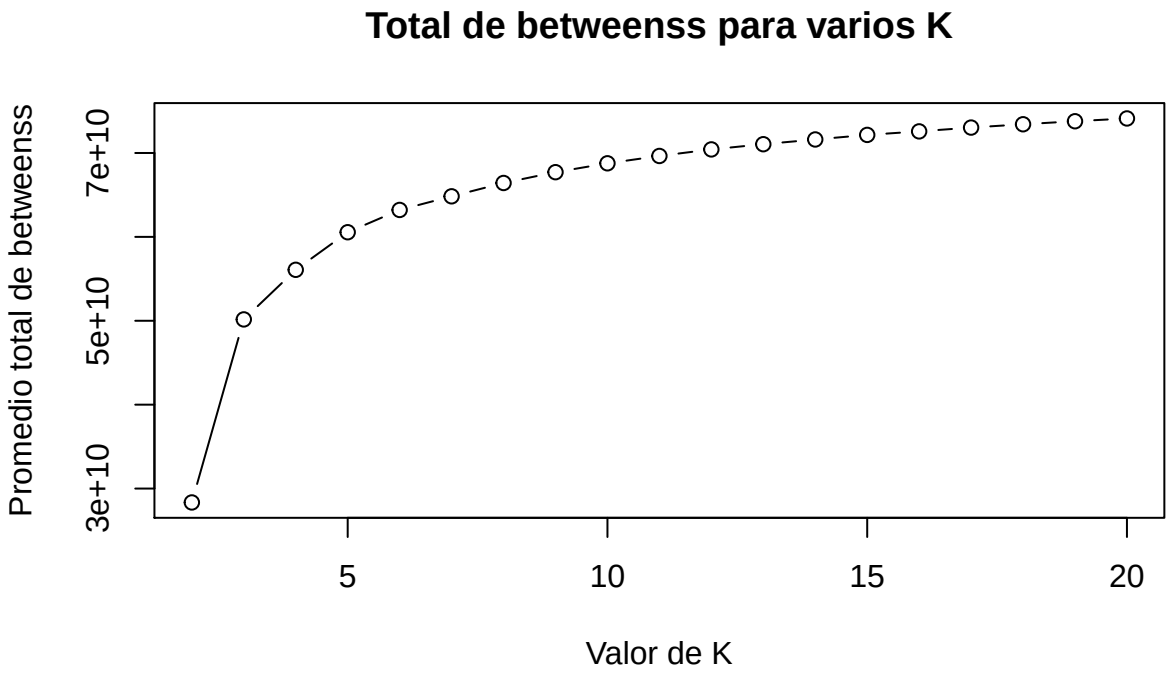
Ejercicio 2

Se aplica el proceso de minería de datos para construir un modelo descriptivo basado en agrupación sobre el conjunto de datos "WHOLESALE CUSTOMERS" para determinar patrones de compra de clientes.

El conjunto de datos contiene algunos outliers que pueden afectar el resultado del algoritmo k-medias, en la etapa de limpieza se procede a eliminar estos clientes. Además es necesario eliminar los atributos "Channel" y "Region" ya que funcionan como identificadores y no son útiles para la agrupación.

Para validar el modelo se observan, para varios modelos con distintos valores de k y varias repeticiones, la suma del cuadrado de las distancias entre los elementos de los grupos y los centros (withinss) y la suma del cuadrado de las distancias entre los centros de los grupos (betweenss).





Se observa en los gráficos que a partir de 5 grupos la ganancia en las distancias se reduce lo suficiente como para seleccionar este valor de K.

Luego se construye el modelo usando el algoritmo K-Means con k igual a 5 y a continuación se muestran los centros de cada uno de los cinco grupos creados y el número de clientes en cada grupo.

```
##       Fresh      Milk   Grocery   Frozen Detergents_Paper Delicassen
## 1 18649.606  3335.586  4497.848 3301.747        1046.5859   1450.566
## 2  5830.214 15295.048 23449.167 1936.452       10361.6429   1912.738
## 3  4238.892  7725.289 11011.747 1336.566        4733.3614   1400.530
## 4  5845.392  2337.319  2878.205 2766.596         660.2952    858.994
## 5 35922.387  4851.806  5862.581 3730.677        1004.6129   1552.161
```

```
##
##   1   2   3   4   5
##  99  42  83 166  31
```

De las características observadas en cada grupo podemos observar lo siguiente para cada uno de estos:

- Grupo 1: Tiene un consumo alto de productos frescos y un consumo medio del resto de las categorías.
- Grupo 2: Tiene un alto consumo de productos lácteos, enlatados, detergentes y delicateses, mientras mantienen un consumo medio en las demás categorías.
- Grupo 3: Tienen un consumo medio en comparación con el resto de los grupos en todas las categorías.
- Grupo 4: Mantiene un consumo medio o bajo en todas las categorías, exceptuando la de productos congelados en donde presentan un mayor consumo respecto a las demás categorías.
- Grupo 5: Tiene el mayor consumo de productos frescos y congelados.

Ejercicio 3

Se aplica el proceso de minería de datos sobre el conjunto de de datos "Bank Marketing" con el fin de realizar un análisis de asociación para encontrar asociaciones frecuentes que caractericen a los clientes que podrían subscribirse o no a los servicios de depósitos bancarios. Se utilizó un umbral para el soporte de 70% y una confianza de 80%.

Luego de cargar el conjunto de datos, es necesario discretizar las variables "age", "duration", "campaign", "pdays", "previous", "emp.var.rate", "cons.price.idx", "cons.conf.idx", "euribor3m", "nr.employed".

Despues de construir el modelo se eliminan las reglas redundantes y con valor lift menor a 1.

```
##   lhs                          rhs    support   confidence lift
## 4 {poutcome=nonexistent}  => {y=no} 0.7844137 0.9171161  1.029880
## 5 {previous=0.000}        => {y=no} 0.7844137 0.9171161  1.029880
## 6 {pdays=[508.80,999.00]} => {y=no} 0.8759408 0.9113412  1.023395
## 1 {}                      => {y=no} 0.8905074 0.8905074  1.000000
```

Según las reglas obtenidas del modelo, las variables tomadas en cuenta son:

- pdays: representa el número de días desde que el cliente fue contactado en la última campaña, si este valor está en el intervalo [507.95,999.00] (999.00 quiere decir que el cliente no fue contactado anteriormente) el cliente no se subscribe al servicio.
- poutcome: Si no existe un resultado de la campaña anterior, lo que indica que el cliente pudo no haber sido contactado, el cliente no se subscribe al servicio.
- previous: si el número de contactos previos está entre cero y 1.18 el cliente de se subscribe al servicio.