PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE MATEMÁTICAS
DEPARTAMENTO DE ESTADÍSTICA

Diplomado en Data Science:
Algoritmos de agrupacioón, K-means, K-mode, K-prototype

Profesora: Valeria Leiva

1. En cada uno de los siguientes apartados, indique si la afirmación es verdadera o falsa. En ambos casos, justifique.

   a) En el algoritmo `k-means`, es posible calibrar el valor de k utilizando un conjunto de datos de validación.

   b) El algoritmo `k-Modes` es la primera extensión del algoritmo `k-Means` orientada al agrupamiento de datos cualitativos.

   c) El algoritmo `k-Prototypes` es un algoritmo de agrupamiento restringido que permite agrupar grandes conjuntos de datos mezclados

2. El archivo `NCI60` contiene información sobre datos de microarray de líneas celulares de cáncer: medidas sobre expresión de 6830 genes (variables) sobre 64 líneas celulares cancerígenas (observaciones). El formato de este set es una lista con dos elementos: una matriz con los valores de expresión génica (data) y un vector con el nombre de los tipos de cáncer (labs). El objetivo del estudio es determinar si las observaciones se agrupan en distintos tipos de cáncer.

   a) Estandarice las observaciones, de modo que todas las variantes genéticas estén en la misma escala.

   b) Implemente el algoritmo `k-means`. Comente.

3. Suponga que la(s) siguiente(s) línea(s) de código(s) representa conjunto de datos de transacciones y datos demográficos de clientes con dos características numéricas (cantidad y frecuencia de las transacciones) y dos características categóricas (grupo de edad y estado civil).

```
# Python
X = np.hstack((np.random.rand(100, 2) * 100,
np.random.randint(4, size=(100, 2))))
# R
X  = cbind(matrix(runif(200, 0, 100),
ncol = 2), matrix(sample(0:3, 200, replace = TRUE), ncol = 2))
```

   Implemente el algoritmo `k-Prototypes` que le permita segmentar a los clientes en tres grupos en función de sus hábitos de gasto y características demográficas. Comente.