MBD – Estadística – Práctica III (CLUSTERING)

Descripción

La compañía AIRETOUCH trabaja en la creación de unos brazos robotizados de gran movilidad. Para imitar la gestualidad humana ha implementado un sistema con varios componentes y sensores que recogen la actividad muscular con el propósito de usar estos datos para poder asignar movimientos específicos al brazo robotizado.

La compañía ha recogido la actividad muscular del brazo humano para cuatro tipos de movimiento: **piedra, papel, tijera** u **ok**. Los 3 primeros corresponden a los gestos del juego clásico y el último es el clásico signo de "*ok*". Cada registro del conjunto de datos tiene 8 mediciones (*M*) consecutivas de cada uno de los 8 sensores (*S*) dando lugar a 64 columnas de datos. La última columna contiene la variable respuesta con el gesto correspondiente.

Predecir el gesto realizado a partir de los sensores ayudará notablemente al desarrollo de nuevos brazos robotizados en el futuro.

Objetivos

1. Agrupar las mediciones de los sensores procurando mantener las máximas similitudes entre clústeres y la máxima heterogeneidad entre ellos. Se usará la técnica de *clustering no supervisado del* ***k-means*** *con* los datos de entrenamiento sin considerar la variable respuesta
2. Construir un modelo predictivo que sea capaz de clasificar los individuos en una de las 4 categorías de la variable respuesta. Para realizar la predicción sobre los datos de test, se deberán usar algunos o todos o combinaciones de 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

Datos

Existen 2 conjuntos de datos:

1. Datos de entrenamiento. Contendrán la variable respuesta (gesto real del humano) en las 4 categorías. Se usarán para construir la parte de clustering NO supervisado y para construir el modelo de clustering supervisado que utilizaremos con la muestra test.
2. Datos test. No contendrá la variable respuesta y únicamente servirá para que el profesor evalúe la capacidad predictiva de los algoritmos.

Variables

*S**x**M**y***: Medida numérica ***y*** proporcionada por el sensor ***x***

y: variable respuesta de 4 categorías

Evaluación

Evaluación de la primera parte. Se valorará, entre otros aspectos:

- La elección del número de clústeres basada en algún criterio
- Alguna representación gráfica que muestre visualmente la bondad del agrupamiento

Evaluación de la segunda parte. Se valorará, entre otros aspectos:

- La capacidad predictiva (porcentaje de acierto) obtenida en el conjunto test.
- El esfuerzo adicional por mejorar el porcentaje de acierto con algún o algunos algoritmos.

En ambas partes, también se considerará la explicación, interpretación, presentación y concisión de los resultados.

Entrega

La fecha límite para realizar la entrega será el día **22/12/2022 a las 23:55** a través del campus. Se deberán entregar 3 ficheros:

1. **Informe**. (Formato: *.docx o pdf o html).* Debe contener como mínimo:
    a. Clustering no supervisado (Extensión aprox.: 2-3 páginas*)*
        i. El número de grupos escogido y razonamiento
        ii. El porcentaje de variabilidad explicada por dicho número de grupos.
        iii. Cualquier representación visual de la agrupación.
    b. Clustering supervisado (Extensión aprox.: 5-6 páginas*)*
        i. Enumeración del algoritmo o algoritmos utilizados.
        ii. Parámetros testados con cada algoritmo (si hubiere)
        iii. Algoritmo y parámetros usados para hacer la predicción final en la muestra test
        iv. Si se hubiese dividido la muestra de entrenamiento, a su vez, en una muestra de entrenamiento y test, indicar la proporción de acierto hallada en esta sub-muestra test (dentro de los datos de entrenamiento)
        v. Se valorará la presencia de gráficos y/o tablas que faciliten la comprensión. Por ejemplo, una tabla con la proporción de acierto por cada algoritmo.
        vi. Se valorará cualquier sistema para identificar variables relevantes dentro del conjunto de datos para realizar las predicciones.
        vii. Se valorará cualquier esfuerzo por mejorar la capacidad predictiva de los algoritmos usados.

2. **Predicciones**. Fichero de texto con una única columna con las predicciones elegidas (nombre de la categoría) para los datos test. Esta columna debe tener cabecera "y". Formato: *.txt.* Es muy importante que este fichero se llame exactamente "**p3.txt**"
3. **Código**. El código comentado utilizado para realizar la práctica. Formato: . *R*

ANEXO: Comentarios adicionales

1. Algunos algoritmos o funciones pueden ser costosas computacionalmente. Si alguna función tarda mucho en ejecutarse, existen alternativas:
    a. Escoge los parámetros de forma conveniente para reducir el coste computacional.
    b. Reduce la dimensionalidad usando componentes principales (*princomp*).
    c. Reduce el número de individuos, haciendo un muestreo de los mismos.
    d. Busca otra función o algoritmo alternativo.

2. Para generar el fichero de texto con las predicciones, se hará con las siguientes sentencias:

```
nombre_objeto <- data.frame(y=pr)     # pr: predicciones
write.table(nombre_objeto,   'p3.txt',   row.names   =   FALSE,
col.names = TRUE, sep='\t', quote = FALSE)
```

3. Podéis comprobar vuestro porcentaje de acierto a través de una shiny-app disponible en: http://shiny-eio.upc.edu/jordi/p3_2019/
   [Es una versión beta que puede proporcionar errores. Por favor, escribid al profesor si tenéis alguna incidencia]