MBD – Estadística – Práctica II (CLUSTERING)

Descripción

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

Datos

Existen 2 conjuntos de datos:

1. Datos de entrenamiento. Contendrán la variable respuesta (actividad real del usuario en las 6 categorías). Se usarán para construir la parte de clustering NO supervisado y para construir el modelo de clustering supervisado que evaluaremos con la muestra test
2. Datos test. No contendrá la variable respuesta y únicamente servirá para que el profesor evalúe la capacidad predictiva de vuestros algoritmos.

Variables

feat_1 a *feat_561*: datos de posicionamiento proporcionados por los sensores del teléfono móvil

subject: identificador del voluntario que tiene el teléfono (1 a 21).

activity: variable respuesta de 6 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 e interpretación de los resultados y la presentación de los mismos.

Entrega

La fecha límite para realizar la entrega será el día **20/12/20 a las 23:55** a través del campus. Se deberán entregar 3 ficheros:

1. **Informe**. Debe contener como mínimo:
    a. Clustering no supervisado (Extensión aprox.: 2 páginas. Formato: *.docx, pdf o html)*
        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. Formato: *.docx, pdf o html)*
        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 (para hacer la predicción final, deberéis usar únicamente un algoritmo)
        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 (sacada de los datos de entrenamiento) para los algoritmos testados. Se recomiena una tabla resumen al final del informe.
        v. Se valorará la presencia de gráficos que faciliten la comprensión.
        vi. Se valorará cualquier sistema para hallar variables relevantes dentro del conjunto de datos para realizar las predicciones.

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 (=”activity”). Formato: *.txt.* Es muy importante que este fichero se llame exactamente “p2.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 la función de forma conveniente para reducir el tiempo de computación.
    b. Reduce la dimensionalidad de los datos 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(activity=pr) # pr: predicciones
write.table(nombre_objeto, 'p2.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/p2/
   [Es una versión beta que puede proporcionar errores. Por favor, escribid al profesor si tenéis alguna incidencia]