Ciencia de los Datos - Práctica Final 2024-2025

Entrega: subir a Moodle, antes del domingo 27 de abril a las 23:59 horas, un documento PDF en el que se incluya, para cada pregunta:

- El enunciado de la pregunta.
- Los comandos de R utilizados para responder a la pregunta.
- Los resultados mostrados por R como respuesta a cada comando.
- Copia, en su caso, de las gráficas generadas por R.
- Cuando se pida, el análisis de los resultados obtenidos.

Valoración: 100% de la nota final.

1 Clustering

Se les ha pedido a dos conductores que realizaran el mismo trayecto con el mismo vehículo tres veces de modo que cada vez aplicasen uno de estos modelos de conducción: 0 un estilo de conducción tranquila, 1 un estilo de conducción normal, y 2 un estilo de conducción agresiva.

Durante estas pruebas se han registrado varias métricas del coche, que están registradas en el fichero "Datosconduccion" (licencia CC-BY 4.0 Yushetf López). Se nos ha informado que durante la toma de muestras se ha visto que los datos de velocidad angular en los ejes X, Y y Z del giroscopio no están siendo bien registrados y muestran valores anómalos. No sabemos si debemos descartarlos.

Se requiere:

- Aplicar K-means para clasificar las muestras y comparar los resultados con los estilos de conducción proporcionados.
- Analizar qué variables tienen más peso en la clasificación.
- Evaluar si los clusters coinciden con los estilos de conducción subjetivos esperados.

Valoración del ejercicio Se valorará el ejercicio considerando la aproximación al problema de cada estudiante, la creatividad, la metodología, el tipo de gráficas usadas, el análisis que realiza, la discusión que sostiene, la claridad y orden en la exposición de los resultados, y cualquier otro factor que permita justificar la madurez crítica del estudiante de master en un análisis de ciencia de los datos.

Algunas ideas del proceso:

1. **Cargar librerías y entornos necesarios:** Indica qué librerías vas a utilizar y por qué.
2. **Importar y explorar datos:** Describe tu análisis de los datos, es decir, cómo lo haces y qué observas. Esto incluye, entre otros, si hay que hacer limpieza de datos nulos, si hay que ajustar decimales que están en distintos sistemas, si todas las columnas tienen el mismo tipo de datos, etc.
3. **Preprocesar los datos:** Elige las variables numéricas relevantes. Valora si hay que hacer escalado de alguna columna y en ese caso qué tipo de escalado realizarías y por qué.
4. **Aplicar K-means:** Explica el proceso y el código usado. Recuerda comentar los resultados del ajuste, qué observas, qué confianza te da el ajuste, cuántos datos hay en cada clúster, qué pasa si cambias el número de centroides, ¿tiene sentido?, ¿qué pasa si restringes las variables analizables?, etc.

5. **Evaluación de los resultados:** Utiliza herramientas que ayuden visualmente a discernir si tiene sentido el ajuste realizado, si no hay mucho solapamiento de datos, intenta evaluar qué peso tiene cada variable en cada cluster.

6. **Comparación con los estilos subjetivos de conducción:** ¿Cómo lo haces? ¿Qué se puede concluir?

2 Series temporales

Los datos sobre los que se va a trabajar en este proyecto proceden de la página Berkeley Earth

http://berkeleyearth.org/data/, sección Quality Controlled, subsección TAVG.

En ella se reúnen **registros de temperaturas en distintos puntos del planeta** durante amplios periodos de tiempo.

Archivos de datos para el proyecto:

1. 'Whole years.txt': 5235 registros de años completos con formato: Station ID,Whole year.

2. 'data series.txt': 68493 registros con formato: Station ID,Series Number,Date,Temperature.

En el primer archivo se indican los años para los que hay datos completos para cada una de las estaciones meteorológicas, donde cada estación está ubicada en una ciudad distinta.

En el segundo archivo se guardan los registros de las temperaturas medias mensuales para cada una de las estaciones/ciudades.

Se pide:

1. Localiza las ciudades que tienen datos completos desde 2000 hasta 2019.

2. Para cada una de ellas, crea una serie temporal con los 240 datos de los 20 años entre 2000 y 2019. Usa las opciones frequency=12 y start=1999.

3. Ajusta a cada una de las series un modelo ARIMA con componente estacional. Razona los parámetros escogidos (d y D, mediante las varianzas de las diversas diferencias, y p, q, P, Q usando las gráficas de ACF y PACF).

4. Utiliza el modelo ARIMA para realizar una predicción de la temperatura media esperada para el mes de junio de 2020 en cada una de las ciudades. Compara la predicción con la media en esa ciudad de la temperatura en los meses de junio de 2000 a 2009.