Mentoría Data Science aplicado a BCI

Consignas TP1: Análisis y Visualización

Pautas generales

Se propone la elaboración de un informe, en el cuál se exploren las respuestas a los interrogantes planteados. Este informe debe presentarse en formato estático (pdf, html), y deberá apuntar a un público técnico, pero sin conocimientos específicos en el tema. Estos documentos deberán ser acompañados por los notebooks o los códigos python con los que los realizaron, estos no se corregirán en detalle pero es importante incluirlos para garantizar la reproducibilidad de sus resultados.

Siempre tener en cuenta que la presentación de los resultados debe ser clara, apoyar las conclusiones extraídas, y ser lo más sintética posible. No necesariamente todo lo que prueben debe estar reflejado en el informe. Ciertamente valoramos el esfuerzo de explorar alternativas de análisis y visualización, pero parte del trabajo también es elegir las más adecuadas.

Introducción

En caso de que no haya quedado completamente claro en la llamada introductoria, o que algo se nos haya pasado, les recordamos que el dataset consiste en 5 (cinco) señales EEG, cada una en un archivo TXT con formato CSV (Comma Separated Values).

Los nombres asignados a cada archivo representan a qué sesión de registro pertenecen: las dos primeras letras refieren al sujeto, y el número al conteo de sesión trabajada con dicho sujeto. Esto quiere decir que si aparecen los archivos JX0, JX1 y NB0, entonces el dataset consistiría en dos sujetos JX y NB, de los cuales se tomaron 2 y 1 sesiones respectivamente. Esto es importante porque muchas veces separaremos las consignas en analisis dentro de un mismo registro, entre registros de un mismo sujeto y entre sujetos.

Dentro del archivo se encuentran 11 columnas correspondientes a:

No.Muestra-Ch1-Ch2-Ch3-Ch4-AccX-AccY-AccZ-Tags-Time-Timestamps

donde **Ch** refiere a canales de electroencefalografía y **Acc** a los acelerómetros.

¡Cualquier otra duda nos contactan por el medio que hayamos definido! ¡Éxitos!

Parte I: Exploración de la base de datos.

A) Leer los datos, eliminar los metadatos innecesarios.
B) Describir las características generales del dataset:
    - Número de registros, diferencias entre los mismos.
    - Definir la conveniencia de trabajar todos juntos como un solo dataset, o por separado.
C) Analizar las columnas presentes en el dataset:

- ¿Todas las columnas son relevantes? ¿Cuáles contienen información útil?
- ¿Qué tipo de datos contienen? ¿Qué variables describen las columnas consideradas? ¿Con qué sensibilidad?
- En el caso de los datos cualitativos. ¿Cuáles son los valores posibles para esta variable? Describa su presencia (frecuencia, intervalos, secuencia, etc.)
- Para todas las columnas, ¿hay datos dañados? ¿valores nulos? ¿Qué estrategia considera más pertinente para abordar esos datos? Justifique.
- Suponiendo que los datos se adquieren a una frecuencia de muestreo exacta 200Hz, ¿cómo se manifiesta esta información en el número de muestras presentes en el registro?
- Determine la forma más adecuada de parsear los datos temporales para poder graficar las señales en el dominio del tiempo.

D) Generar visualizaciones de ejemplo para las series temporales provistas. Determinar los intervalos de tiempo más adecuados para generar visualizaciones claras que permitan comparar las señales en los siguientes escenarios:
   - Un sujeto - todos los canales
   - Un mismo canal - todos los sujetos
   - Un mismo canal - mismo sujeto en diferentes sesiones.
   - Un mismo sujeto y canal - diferentes estados
   - Mismo estado - diferentes sujetos

Al analizar estas visualizaciones, ¿extrae alguna información que considere relevante para el problema? ¿Se observa algún fenómeno distinguible a primera vista?

Parte II: Dominio del tiempo.

A) Nivel Segmento/Estado: Seleccione los datos correspondientes a un paciente y un canal, y para él defina un conjunto de señales para cada estado presente en el dataset. Para cada uno de ellos estudie los siguientes elementos y luego compárelos.
   a) ¿Presenta los valores de voltaje una distribución normal? Utilizar un criterio gráfico y un test para probarlo. Si la distribución normal no se ajusta, ¿a qué distribución se asemejan?
   b) Realice un resumen estadístico de los valores de voltaje en el intervalo de tiempo considerado. ¿Qué estimador de posición central usaría para describir los valores? ¿Y de dispersión?
   c) En adición a los datos dañados encontrados en la parte I, ¿Encuentra outliers a este nivel de análisis? ¿Estos outliers deberían ser tratados de forma diferencial? ¿De qué manera?
   d) ¿Existe una diferencia estadísticamente significativa para considerar que los estimadores de posición central son diferentes entre los estados? Use un test de hipótesis para probarlo al menos entre dos estados.
   e) Resuma las principales conclusiones de este nivel de análisis.

B) Nivel Paciente - un canal: Seleccione los datos correspondientes a un paciente y un canal de adquisición y para ese caso estudie los siguientes elementos:

a) Considere el conjunto completo de valores de voltaje correspondientes a cada uno de los estados a lo largo de todo el registro y repita los elementos del apartado II-A).
b) Ahora que dispone de más datos, ¿son variables independientes el estado registrado de la señal y su voltaje? Use herramientas cuantitativas y cualitativas para justificar su respuesta.
c) Para cada uno de los estados, los valores de voltaje a lo largo del tiempo, ¿varían con alguna tendencia?
d) Resuma las principales conclusiones de este nivel de análisis.

C) Nivel Paciente - multicanal: Seleccione los datos correspondientes a un paciente y para ese caso estudie los siguientes elementos:
a) Las señales de voltaje en función del tiempo para cada canal, ¿son variables independientes entre sí? Use herramientas cuantitativas y cualitativas para justificar su respuesta. (Ejemplo, matriz de correlación)
b) Tomando los puntos que considere relevantes del apartado II-B) para cada canal y considerando la respuesta anterior. ¿Considera relevante trabajar con todos los canales disponibles o podría quedarse con un subconjunto? Si elige el subconjunto, ¿qué canales elegiría y por qué?
c) Opcional: Tomando un par de canales a elección, analice la distribución conjunta de los valores de voltaje para cada estado de forma cualitativa. (Ejemplo, Heatmap, scatterplot, 3D-mesh, etc.)
d) Resuma las principales conclusiones de este nivel de análisis.

D) Nivel Multi-Paciente.
a) A partir de las conclusiones extraídas de los niveles de análisis anteriores. Decida cuáles son los aspectos más importantes a analizar de los registros de un paciente y compárelos entre pacientes. ¿Encuentra diferencias significativas? ¿Qué variables pueden identificar esas diferencias?

   A modo de ejemplo: los valores de voltaje medios para cada estado de un paciente, ¿difieren significativamente entre pacientes?