Ciencia de Datos 2022

Lista 04

03.mayo.2022

1. (No entregar)
   Toma el libro *Applied Multivariate Analysis* de Everitt y Hothorn. Leer y reproducir los ejemplos de la sección 6.1 a 6.5.

2. (No entregar)
   Lee las secciones 1.1, 1.2 y 1.3 del libro de Giraud *Introduction to High-Dimensional Statistics*. (Ver también la lectura de Raúl Rojas sobre el tema de la *maldición de la dimensionalidad* que se indica en la lectura del 11.03.2021)

   a) Replica las gráficas de las figuras 1.4 y 1.5 del libro.
   b) Prueba o da evidencia empírica del siguiente hecho: En altas dimensiones, si elegimos dos vectores aleatorios (en el cubo unitario), la probabilidad de que sean ortogonales aumenta con la dimensión d. A esto le llamamos la *propiedad de casi ortogonalidad*.

3. ¿Cuál método es más sensible a datos atípicos: k-medias o agrupamiento jerárquico? Motiva ampliamente tu respuesta con ejemplos.

4. Haz un análisis de agrupamiento para los datos del heptatlón. Están disponibles en el archivo `heptatlon.csv`. Relaciona los resultados con otros patrones en el comportamiento de los datos.

5. Considera los datos del proyecto de la Universidad de Oxford sobre las diferentes medidas que los gobiernos tomaron para enfrentar COVID-19: `https://covidtracker.bsg.ox.ac.uk/`

   Se pueden descargar los datos desde:
   `https://raw.githubusercontent.com/OxCGRT/covid-policy-tracker/master/data/OxCGRT_latest.csv` con mayor información en
   `https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/codebook.md`

   Haz un análisis de datos, principalmente un análisis de agrupamiento por país, con las medidas vigentes al inicio de enero 2021, limítate a las variables del grupo *Containment and closure policies*.