

# Aprendizaje Estadístico 2024

Lista 04

06.marzo.2024

1. Leer 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)
  - a) Replica las gráficas de las figuras 1.4 y 1.5 del libro.
  - b) Probar 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**.

2. ¿Cuál método es más sensible a datos atípicos:  $k$ -medias o agrupamiento jerárquico? Motivar ampliamente su respuesta con ejemplos.

¿Cómo se puede resolver el problema de los datos atípicos a la hora de realizar un método de agrupamiento. Exponga sus ideas y evidéncielas a través de un ejemplo.

3. Realizar el ejercicio del sistema de recomendación, con los datos de `recommender.ipynb`, `movies.csv`, `ratings.csv` usados en el aula del 07 de febrero.
4. Hacer un agrupamiento jerárquico con los datos de los países `countries_binary.xlsx` de la clase de 29 de febrero.

Visualizar los resultados o dendrogramas de diferentes métodos de agrupamiento, variando los siguientes:

- el método de agrupamiento (simple, completo, promedio, Ward)
- la métrica utilizada (euclídeana, Hamming).

5. Realizar un análisis de agrupamiento  $k$ -means, nuevamente para los datos de los países, que están disponibles en el archivo `countries_binary.xlsx`.

Con estos datos, hacer la proyección PCA de los datos a 2 componentes principales, e ilustrar cómo se ven las regiones de clasificación obtenidas con su método de agrupamiento en este espacio 2-dimensional.

Ilustrar cómo cambian las regiones de clasificación cuando se utilizan diferentes métricas en el  $k$ -means.

6. Con una (o varias) imágenes RGB de su elección, realizar un algoritmo de cuantización de colores para la imagen, usando como base un algoritmo de  $k$ -medias.

Ilustrar los resultados obtenidos comparando qué ocurre:

- al aumentar el número  $k$  de grupos.
  - al modificar la métrica usada en el algoritmo de  $k$ -means.
-