UNIVERSIDAD EAFIT

Escuela de Ciencias

Maestría en Ciencias de los Datos y Analítica

Álgebra para Ciencias de los Datos

Taller 1

NOTA: _____

NOMBRE: _____ CÓDIGO: _____

GRUPO: ____ PROFESOR: Henry Laniado. Fecha de entrega: Viernes de la tercera sesión. En grupos de a 5

---

1. Pruebe que:.

   a) Que una sucesión creciente de conjuntos es convergente. Describa un ejemplo de una sucesión de conjuntos que sea convergente y otro de una divergente.

   b) Cualquier bola abierta es un conjunto abierto

   c) La suma finita de métricas es una métrica

   d) Una combinación lineal convexa de métricas es una métrica.

   e) La distancia de Mahalanobis es una métrica

   f) Si $d : X \times X \longrightarrow \Re$ es una métrica, entonces $\bar{d}(x, y) = \frac{d(x,y)}{1+d(x,y)}$ también lo es.

   g) Si $d : X \times X \longrightarrow \Re$ es métrica, entonces $\bar{d}(x, y) = \text{mín}\{1, d(x, y)\}$ lo es.

   h) Si $A \subset B$, ambos subconjuntos de $\Re^n$, entonces para cualquier $x \in \Re^n$ y d una métrca, se tiene que $d(x, A) \geq d(x, B)$

   i) Que la norma de Frobenius satisface las condiciones para ser norma matricial.

   j) Que la norma 2 matricial de A es la raiz cuadrada del mayor autovalor de A

2. Defina qué es una pseudométrica y muestre algunos ejemplos de pseudométricas

3. Considere A un subconjunto de un espacio métrico X y sea $\epsilon > 0$. Interprete el conjunto

$$A_\epsilon = \{x \in X \mid d(x, A) < \epsilon\}.$$

   Cómo relaciona A_{ϵ_1} y A_{ϵ_2} cuando $\epsilon_1 < \epsilon_2$. Calcule X_ϵ y ϕ_ϵ

4. Simule 10000 aleatorios de una distribución Normal bivariante. Para cada una de las métricas de la diapositiva 26, calcule todas las distancias de cada dato a su media. Pinte de rojo los puntos cuya distancia a la media se encuentra en el 10 % de las mayores distancias.

5. Consulte y programe el algoritmo de $k-$ means. Prográmelo siendo los inputs, el número de grupos, la distancia p con $p = 0, 1, 2, \ldots$, y el número de iteracciones hasta la parada. Suponga que $p = 0$ es la distancia de Mahalanobis. Simule tres muestras aleatorias de distribuciones normales bivariantes con distintas medias, luego haga un análisis de qué distancia tiene un mejor desempeño para clasificar las muestras. (enviar código)

6. Programe la distancia de Mahalanobis utilizando la covarianza habital, luego la covarianza bajo el shrinkage de Ledoit and Wolf. ($cov1para.m$), y la covarianza y vector de medias robustos obtenida bajo el método de mínima curtosis ($kurmain.m$). Ilustre ejemplos concretos donde el shrinkage y el método robusto presenta un mejor rendimiento y comente los resultados.

7. El fichero (porfolio100.txt) tiene rentabilidades mensuales para 100 sectores económicos, desde julio de 1963 hasta febrero de 2019, un total de 668 registros. Realice un análisis de identificación de outliers utilizando los tres métodos del punto anterior y el método de mínima curtosis con la salida idx. Comente los resultados.

8. En el fichero (porfolio100.txt), saque la media de cada fila, es decir el vector de media será un vector de 668×1. Defina la variable binaria como 1 donde el vector de medias es positivo y cero en otro caso. Defina la matriz binaria que vale 1 si en portfolio100 hay un valor positivo y cero en otro caso. Con las métricas binarias (Pearson, Jaccard y Dice) identifique los 10 activos más parecidos y los 10 menos parecidos al vector binario de medias.

9. Describa y ejecute un proceso de identificación de outliers en variables binarias. Utilícelo para identificar qué meses en portfolio100, versión binario siendo 1 si hay valor positivo, son considerados meses atípicos. Cómparelos con los meses identificados en el punto 7.

10. Saque una foto suya y defina una sucesión de imágenes que sea convergente a su foto. Muestre los 10 primeros elementos de la sucesión y el elemento 1000-ésimo. Como la sucesión de imágenes es convergente, obtenga la imagen tal que las imágenes siguientes de la sucesión tienen una distancia (en norma (1,2, ∞ y Frobenius)) menor a 0.01.

11. Envíe al grupo 4 imágenes de su rostro. Pase cada imagen a escala de grises. Para las cuatro normas matriciales discutidas en clase (1,2, ∞ y Frobenius). Calcule, con la métrica inducida por las normas, la distancia de cada persona a todas las personas. Defina un indicador de lejanía del individuo j como el promedio de las distancias del individuo j a todos. Un concepto sencillo de imagen mediana sería aquel individuo cuyo indicador de lejanía es el menor. Obtenga con las cuatro métricas quién de ustedes es la mediana, es decir, el más típico.

12. Con las cuatro métricas, determine quien del grupo es el más parecido a usted y discuta qué métrica es la más conveniente para identificarlo. Ensaye introduciendo en la base un par de fotos suyas más. Construya una vecindad con centro en usted y un radio tal que la vecindad tenga 5 imágenes. Muestre las imágenes. Explique con buenos argumentos si su imagen es punto interior, punto frontera o punto de acumulación del conjunto de imágenes del grupo.

13. Calcule el número condición de la matriz de covarianzas de portfolio100. Implemente alternativas para mejorarlo o reducirlo a la mitad.

14. Sea $H_n(i,j) = \frac{1}{2i+2j-1}$, llamada la matriz de Hilbert. Realice una gráfica de n en el eje x con el número condición en el eje y. Qué tipo de comportamiento observa. Haga lo mismo para su determinante. Realice lo mismo utilizando el shrinkage de Ledoit and Wolf pero cambiando en la linea 18 de cov1para.m $sample = (1/t).*(x'*x)$; por $sample = H_n$. Compare los resultados. Haga una análsis gráfico y de visualización donde se observe si al final el shrinkage mejora el número condición.

15. Considere $x = [1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12]$. Defina $b = H_{12}x$. Resuelva el sistema con la forma $x = H^{-1}b$. Qué conclusión obtiene. Busque alternativas para resolver el problema observado.

16. Realice lo mismo que los dos puntos anteriores pero con la matriz de Vandermonde en lugar que la Hilbert. Consulte qué tipos de problemas útiles de aplicaciones matemáticas en analítica de datos usan la matriz de Vandermonde y la de Hilbert. Tiene alguna sugerencia para contribuir a mejorar estas aplicaciones matemáticas? ya que son problemas muy mal condicionados.