Ciencia de Datos 2021

Lista 06

02.mayo.2021

1. (No entregar)
   Existen ciertas relaciones entre PCA y métodos de agrupamiento. Para aquellos que se interesen en eso, les recomiendo:
   `http://ranger.uta.edu/~chqding/papers/KmeansPCA1.pdf`

   Bajo ciertas restricciones, calcular el primer componente de PCA y aplicar la funcion $sign()$, genera la partición obtenida con K-medias. Contrario a agrupamiento, es un resultado exacto.

2. Implementa el algoritmo EM para encontrar grupos en un conjunto de datos en el plano $\mathbb{R}^2$ con una mezcla de K distribuciones Gausianas. Puedes tomar $K = 2$ (o más). Evalúa tu algoritmos y cómpáralo con K-medias en los siguientes tres experimentos:

   i) Dos o más gaussianas considerablemente separadas
   ii) Dos o más gaussianas no tan separadas
   iii) Datos formando alguna estructura particular (círculos, lunas, o similares).

   (Puedes usar conjuntos de datos ya existentes o generas datos sintéticos).

3. Supongamos que (X, Y) son variables aleatorias discretas con la siguiente distribución conjunta: Queremos predecir Y con

	$X = 1$	$X = 2$	$X = 3$	$X = 4$
$Y = 0$	0.1	0.05	0.05	0.15
$Y = 1$	0.12	0.1	0.25	0.18

   base en el valor observado para X.

   a) Calcula el clasificador Bayesiano óptimo si equivocarse de categoría tiene costo 1 y no equivocarse tiene costo 0. ¿Cuál es el costo (error) promedio para este clasificador?
   b) Calcula el clasificador Bayesiano óptimo si clasificar una observación mal cuando el verdadero valor es $Y = 1$ tiene un costo 3 y en el otro caso tiene costo 2.

4. Deriva el clasificador Bayesiano óptimo para el caso de tres clases y una función de costo simétrica cuando:

$$X \mid Y = 1 \sim \mathcal{N}(\mu_1, \Sigma), \quad X \mid Y = 2 \sim \mathcal{N}(\mu_2, \Sigma), \quad X \mid Y = 3 \sim \mathcal{N}(\mu_3, \Sigma),$$

   y

$$\mathbb{P}(Y = 1) = 2\mathbb{P}(Y = 2) = \mathbb{P}(Y = 3).$$

   Ilustra con un diagrama (en $\mathbb{R}$ ó $\mathbb{R}^2$), cómo se ven las regiones de clasificación para este caso.

5. En este ejercicio trabajamos con los datos de Iris que Ronald Fisher usó para introducir LDA.

   Construye un clasificador LDA. Compara los resultados con un clasificador K-nn. Para ello, selecciona un subconjunto de datos para entrenamiento y otros subconjunto para propósitos de evaluación. Compara ambos clasificadores de forma visual, y en términos de métricas: matriz de confusión, *accuracy*, F_1-*score*, u otras métricas similares.
   Discutir los resultados.

   Es interesante echar un ojo al artículo original de Fisher:
   `https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1469-1809.1936.tb02137.x`