Por favor, para las próximas tareas integrar
la entrega en un único documento pdf.

Reconocimiento Estadístico de Patrones
Tarea 4

Nota: 9/10

Randy Osbaldo Ibarra Cayo

B. Preguntas Cortas

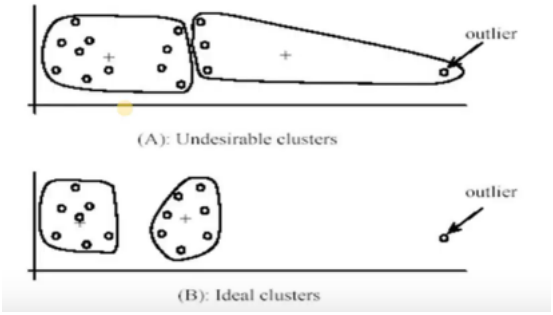
Pregunta 1

Nota: 2/2

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

Solución

Dado que la media es sensible a datos atípicos, se sigue que k-medias al usar la media de cada clúster es sensible a datos atípicos. En la figura se presenta un ejemplo en el que al agregar un solo dato atípico el resultado de k-means cambia bastante, pues podemos notar que la presencia de ese punto hace que uno de los clusters se divida debido a que la media ahora está más alejada.



Ok

Por otro lado, agrupamiento jerárquico no tiene el problema de sensibilidad de la media pues se basa en unir en cada iteración a los dos clusters más cercanos. Dado que los datos atípicos son datos que están "alejados" de los otros datos, estos se uniran a otro cluster una vez que sean considerados cercanos a alguno, lo cual pasará en la iteraciones finales.

Se concluye que k-medias es más sensible a datos atípicos.

En realidad ambos métodos son sensibles. Es posible encontrar ejemplos
donde el método jerárquico destruye los grupos en presencia de atípicos.

Pregunta 2 **Nota: 2/2**

Verifica la propiedad de casi ortogonalidad de dos vectores elegidos al azar (`recpat10.pdf`, pag. 12): Definimos

$$X = \frac{1}{\sqrt{d}}[X_1, ..., X_n]$$
$$Y = \frac{1}{\sqrt{d}}[Y_1, ..., Y_n]$$

donde X_i y Y_i son muestras de Z, con $E(Z) = 0$ y $Var(Z) < \infty$, entonces

$$P(|\langle X, Y\rangle| > \epsilon) \to 0$$

si $d \to \infty$.

Solución

Definimos W_d como

$$\bar{W}_d = \frac{1}{d}\sum_{i=1}^{d} W_i \qquad \text{Ok}$$

donde $W_i = X_i Y_i$. Dado que las muetras son independientes, se sigue que W_i es una sucesión de variables aleatorias independientes y dado que cada muestra proviene de la misma variable aleatoria Z, se sigue que W_i son identicamente distribuidas. Entonces, se tiene que W_i es una sucesión de v.a.i.d. con media μ_w entonces, por la Ley (débil) de los grandes números, para cualquier número positivo ϵ se tiene

$$\lim_{d\to\infty} P(|\bar{W}_d - \mu_w| < \epsilon) = 1 \qquad \text{Ok}$$

Por otro lado, notemos que $E(W_i) = E(X_i Y_i)$, entonces por independencia de X_i y Y_i se sigue que $E(X_i Y_i) = E(X_i)E(Y_i)$ y dado que $E(X_i) = E(Y_i) = E(Z) = 0$, entonces $E(X_i)E(Y_i) = 0$, es decir $E(W_i) = \mu_w = 0$, entonces

$$\lim_{d\to\infty} P(|\bar{W}_d| < \epsilon) = 1$$

es decir

$$\lim_{d\to\infty} P(|\bar{W}_d| > \epsilon) = 0$$

pero

$$\bar{W}_d = \frac{1}{d}\sum_{i=1}^{d} W_i = \frac{1}{d}\sum_{i=1}^{d} X_i Y_i = \sum_{i=1}^{d}(\frac{1}{\sqrt{d}}X_i)(\frac{1}{\sqrt{d}}Y_i) = \langle X, Y\rangle \qquad \text{Ok}$$

por lo tanto

$$\lim_{d\to\infty} P(|\langle X, Y\rangle| > \epsilon) = 0 \qquad \text{Bien!}$$

Bonita prueba

B. Análisis de Datos

Ejercicio 1

Haz un análisis de agrupamiento para los datos del heptatlon. Están en `library("MVA")`.

Solución

Anexa en notebook correspondiente a este ejercicio.

Ejercicio 2

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 bajar los datos desde

`https://raw.githubusercontent.com/OxCGRT/covid-policy-tracker/master/data/OxCGRT_latest.csv`

Mayores informes en

`https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/codebook.md`

Haz 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.

Solución

Anexa en notebook correspondiente a este ejercicio.