

Clusters. Test de Lectura

Test de Lectura

Como discutiesemos, estos *tests de lectura* son una serie de preguntas simples que permiten monitorear si estan siguiendo efectivamente las clases. No seran corregidos ni evaluados, los administraran directamente ustedes. Cualquier duda que tengan sera discutida en nuestras sesiones sincronicas, la semana siguiente a publicado el material, luego de que hayan hecho un esfuerzo en trabajarlo, no antes. Es muy importante que trabajen este material y lo discutan entre ustedes.

1. Intuitivamente, como se define un grupo en analisis de clusters?
2. Verifica la formula de descomposicion del slide 8.
3. V o F: un encoder que minimiza la variabilidad within maximiza la variabilidad between.
4. Este ejercicio les dara una idea de lo que sucede con la formula del slide 10 del metodo de K -medias. Demostrar que

$$V(Y) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{2} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n (Y_i - Y_j)^2$$

Esta expresion da una interpretacion alternativa a la varianza en terminos de ‘dissimilarities’. Cual es?

5. Supongamos que hay una sola variable (peso, por ejemplo) y que hay 100 personas. Describi paso a paso como opera el mecanismo de k -medias para armar dos grupos en base al peso. (Tip: lo cuento en *Big Data*).
6. Por que $W(C)$ cae con la cantidad de grupos?
7. Supongamos que en base a las mas de cincuenta variables de la Encuesta Permanente de Hogares, armamos dos grupos, en base a analisis de clusters. Por que no hay ninguna garantia de que uno de esos grupos sean efectivamente los pobres?
8. Dicho de otra manera, que es lo unico que garantiza K -medias acerca del los dos grupos resultantes?
9. Porque el analisis de clusters es un metodo *no supervisado*?