Universidad de Santiago de Chile
Departamento de Ingeniería Informática
Análisis de Datos

Profesor: Dr. Max Chacón Pacheco
Ayudante: Javier Arredondo Contreras

Laboratorio 2 - Agrupamiento K-medias

1. Objetivos

- Extraer el conocimiento del problema asignado, mediante el uso del software R, utilizando el algoritmo de *clustering K-means* y realizar el análisis respectivo.
- Comparar los resultados con lo expuesto en la literatura encontrada y ver si se sustenta el conocimiento obtenido.
- Analizar por grupo e identificar aquellas características más relevantes, si clasifica mejor a una clase que otra e inferir conocimiento respecto a ello.

2. Aspectos importantes a considerar

Para obtener los resultados y cumplir los objetivos del laboratorio, se debe tener en cuenta los siguientes puntos:

- Para trabajar con R deben utilizar el paquete Cluster y utilizar la función *pam* u otro paquete a conveniencia.
- Deben decidir y justificar la normalización de las variables.
- Deben justificar el criterio de proximidad a usar con el algoritmo.

3. Informe

El informe se debe regir por el reglamento de titulación v 1.3, apéndice C, apartado C.3 y contener los siguientes puntos:

- Ortografía, redacción y formato **(2 %)**.
- Introducción (máximo 1 plana) **(3 %)**.
- Marco Teórico: Clustering, algoritmo K-means y distancias utilizadas (máximo 2 páginas) **(15 %)**:
- Pre-procesamiento: Se deben definir criterios para eliminar registros o (máximo 12 páginas) columnas que presenten datos perdidos, outliers o que no aporten información relevante para el estudio del problema. En caso de normalizar datos fundamentar la decisión (Máximo 6 páginas) **(15 %)**.
- Obtención del Clúster: Variar parámetros de la función a utilizar de manera que se genere un clúster adecuado en base a las métricas de eficiencia en la clasificación, justifique la utilización de aquellos parámetros, además del criterio para seleccionar el clúster más adecuado. Decida y fundamente el criterio de proximidad de acuerdo a los datos (Máximo 6 páginas) **(10 %)**.
- Análisis de los resultados: Analizar el clúster e identificar aquellas características que sean más interesantes, entregando su significado en el dominio del problema y contrastar esta información con lo expuesto en la literatura **(30 %)**..
- Conclusiones: Respecto a los resultados obtenidos, el desarrollo del laboratorio y el método utilizado. Menciona aspectos positivos y a mejorar en este desarrollo. (Máximo 2 páginas) **(20 %)**.
- Referencias (usar formato APA 6) **(5 %)**

.

4. Observaciones

- Todas las consultas deben ser realizadas al correo javier.arredondo.c@usach.cl
- La entrega debe ser subida al sitio Campus Virtual hasta las 23.55 hrs el 12 de junio del 2020.
- La información de las bases de dato se encuentra en la página `http://archive.ics.uci.edu/ml/`

- Es necesario realizar **TODAS** las experiencias para aprobar el laboratorio
- Cualquier página más allá del máximo permitido no será revisada.
- Una entrega atrasada será evaluada con dos puntos menos y un punto menos por cada día extra de retraso.
- La detección de copia será evaluada con la nota mínima