Pontificia Universidad Católica de Chile
Departamento de Ciencia de la Computación
IIC2433: Minería de Datos

Control 6 - IIC2433

Jueves 26 de Noviembre de 2020
Clustering

Indicaciones

- El control es **individual**. La copia será castigada con nota 1.1 al curso, además de las sanciones disciplinarias correspondientes.
- El control tiene cuatro preguntas, para obtener el 7 deben contestarlas todas de forma correcta. Cada una tiene su propia ponderación.
- Para cada pregunta, si esta contiene más de una subpregunta, **debe contestar cada una de las subpreguntas para obtener el puntaje total en la pregunta.**
- La entrega de este control se realizará a través de un buzón en Canvas, que permanecerá abierto hasta las 23:59 del día Jueves 26 de Noviembre. **No se permitirán entregas atrasadas**.
- El control debe ser entregado en formato **PDF**. En caso de entregar con otro formato, este no se corregirá.

Preguntas Algoritmos Clustering (2.0 pts)

Para esta sección consideraremos el siguiente conjunto de datos. Son productos de los cuales se tiene un par de mediciones pero se desconoce su tipo

ID	Atributo 1	Atributo 2
Producto 1	1	1
Producto 2	1	2
Producto 3	4	5
Producto 4	6	5
Producto 5	0	2

1. (0.8 pts.) Para estos datos, ejecute el algoritmo K-means, con K = 3 y medias iniciales (0,0) (3,3) (4,5). Vaya describiendo y calculando el paso a paso del algoritmo, puede apoyarse en dibujos si es necesario. Finalmente entregue los conjuntos de datos pertenecientes a cada cluster.
2. (0.8 pts.) Sobre el mismo conjunto de datos, ejecute el algoritmo mean shift con radio 1 (distancia euclidiana). Vaya describiendo y calculando el paso a paso del algoritmo, puede apoyarse en dibujos si es necesario. Finalmente entregue los conjuntos de datos pertenecientes a cada cluster.

3. (0.4 pts.) ¿Cuál algoritmo diría que dio los mejores resultados? Justifique su respuesta. Dentro de los dos algoritmos utilizados, utilizaría alguno de ellos con otra configuración para obtener mejores resultados. Justifique su respuesta, si es que hay una mejor versión, diga el algoritmo y su configuración. Si es que cree que no, justifique por qué no.

Preguntas Gaussian Mixture Models (1.5 pts)

1. (0.9 pts.) Que beneficios puede tener realizar clustering con un modelo probabilístico como es el GMM. Mencione y justifique al menos 3. ¿Siempre es recomendable utilizar este modelo en vez de K-means? Justifique su respuesta.

2. (0.6 pts.) Al igual que K-means, GMM puede terminar con diferentes clusters para la misma configuración para diferentes ejecuciones. Dibuje un dataset en donde pueda ocurrir que el algoritmo para un k = 3, tenga dos posibles clusters. Dibuje las gaussianas encontradas sobre los datos en las dos ejecuciones del algoritmo.

Preguntas validación de clustering (2.5 pts)

1. (0.9 pts) ¿Qué problemas pueden ocurrir si no realizamos validación de nuestros resultados de clustering? Mencione y explique al menos 3 problemas.

2. (0.8 pts) ¿Tener un alto SSE implica tener un buen resultado de clustering? Justifique su respuesta. Explique el procedimiento para poder encontrar buenos clusters utilizando el SSE y distintas configuraciones del algoritmo utilizado.

3. (0.8 pts) Dibuje y justifique un ejemplo donde se esté cometiendo un error al utilizar clustering. ¿Siempre existirán patrones en los conjuntos de datos? Justifique su respuesta. De no existir patrones, ¿qué se podría hacer sobre el conjunto de datos?

Pregunta meme (0.3 pts) BONUS

- Se entregará 1 décima si sube un meme con la temática de cualquier contenido del curso. Este meme no puede ser uno subido anteriormente por el profesor o los ayudantes.

- Se entregarán 2 décimas adicionales **a criterio del cuerpo docente** para los mejores memes. Este punto no es posible recorregir.