INTELIGENCIA DE NEGOCIOS
Dr. Mario Garza Fabre
Mayo – Agosto 2018

PROYECTO UNIDAD 3: Clustering

Las técnicas de clustering (agrupamiento) representan una de las herramientas fundamentales para el análisis de datos, siendo de indiscutible relevancia para la solución de problemas prácticos en diferentes ámbitos. El objetivo de este proyecto es que el alumno se familiarice con el tema de clustering mediante la aplicación y el análisis comparativo de diferentes algoritmos de clustering en diversos conjuntos de datos de prueba.

ALGORITMOS A COMPARAR

Se comparará el desempeño de los siguientes cuatro algoritmos de clustering:

1. *k*-Means
2. Clustering jerárquico aglomerativo - distancia promedio (average linkage)
3. Clustering jerárquico aglomerativo - distancia mínima (single linkage)
4. Clustering jerárquico aglomerativo - distancia máxima (complete linkage)

DATOS DE PRUEBA

La siguiente tabla muestra los 25 conjuntos de datos de prueba a utilizar, describiendo su tamaño (N), dimensionalidad (D) y número de clusters (K).

#	Dataset	N	D	K
1	Aggregation	788	2	7
2	Compound	399	2	6
3	D31	3100	2	31
4	Flame	240	2	2

5	Glass	214	9	6
6	Iris	150	4	3
7	Jain	373	2	2
8	Long3	1000	2	2
9	Pathbased	300	2	3
10	R15	600	2	15
11	S4	5000	2	15
12	Smile	400	2	4
13	Spiral2	1000	2	2
14	Spiral3	312	2	3
15	Tevc1	1477	2	4
16	Tevc2	882	2	4
17	Triangle	1000	2	4
18	UKC1	29463	2	11
19	UKC2	26739	2	10
20	UKC3	31929	2	11
21	UKC4	29149	2	10
22	UKC5	30688	2	11
23	UKC6	31191	2	11
24	UKC7	31666	2	11
25	UKC8	34654	2	12

EXPERIMENTOS

- Se aplicará cada uno de los 4 algoritmos a los 25 problemas de prueba.

- Independientemente de la naturaleza determinista o no determinista de los algoritmos de clustering a utilizar, se realizará una única ejecución de cada uno de ellos (utilizando parámetros por default).

- El resultado (clustering) obtenido por cada algoritmo será evaluado con respecto al agrupamiento real de los datos (ya conocido para los 25 conjuntos de datos de prueba). Para ésto, se utilizará la métrica llamada ***Adjusted Rand Index*** (ARI). ARI se encuentra ya disponible en la mayoría de las bibliotecas para minería de datos/aprendizaje máquina (por ejemplo, adjusted_rand_score es la implementación de ARI en Scikit-Learn, Python).

- Adicionalmente, los resultados obtenidos por los diferentes métodos serán evaluados de manera gráfica para todos los problemas de prueba bi-dimensionales, es decir, donde **D = 2**.

- Se recomienda **normalizar los datos** antes de aplicar los algoritmos de clustering. Por ejemplo, en Python bastará con transformar los datos utilizando: **datos = StandardScaler().fit_transform(datos)**.

REPORTE DE RESULTADOS

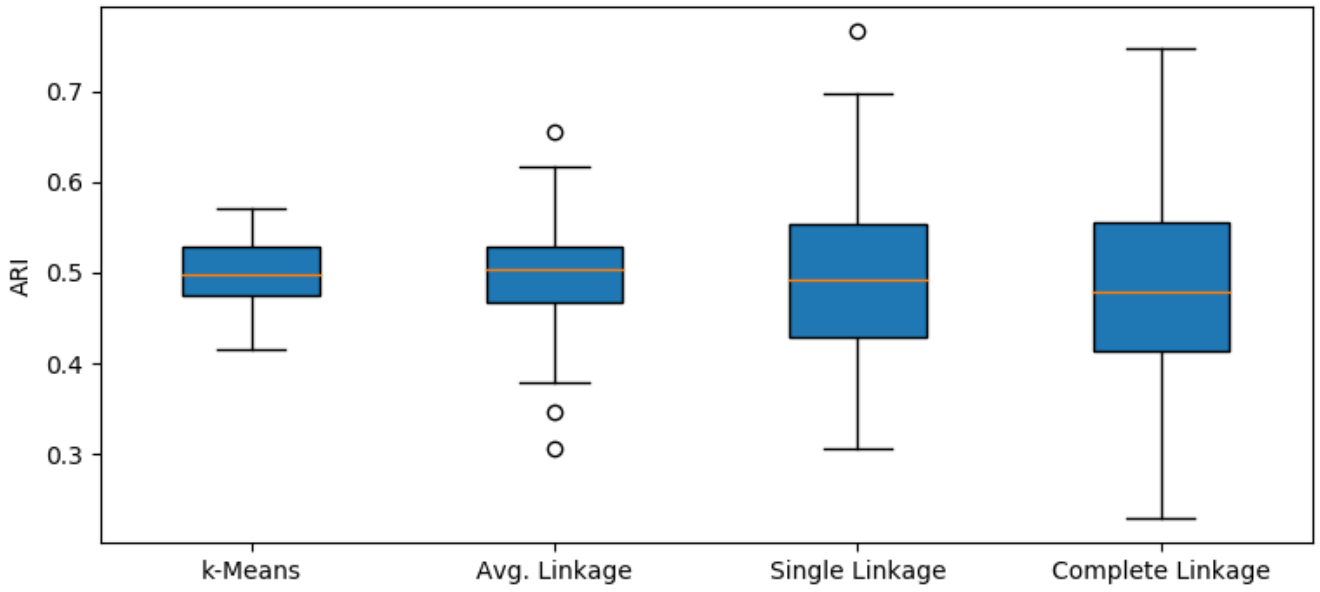
Se entregará un reporte breve con el siguiente contenido:

1. **Portada.** Importante incluir los nombres de los integrantes del equipo de trabajo.

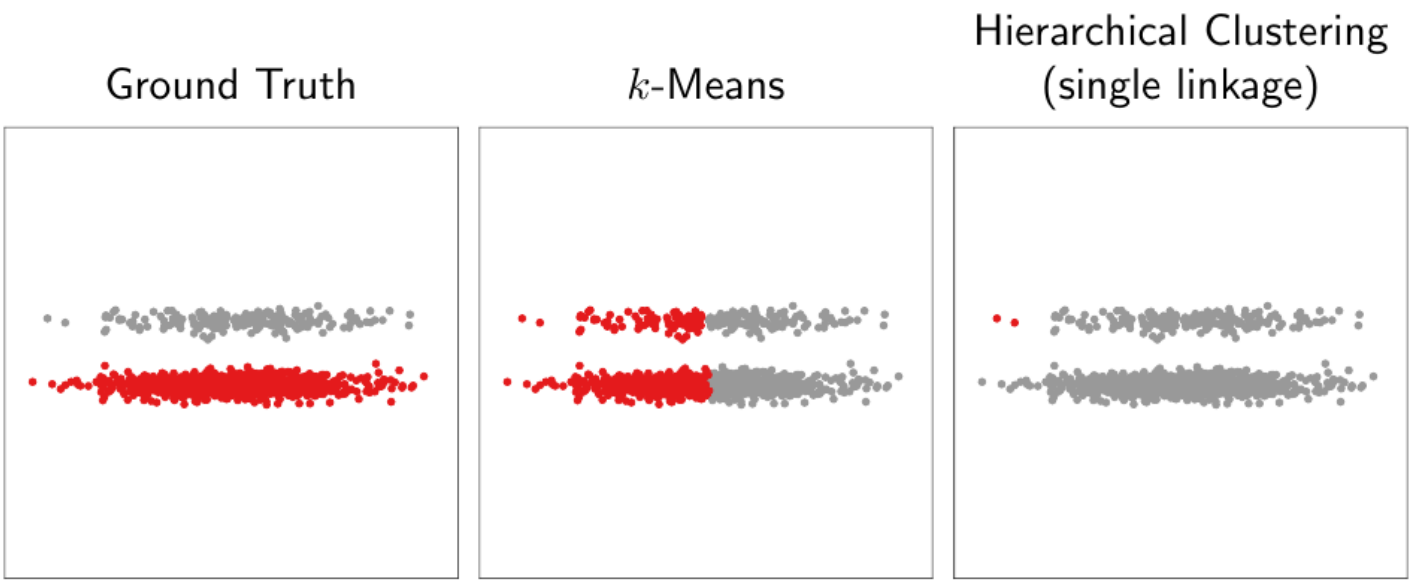
2. **Tabla de resultados.** Para cada conjunto de datos de prueba, se mostrará el ARI calculado para los resultados obtenidos por los diferentes algoritmos evaluados. Para cada problema, sombrear la celda y/o resaltar el texto en negrillas para el mejor resultado obtenido (valor más alto de ARI).

Dataset	k-Means	Avg. Linkage	Single Linkage	Complete Linkage
Aggregation	**0.8**	0.7	**0.8**	0.75
Compound				
...				
UKC8				

3. **Gráfica de estadísticas.** Generar una gráfica de cajas (boxplot) para ilustrar el desempeño general de cada método en los diferentes conjuntos de datos de prueba. Para generar esta gráfica, **NO** será necesario normalizar los resultados, ya que ARI ya se encuentra definido en el rango [0, 1].



4. **Visualización de resultados.** Para cada problema bi-dimensional (D = 2), incluir un diagrama/gráfica de dispersión donde se ilustren los resultados obtenidos por los diferentes algoritmos. Ilustrar también el agrupamiento real de los datos. Utilice diferentes colores para resaltar los diferentes clusters.



EVALUACIÓN Y ENTREGABLES DEL PROYECTO

- **IMPORTANTE:** Eclass será el único medio a través del cual se podrán enviar los proyectos. Sin excepción, **cualquier proyecto enviado por correo electrónico será ignorado por el instructor**.
- Cada equipo de trabajo deberá entregar un único archivo (.zip o tar.gz) incluyendo código fuente y un documento o archivo de texto donde se proporcionen las instrucciones de compilación (si aplica) y funcionamiento del programa, así como el reporte de resultados.
- Utilice el lenguaje de programación y/o herramientas de su elección.
- El 100 % del valor de este proyecto será distribuido y evaluado de la siguiente manera:

Utilización de algoritmos solicitados	50 %
Experimentación y reporte de resultados	50 %
Total	**100 %**

FECHA Y HORA LÍMITE DE ENTREGA

La fecha y hora límite de entrega será: **Miércoles 4 de julio de 2018, 23:55 horas.**

NOTA: La puntuación máxima de un proyecto se reducirá **10 %** por cada día hábil de retraso en su entrega.