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

PROYECTO UNIDAD 4: Clustering – Parte II

El proyecto anterior consideró problemas sintéticos para los que ya se conoce la solución correcta y el número de clusters a identificar. Aunque tal escenario es útil para la evaluación objetiva durante el desarrollo de nuevos algoritmos de clustering, en la práctica generalmente no disponemos de este tipo de información. Este proyecto tiene como objetivo que el alumno profundice sus conocimientos en el área de clustering de datos, considerando ahora problemas para los que se ignora cuál es la solución correcta y el número de clusters.

CONJUNTOS DE DATOS

Utilizando los mismos datos considerados para el proyecto de la Unidad 2, el alumno construirá diversos conjuntos de datos basados en diferentes atributos que describen a las 42 fuerzas policiacas del Reino Unido. Todos los conjuntos de datos tendrán, por lo tanto, un total de N=42 observaciones, pero la dimensionalidad (número de atributos) del problema podrá cambiar de acuerdo con las siguientes especificaciones:

#	Conjunto de datos	Atributos	% Eval.
1	**Ubicación de las fuerzas policiacas.** Utilizar el centroide de las coordenadas (latitud y longitud) para representar la ubicación de cada fuerza policiaca. Utilizar archivos KML.	2	20
2	**Total de crímenes y área de la fuerza policiaca.** Número total de crímenes reportados (2011-2017) por cada fuerza y tamaño de su área de cobertura. Utilizar archivos KML.	2	20
3	**Ubicación y área de las fuerzas policiacas.** Coordenadas de los centroides y tamaño del área de cobertura de cada fuerza policiaca. Utilizar archivos KML.	3	20
4	**Total de reportes para cada tipo de crimen.** Número total de reportes (2011-2017), separado para cada tipo de crimen.	16	13.333

5	**Total de reportes anuales.** Número total anual de crímenes en los 7 años comprendidos desde 2011 hasta 2017.	7	13.333
6	**Total de reportes mensuales.** Número total mensual de crímenes en los 84 meses comprendidos desde enero de 2011 hasta diciembre de 2017.	84	13.333
		TOTAL:	**100**

EXPERIMENTOS

- Se comparará el desempeño de **cuatro algoritmos** al resolver cada uno de los problemas propuestos:
  - *k*-Means
  - Clustering jerárquico aglomerativo - distancia promedio (average linkage)
  - Clustering jerárquico aglomerativo - distancia mínima (single linkage)
  - Clustering jerárquico aglomerativo - distancia máxima (complete linkage)

- Dado que no conocemos el número de clusters (k) que presenta cada conjunto de datos, se considerarán **TODOS los valores de k en el rango [2, 41]**.

- 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 para cada problema y valor de k.

- El resultado (clustering) obtenido por cada algoritmo será evaluado utilizando un índice (criterio de clustering) interno: **Mean Silhouette Coefficient**. Esta métrica se encuentra ya disponible en la mayoría de las bibliotecas para minería de datos/aprendizaje máquina (por ejemplo, **silhouette_score** es la implementación disponible en Scikit-Learn de Python). Esta medida recibe como entrada el conjunto de datos y la membresía de cada observación a un determinado cluster (etiquetas obtenidas por un algoritmo de clustering). La medida calcula la calidad del clustering y regresa un valor en el rango [-1, 1], donde 1 corresponde al mejor resultado posible y -1 al peor.

- Para cada problema, los resultados obtenidos se reportarán en una **gráfica de curvas.** Cada curva representará un algoritmo diferente y mostrará los resultados obtenidos para cada valor de k considerado (ver ejemplo proporcionado en la descripción del reporte).

- Adicionalmente, para los **problemas con 2 o 3 atributos** se incluirá una **gráfica de dispersión** ilustrando el mejor resultado obtenido por cada algoritmo (número de clusters k para el que se obtuvo el valor más alto para la medida de evaluación utilizada).

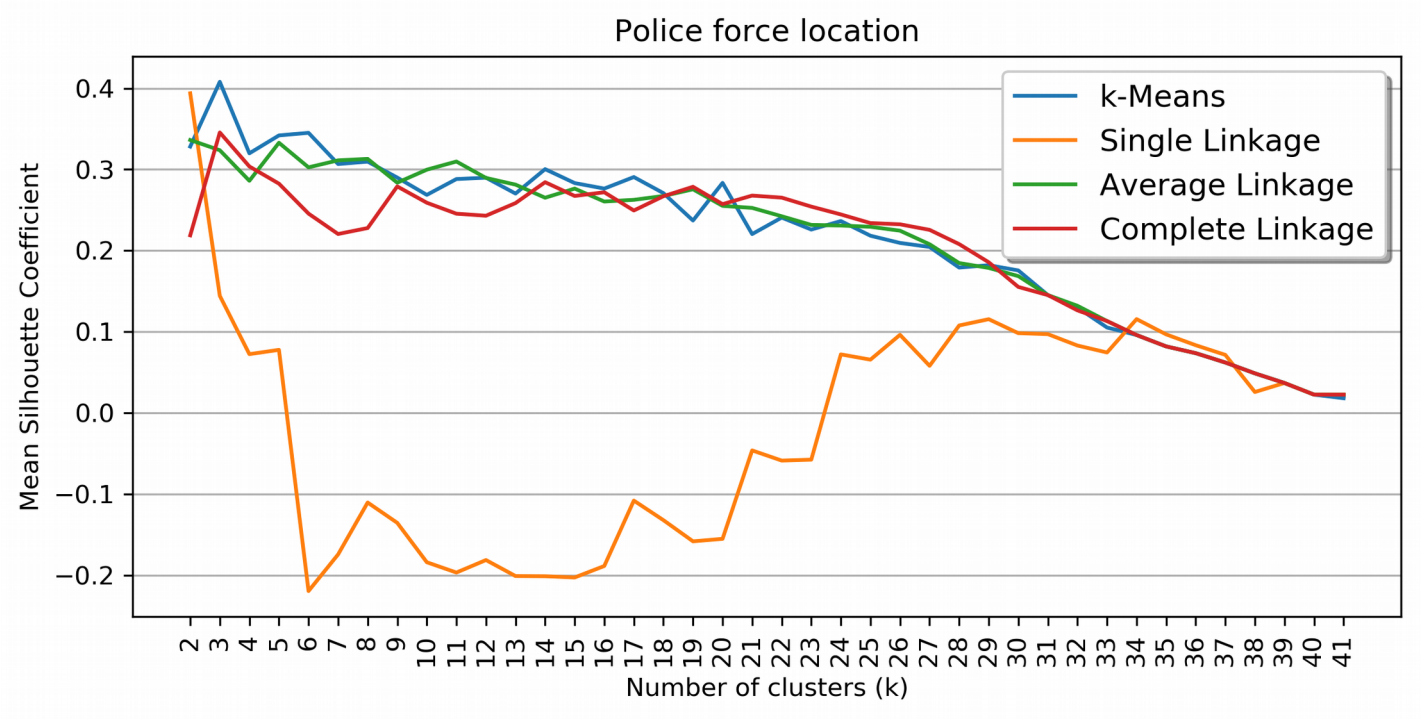
- Se recomienda **normalizar los datos** antes de aplicar los algoritmos de clustering. Por ejemplo, en Python bastará con transformar los datos utilizando: **datos = StandardScaler().fit_transform(datos)**.

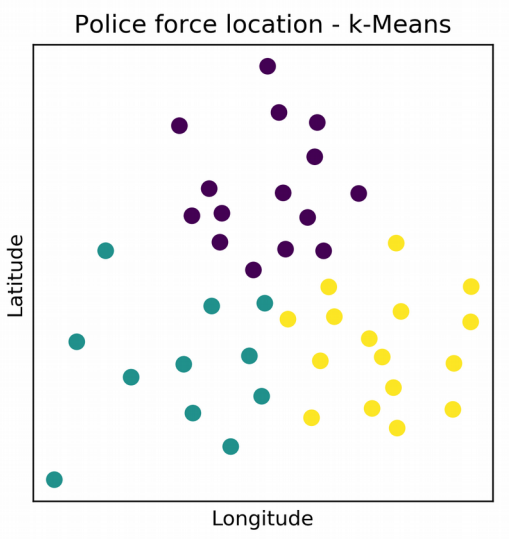
REPORTE DE RESULTADOS

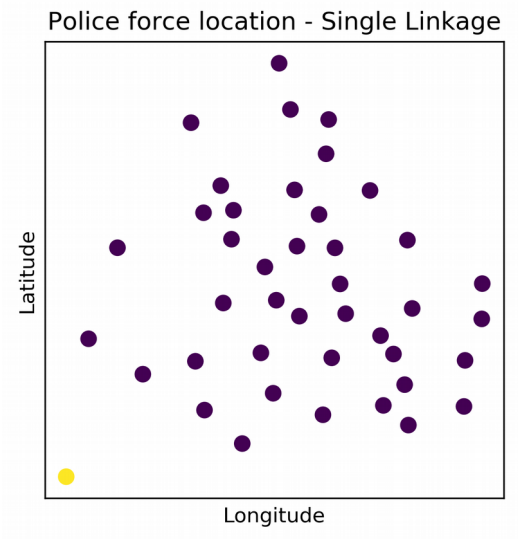
Se entregará un reporte breve con el siguiente contenido:

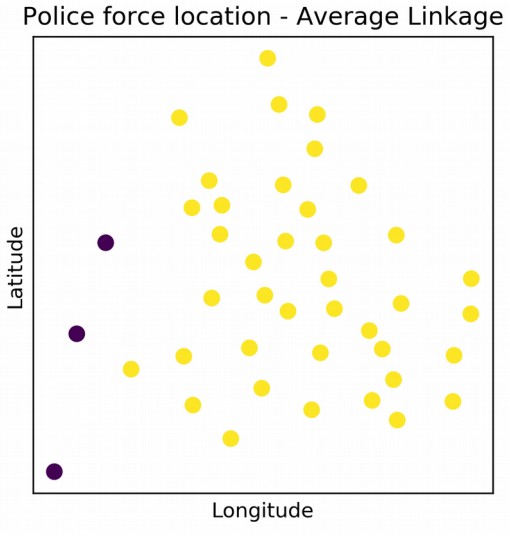
1. **Portada.** Importante incluir los nombres de los integrantes del equipo de trabajo.

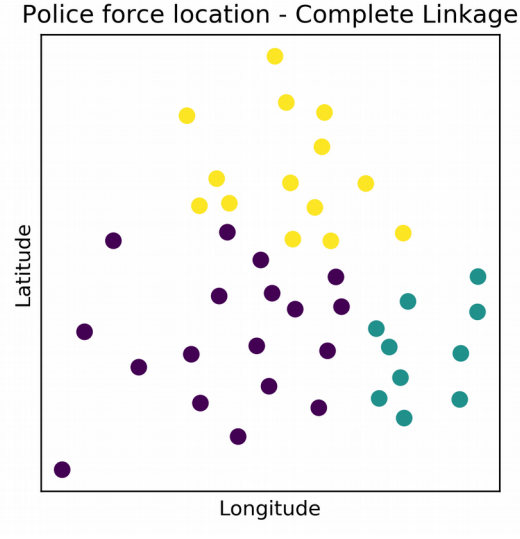
2. **Visualización de resultados.** Para cada problema se presentará la gráfica de curvas mostrando los resultados obtenidos. Para problemas con 2 o 3 atributos, se incluirán también gráficas de dispersión ilustrando los mejores resultados obtenidos por cada uno de ellos. Ver ejemplos a continuación.


Police force location
Mean Silhouette Coefficient
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
k-Means
Single Linkage
Average Linkage
Complete Linkage
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
Number of clusters (k)


Police force location - k-Means
Latitude
Longitude


Police force location - Single Linkage
Latitude
Longitude


Police force location - Average Linkage
Latitude
Longitude


Police force location - Complete Linkage
Latitude
Longitude

EVALUACIÓN Y ENTREGABLES DEL PROYECTO

- Cada equipo de trabajo entregará un único archivo (.zip o tar.gz) incluyendo código fuente y reporte.
- Utilice el lenguaje de programación y/o herramientas de su elección.
- El 100 % del valor de este proyecto será distribuido como se indica en la tabla de especificaciones para los conjuntos de datos.
- Se tomará en cuenta la calidad de las gráficas generadas para el reporte de resultados. Las gráficas deberán ser suficientemente descriptivas y utilizar colores y tamaños/tipos de fuente adecuados.

FECHA Y HORA LÍMITE DE ENTREGA

La fecha y hora límite de entrega será: **Viernes 13 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.