Mentoría: Coronavirus en Argentina y el mundo
Práctico 5: Aprendizaje no supervisado.

Matías A. Bettera Marcat Lucas C. Didoné Daniel Peralta

Diciembre 17, 2020

1 Práctico Aprendizaje no supervisado

Objetivos:
Realizar un k-means clustering para clasificar los países (o regiones) según alguna(s) variable(s) de interés.

Consigna:
En la notebook tienen un ejemplo de k-means clustering para series de tiempo utilizando la variable casos_nuevos de cada provincia. Revisen el modelo propuesto y luego discutan/ debatan/ respondan/ propongan:

¿Qué número de clusters k les parece más conveniente utilizar y por qué?
Correr el modelo con ese valor de k.
Interpretar los resultados (clusters) obtenidos, ¿tienen sentido con lo que sabemos de la realidad o lo que nos dice la intuición al revisar los datos?
Probar el modelo para distintos valores de:
la variable (casos_total, casos_nuevos, muertes_total y muertes_nuevos)
la distancia (DTW, soft-DTW, ...)
Tener en cuenta en cada caso cuál sería el k óptimo. Quizás sea conveniente tener las variables por cada millón de habitantes en cada provincia (quizás el modelo mejora, prueben!!). La idea es encontrar el modelo que produzca la mejor agrupación posible de provincias.

¿Qué otros modelos de aprendizaje no supervisado se podrían utilizar con este tipo de datos? Si se animan propongan otro y si se animan aún más lo prueban (si llegan, sino todo bien, la idea es aunque sea pensarlo y debatirlo).

2 Introducción

Para la implementación de algoritmos de Aprendizaje no supervisado, en este práctico se trabajó con el dataset publicado por el Ministerio de Salud de Argentina:

http://datos.salud.gob.ar/dataset/covid-19-casos-registrados-en-la-republica-argentina/archivo/fd657d02-a33a-498b-a91b-2ef1a68b8d16

3 Limpieza y Feature Engineering

Antes de comenzar a trabajar en la implementación del algoritmo, fue necesario aplicar ingeniería de características en las variables "País" y "Provincia":

- **País:** sólo se consideraron casos cuyo país de origen es Argentina.
- **Provincia:** se excluyeron del dataset aquellos casos donde la provincia de residencia figuraba como "SIN ESPECIFICAR".

4 Clustering

Pre-procesamiento del dataset

A fines de realizar un análisis de clustering por regiones, trabajamos en el agrupamiento previo de los datos basándonos en cuatro variables de interés:

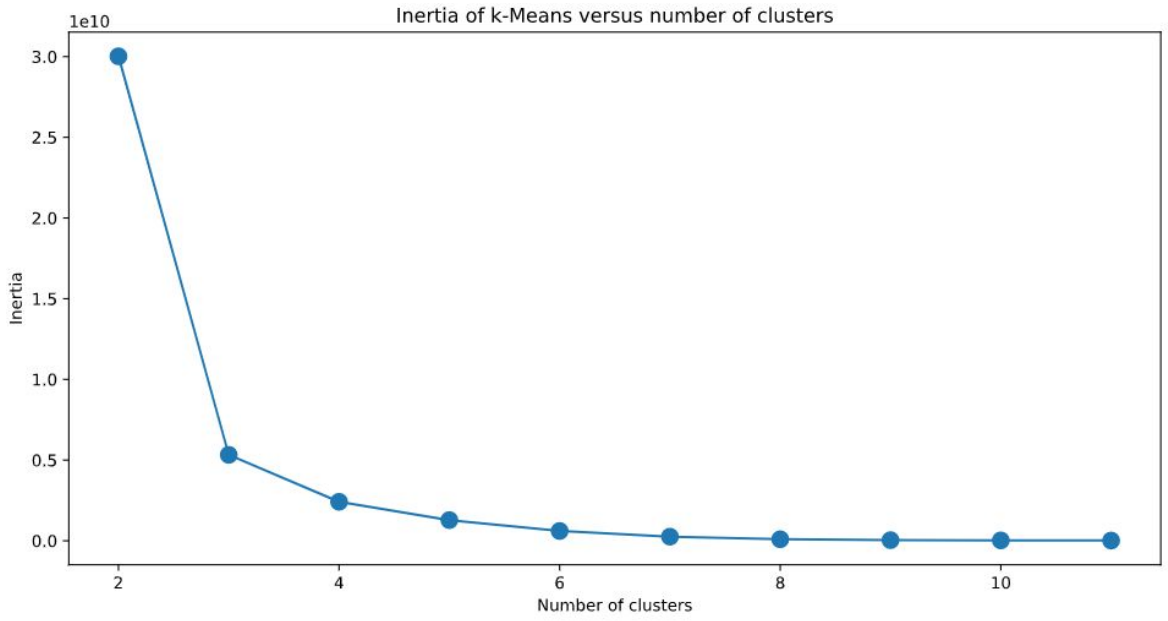
- Casos confirmados
- Fallecidos
- Cantidad de casos en cuidados intensivos
- Cantidad de casos con asistencia respiratoria mecánica

Para realizar este agrupamiento, utilizamos 4 sub-dataset (uno por cada variable) y luego lo unificamos a través de una operación "merge". De esta forma, el dataset final quedó con la siguiente estructura:

	residencia_provincia_nombre	confirmados	fallecidos	edad_promedio	cuidado_intensivo	asistencia_respiratoria_mecanica
0	Buenos Aires	535333	20381	71.615468	13948	5405
1	CABA	143690	6045	76.649254	5178	2169
2	Catamarca	924	1	97.000000	35	10
3	Chaco	12743	560	65.517857	1004	443
4	Chubut	13972	202	70.282178	177	79

Análisis del número de clusters "k"

Para buscar el valor más conveniente del número de clusters utilizamos el método del codo. Tal como se puede observar en la siguiente figura, el valor "k" mas conveniente es "3":



Implementación de K-Means

Para la implementación del algoritmo K-means utilizamos el siguiente array de variables:

["confirmados", "fallecidos", "edad_promedio", "cuidado_intensivo", "asistencia_respiratoria_mecanica"]

Al ejecutar K-Means obtuvimos como resultado 3 clusters agrupados de la siguiente forma:

	confirmados	fallecidos	edad_promedio	cuidado_intensivo	asistencia_respiratoria_mecanica
cluster					
1	13405.3	346.2	68.9	271.8	126.4
2	535333.0	20381.0	71.6	13948.0	5405.0
3	110622.0	2970.3	74.2	2915.0	1133.0

Luego, visualizando este agrupamiento por provincias a través de un Treemap se muestra de la siguiente forma:

Si bien se observa en principio que los clusters estarían definidos por la cantidad de casos, podríamos interpretar alguna relación con la ubicación geográfica de las provincias y la circulación entre las mismas. Esto puede verse más que todo en el cluster "3" (CABA-Santa Fe-Córdoba).
Siguiendo con este enfoque, podríamos considerar a futuro la posibilidad de incorporar a este set de datos información relacionada a la densidad poblacional de cada provincia.

Otros algoritmos que podrían implementarse

Si bien en este caso nosotros realizamos una implementación de K-Means trabajando con un agrupamiento previo de los datos, el enfoque propuesto por Lucía Pappaterra de trabajar con una adaptación de análisis de series de tiempos resulta muy interesante.
Por otro lado, en forma alternativa podrían realizarse pruebas de implementaciones con Mean Shift y DBScan sobre el mismo set de datos utilizado en este caso.