EL4106 Inteligencia Computacional
Tarea 4

Profesores: Pablo Estévez y Pablo Huijse
Auxiliar: Ignacio Reyes
Ayudante: Rodrigo Carrasco
Semestre: Primavera 2017

La base de datos *world.p* contiene información de 57 países referente a su población, calidad de vida, desarrollo económico y tecnológico (fuente: Banco Mundial www.worldbank.org). En este archivo cada una de las filas corresponde a uno de los países incluidos y las columnas a una característica de este (21 características). El detalle de los atributos se encuentra en el Anexo de esta tarea. Se pide realizar un análisis de clustering de estos países de modo de encontrar grupos de similitud entre ellos. Para cada una de las actividades a realizar se han de considerar los siguientes criterios de agrupación:

a) Según características geográficas y poblacionales. Utilice las columnas 1, 2, 3 y 4.

b) Según su desarrollo económico. Utilice las columnas 5, 6, 7 y 8.

c) Según su desarrollo tecnológico. Utilice las columnas 9, 10, 11, 12, 13 y 14.

d) Según calidad de vida y educación. Utilice las columnas 15, 16, 17, 19, 20 y 21.

e) Global. Utilice toda la información (columnas 1 a 21) y realice una proyección en componentes principales que retenga al menos un 90 % de la varianza de los datos. Entrene el mapa auto-organizativo usando los datos proyectados.

A continuación se especifica lo que usted deberá desarrollar en esta tarea:

1. Responda ¿Qué significa que un algoritmo de aprendizaje de máquinas sea no-supervisado? Describa brevemente el algoritmo SOM y explique como se interpreta la visualización mediante matriz-U.
2. Para cada uno de los criterios de agrupación mostrados anteriormente:
   i) Construya un mapa auto-organizativo de Kohonen (SOM), visualice los resultados usando la matriz-U de distancias y los mapas de prototipos para cada característica.
   ii) Identifique y describa los grupos de similitud o clusters entre los países. Estime el número de clusters. ¿Se observan outliers? Complemente su análisis con la información obtenida de la proyección en componentes principales y el clustering de k-medianas (usando el número estimado de clusters).
   iii) Comente sobre la relación particular de Chile con respecto a otros paises (a) de América del Sur y (b) de todo el mundo.

IMPORTANTE: El día Jueves 26 de Octubre en horario de cátedra se realizará una sesión de laboratorio en la sala de computación del segundo piso del DIE donde se desarrollará la tarea. La asistencia es OBLIGATORIA. Podrán trabajar en grupos de dos personas. No olvide poner los nombres de ambos integrantes en el informe. Se recomienda PUNTUALIDAD.

IPYTHON NOTEBOOK: Para iniciar un servidor de *ipython notebook*, abra una consola de anaconda (Menú de inicio, anaconda, open anaconda shell) y escriba *ipython notebook*. Se debería abrir un browser en la dirección http://127.0.0.1:8888/. Una vez en la interfaz del *ipython notebook* utilice el explorador de carpetas y navegue hasta encontrar el archivo ipynb adjunto o utilice el botón *upload*. Para ejecutar cada bloque de código presione SHIFT+ENTER.

Anexo

El nombre de las características incluidas en la base de datos:

Característica	Nombre
1	Superficie (km^2)
2	Tasa de crecimiento promedio anual de la población
3	Densidad poblacional (personas por km^2)
4	Población total
5	PIB ($ US)
6	PIB per cápita ($ US)
7	Desempleo total (% fuerza laboral total)
8	Índice de Gini[1]
9	Suscritos a planes de banda ancha (por 100 habitantes)
10	Suscritos a planes de telefonía móvil (por 100 habs)
11	Exportación de alta tecnología (% de todas las exportaciones)
12	Investigadores (por millón de habitantes)
13	Consumo eléctrico (kWh per cápita)
14	Emisiones de CO2 (tonelada métrica per cápita)
15	Gasto Total en Salud (% del PIB)
16	Esperanza de Vida (años)
17	Tasa de mortalidad adulta femenina (por cada 1000 mujeres)
18	Tasa de mortalidad adulta masculina (por cada 1000 hombres)
19	Alumnos matriculados, educación primaria[2]
20	Alumnos matriculados, educación secundaria (ver nota del ítem anterior)
21	Alumnos matriculados, educación terciaria (ver nota del ítem anterior)

[1] El índice de Gini mide cuan distinta es la distribución del ingreso entre individuos en comparación a una distribución perfectamente equitativa. Un índice cero corresponde a equidad perfecta, mientras que un índice de 100 indica máxima inequidad.

[2] Se mide como porcentaje respecto al total de individuos en el rango de edad correspondiente (primaria, secundaria, terciaria).