PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
FACULTAD DE CIENCIAS E INGENIERÍA

APLICACIONES DE CIENCIAS DE LA COMPUTACIÓN

4.ª práctica (tipo b)

(Primer semestre 2015)

Evaluación Individual (12 ptos):

Limpieza de datos

1. **Normalización (2 puntos):** descargar el archivo "diabetes_Test_268.txt" y normalizar cada atributo (columna) entre [-1, 1] (Ver Eq. 1) . Guardar los atributos normalizados, separados por comas, en el archivo "diabetes_norm_268.txt".

$$x'_i = -1 + 2\left[\frac{x_i - min(x)}{max(x) - min(x)}\right] \quad \textbf{(Eq. 1)}$$

donde:

x_i = valor a normalizar
min(x) = valor mínimo para ese atributo (columna)
max(x) = valor máximo para ese atributo (columna)
x'_i = valor normalizado

Clustering: K-means

2. Implementar dos nuevas medidas de distancia:

   a. **Distancia Manhattan (3 puntos):** la distancia entre dos puntos es la suma de las diferencias (absolutas) de sus coordenadas. También se conoce como distancia rectilínea, distancia L_1, distancia Manhattan. *El último nombre alude al diseño de grilla de la mayoría de las calles de la isla de Manhattan, lo que causa que el camino más corto que un auto puede tomar entre dos puntos de la ciudad tengan la misma distancia que dos puntos en geometría Taxicab.*

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|,$$

donde:

$\mathbf{p} = (p_1, p_2, \ldots, p_n)$ y $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ son vectores.

b. **Distancia Canberra (3 puntos):** es una medida de distancia entre pares de puntos en un espacio vectorial. Es una versión ponderada de la distancia L_1 (Manhattan).

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|},$$

donde:

$\mathbf{p} = (p_1, p_2, \ldots, p_n)$ y $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ son vectores.

Pruebas (4 puntos)

3. Aplicar el algoritmo K-Means, para K=2, usando como archivo de prueba el archivo "jain-2dim.txt" (ver **Fig. 1**), utilizando las tres medidas de distancia implementadas anteriormente (Euclidiana, Manhattan y Canberra). Guardar los resultados obtenidos en un archivo .csv ("2_output_euclidiana.csv"/ "2_output_manhattan.csv"/ "2_output_canberra.csv"), siguiendo el formato de archivo de salida pedido para la evaluación grupal.

Recordando …

- El resultado de la clusterización debe imprimirse en un archivo de salida CSV: "**k_output.CSV"**, donde "**k**" es el número de clusters que se solicitó generar.
- El formato de salida del archivo CSV debe ser el siguiente:

```
x1,y1,c1
x2,y2,c2
…
```

donde:

xi, yi: atributos

ci: cluster al que pertenece (con valores desde **1** hasta "**k**")

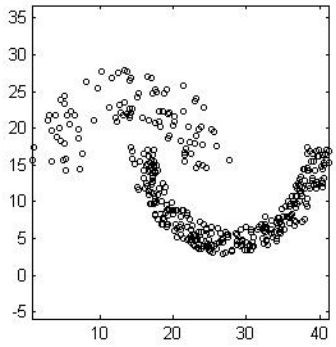

Figura 1- A.K. Jain's Toy Problem