Tarea 1

ENTREGA: Miércoles 17 de Abril 2019 hasta las 23:55 hrs vía `moodle`.

Instrucciones

- Puede realizar la tarea de forma individual o en grupos de hasta 3 personas.
- Debe adjuntar todo el código fuente utilizado.
- Puede programar en cualquier lenguaje y utilizar las librerías que guste.
- Subir un informe en formato pdf por grupo con los nombres y resultados obtenidos.
- Si el lenguaje lo permite, se acepta un Jupyter Notebook que incluya el código y el informe.

[80 %] Parte I

El objetivo de la primera parte es probar los métodos de clustering vistos en clases sobre tres conjuntos de datos diferentes y analizar que algoritmo se comporta mejor y peor en cada caso.

0.1. Algoritmos

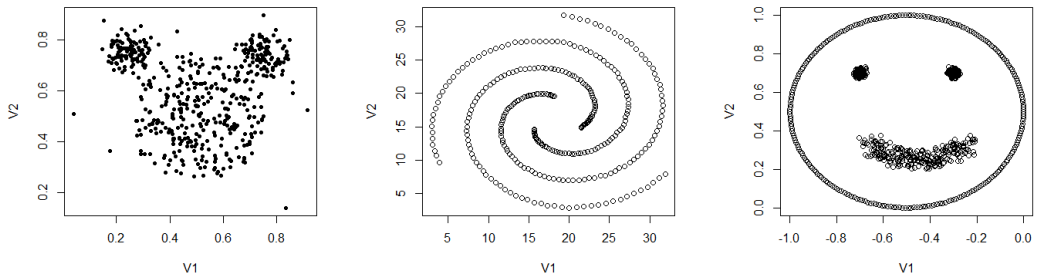
Se estudiarán los siguientes algoritmos:

- K-Means
- Agglomerative Hierarchical Clustering:
  - Single Link
  - Complete Link
- DBSCAN
- Mean-shift
- Spectral clustering
- Fuzzy C-Means

Realice ajuste de parámetros hasta lograr una solución óptima (aceptable) para cada algoritmo. Muestre una visualización (plot) en 2 dimensiones del clustering obtenido en cada dataset. Explique los resultados, señalando los parámetros ideales de los algoritmos. ¿Qué método obtuvo la mejor y peor agrupación según su criterio? Justifique y comente la razón.

0.2. Datasets

Se analizarán tres conjuntos de datos diferentes:



(a) mouse (b) spiral (c) smile

Figura 1: Datasets de la **Parte I**

[20 %] Parte II

Responda las siguientes preguntas:

(a) **[10 %]** Se tiene un conjunto de datos con 100 objetos. Se le pide realizar clustering utilizando K-means, pero para todos los valores de k, $1 \leq k \leq 100$, el algoritmo retorna que todos los clusters están vacíos, excepto uno. ¿En qué situación podría ocurrir esto? (analice los datos y no los parámetros del algoritmo, i.e., iteraciones). ¿Qué resultado tendría single-link y DBSCAN para este tipo de datos?

(b) **[10 %]** Considerando single-link y complete-link hierarchical clustering, ¿es posible que un objeto esté más cerca (en distancia Euclidiana) de los objetos de otros clusters en relación a los de su propio cluster? Si fuese posible, ¿en qué enfoque (single y/o complete) esto podría ocurrir? Justifique con un ejemplo en cada caso.