Tarea 1 - Beer Clustering
Reconocimiento de Patrones en Minería de Datos

Profesor: Marcelo Mendoza - mmendoza@inf.utfsm.cl
Ayudante: Ignacio Espinoza - ignacio.espinoza@alumnos.inf.utfsm.cl

23 de marzo de 2017

1. Introducción

En esta tarea se trabajará con el dataset beeradvocate[1], el cual consiste en una gran colección de reseñas de cervezas del sitio BeerAdvocate[2]. Dentro de cada una de estas reseñas se encuentra una evaluación en diferentes aspectos de la cerveza, además de información de procedencia, tipo de cerveza, entre otros. Para poder hacer más fácil el manejo de los datos, se ha eliminado la columna *reseña*, reduciendo considerablemente el tamaño, en bytes, del dataset.

Usted deberá implementar y probar cinco algoritmos de clustering vistos en clase, mostrando visualizaciones de los resultados obtenidos por cada uno, en un manifold euclidiano 2D. La tarea consta de los siguientes pasos:

a) Descargue el archivo **beer_reviews.tar.gz**, ubicado en la sección **Entregas**, y cárguelo en Python.

b) Describa brevemente el dataset a utilizar (cantidad de datos, tipo de atributos, etc.) en el Notebook Jupyter/IPython especificado en ."Entregable y consideraciones"

c) Probar los siguientes algoritmos sobre el dataset:

- k-means
- Minibatch k-means
- HAC Complete
- Ward
- DBScan

Muestre una visualización por cada algoritmo. Realice ajustes a los parámetros de cada uno, hasta que logre obtener una buena solución. Justifique el procedimiento y elección de estos parámetros. Comente cada resultado obtenido.

d) ¿Qué atributo, Nombre de cervecería o Tipo de cerveza, describe mejor a los cluster como etiquetas de clase, según los resultados obtenidos previamente? ¿Hay mejores marcas que otras en relación a las evaluaciones obtenidas? ¿Se puede definir algún criterio para determinar el mejor tipo de cerveza? Comente.

[1] http://snap.stanford.edu/data/web-BeerAdvocate.html
[2] www.beeradvocate.com

2. Entregable y consideraciones:

- Los equipos de trabajo pueden ser de 2 a 3 personas.
- Se deberá construir el informe en Jupyter/IPython notebook que explique paso a paso la actividad realizada y las conclusiones del trabajo.
- Deben mantener un respaldo de todo código utilizado e informe en GitHub.
- **Entrega:** envío del link del repositorio de GitHub al correo del ayudante, con copia al profesor, especificando el asunto [Tarea1-DM-2017-1].
- **Entrega: Jueves 13 de Abril, hasta las 23:55 Hrs**. Cualquier correo enviado después de la fecha de entrega tendrá un descuento de **20 puntos** por día de atraso.
- Una vez entregada la tarea **No podrá seguir realizando commits a su repositorio**.