Tarea 3: Aprendizaje no supervisado (Clustering)

F.C | GD M.D - II-2015

Encontrando estructuras

En muchas tareas en el proceso de Minería de Datos encontramos conjuntos de datos en dónde desconocemos la clase de las instancias. El proceso de aprendizaje cuando desconocemos la columna clase se denomina en la literatura **aprendizaje no supervisado**.

En algunos de estos casos, podemos aplicar algoritmos que nos permitan encontrar, de existir, estructuras que denominamos clústers para que, una vez encontrados, podamos analizarlos uno a uno hasta conseguir características de relevancia que nos permitan solucionar el problema planteado originalmente. Este subconjunto de tareas donde podemos encontrar estructuras con alguna relación se denomina **clustering**.

En la presente tarea, se le dará a cada analista 8 datasets en dónde no solo deberá escoger el mejor método de clustering visto en clases (de usar un método visto en clases) sino que, para algunos datasets, deberá solucionar problemas específicos planteados en aras de usar herramientas vistas en clases que le permita encontrar una solución apropiada.

Objetivos

Objetivo Principal

Escoger, en base a los conocimientos adquiridos en clase, el mejor algoritmo de clustering según su criterio para distintos datasets dados por el grupo docente.

Ayuda: Recuerde que debe usar, al menos, kmedias (y derivados) y clustering jerárquico (con distintas distancias).

Objetivos específicos

1. Usar métodos exploratorios que corroboren su decisión y de tal manera que la expliquen de mejor manera al usuario final.
2. Elaborar soluciones a problemas específicos de cada dataset usando conocimientos adquiridos en teoría y laboratorio.
3. Aprender a usar herramientas de clustering en R o en python.

Sobre la última columna de cada dataset

Salvo el caso de **guess.csv**, asuma que la **última columna** de cada dataset hace referencia a la clase de cada instancia para que pueda verificar los métodos (hacer la matriz de confusión).

En algunos datasets, los valores son números reales en rangos, por ejemplo, [-10, 10]. En estos casos, haga reglas para asignar clases a la hora de comparar el rendimiento de cada método. Un ejemplo de una regla puede ser:

```
# Ejemplo de regla para asignar clases

definir_clase = function(numero){
  # Si quisiera 5 clusters entonces selecciono 4 cortes.
  # Recuerde que, en el caso de R, cada número es relativo a un color.

  if(numero < -4.0)
        return(1)
    else if(numero < -2.0)
        return(2)
    else if(numero < 0.0)
        return(3)
    else if(numero < 4.0)
        return(4)
    else
        return(5)
  }
```

Hay datasets cuyo valor de última columna son enteros que están escritos como reales y no necesitan reglas para definir las clases. Estos datasets son:

1. a.csv
2. a_big.csv
3. good_luck.csv
4. moon.csv

Al hacer el informe, explique de manera concisa el porqué de las reglas usadas para asignar las clases en el caso de los archivos:

1. h.csv
2. help.csv
3. s.csv

Consideraciones especiales para distintos datasets

Tenemos tareas adicionales para algunos datasets más allá de seleccionar el mejor método de clustering.

Para guess.csv

El dataset **guess.csv** es el único sin la última columna haciendo referencia a la clase de la instancia.

1. Implemente el método Codo de Jambu para determinar, bajo su criterio, el número de clústers. Justifique su respuesta.
2. Explique el gráfico asociado al método y úselo como referencia a la hora de escoger el k.
3. **Ayuda:** Puede usar la implementación de los métodos propias de cada lenguaje.

Para a_big.csv:

1. **Implemente** el algoritmo kmedias en el lenguaje de preferencia (R o python) y asuma la norma p=2 como medida de distancia (distancia euclidiana).
2. Programe una solución para contrarrestar el tamaño del dataset y encontrar los centroides en un tiempo considerablemente menor a no usar su estrategia. Explique detalladamente su enfoque en el **.Rmd** o en el **informe.pdf** en caso de usar R y python respectivamente.
3. **Ayuda:** Comparar los centroides en **a.csv** y **a_big.csv**.

Para help.csv

El análisis exploratorio en este dataset es de suma importancia. Se recomienda usar un graficador en 3 dimensiones antes de aplicar las técnicas de clustering. Dicho esto:

- Cuántos clústers ve en el dataset **help** ?
- Qué pasa al aplicar la regla de asignación de clases en este dataset?
- Qué solución daría para asignar de **manera correcta** los valores de las clases y pueda analizar el desempeño del algoritmo de clustering de manera correcta?
- **Ayuda:** El nombre del dataset es un hint de la forma del dataset.

Consideraciones generales

Consideraciones de forma:

Ingrese a la dirección AprendizajeNoSupervisado y haga *fork* del repositorio donde encontrará los archivos csv y un README.

Este repositorio será propiedad de usted. En consecuencia, solo podrá realizar cambios en el mismo. El repositorio debe poseer lo siguiente:

1. Script .py intradocumentado o .Rmd reproducible y documentado.
2. README.md explicando la configuración del ambiente en el cual trabajó:
   - Ejemplo: README.md de Bootstrap
   - En el caso de usar python, hacer pip freeze de su versión para conocer librerías instaladas.
3. Puede crear un cuaderno con IPython Notebook de ser de su agrado. Sin embargo, es preferible un breve informe (**informe.pdf**) en el caso de usar Python.

Consideraciones de contenido:

- En el caso de usar python, se recomienda el uso de las funcionalidades de preprocesamiento provista por el paquete scikit-learn, así como el uso de las funcionalidades de los paquete pandas, numpy y scipy.
- La tarea es **estrictamente** individual. Se promueve la participación y discusión de la misma en un ambiente responsable. Sin embargo, cualquier evidencia de copia será severamente sancionada colocando una nota mínima de cero (0) puntos según lo establecido en la Ley de Universidades. **Cualquier tarea entregado** debe ser fruto de su propio trabajo.
- Fecha de Entrega: **Viernes 10 de abril de 2016**.
  - Hasta este día se aceptarán push's en los repositorios.
  - No se recibirá ninguna tarea por correo electrónico.
  - La regla de extensión de entregas se aplicará hasta el día **Martes 12 de abril de 2016**