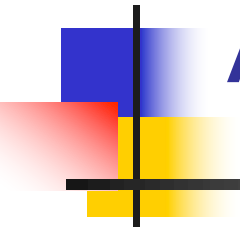
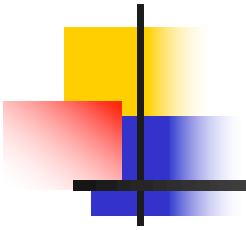
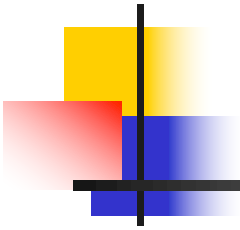


Aprendizaje de Máquina



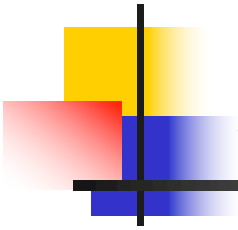
Menú

- Aprendizaje no-supervisado
- “Clustering”: Agrupamiento y segmentación de datos
- Medidas de Similitud
  - Transformación variables ordinales, nominales y categóricas
- Técnicas
  - Métodos de Partición
    - EM: k-medias
  - Métodos de Densidad
  - Métodos Jerárquicos



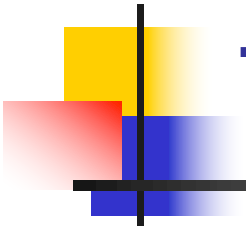
Objetivos

- Agrupar los datos en categorías o “clusters” de manera que los datos que estén más estrechamente relacionados pertenezcan al mismo grupo
- En ocasiones también se busca ordenar los grupos en una jerarquía
  - Agrupar categorías similares en un mismo grupo. (E.g. taxonomía)



Aprendizaje No-Supervisado

- Los algoritmos de agrupamiento son algoritmos de aprendizaje no-supervisado
  - No existen ejemplos previamente clasificados a partir de los cuales se realiza el modelo
- Se utilizan:
  - Cuando etiquetar datos es muy costoso
  - Cuando las categorías de las instancias cambian con el tiempo
  - Para encontrar patrones no sospechados que sean útiles para clasificar
  - Descubrir propiedades/relaciones de los datos
  - …..



Técnicas de Agrupamiento

- Fundamental para las técnicas de agrupamiento es un criterio de similitud entre los datos en cuestión
  - Normalmente depende de la aplicación
    - ¿Qué tan similares son México y Uganda?
  - ¿Qué tan similar es un 4 y un 5; un 4 y un 4.1?

Similitud de Datos
Matriz de Similitud

- En ocasiones se cuenta con una matriz que establece las similitudes entre parejas de datos

	México	Uganda	Holanda
México	1	0.4	0.3
Uganda	0.4	1	0.2
Holanda	0.3	0.2	1

- Estas matrices suelen ser simétricas
- Algunos algoritmos requieren matriz de diferencias.

Similitud de Datos
Similitud de Atributos

- Si no tenemos dicha matriz
  - Definir una medida de similitud entre los valores de cada atributo o variable
  - Definir una manera de combinar las diferentes similitudes entre los atributos de una pareja de datos (distancia entre datos)
  - Eg. <México, 25> y <Uganda, 30>
    - La distancia (similitud) entre México y Uganda es 0.4. Y podemos, por ejemplo, tomar la distancia entre 25 y 30 como |25-30|=5
    - Necesitamos una manera de generar un solo número a partir del 0.4 y el 5 (esto lo vemos más adelante)

Similitud de Datos
Similitud de Atributos

- Variables Cuantitativas
  - Los más común es definir la similitud (o diferencia) como:
    - $(X_{i,k} - X_{j,k})^2$
    - La diferencia al cuadrado del atributo k de los datos i y j
    - Existen, por supuesto, otras
      - Valor absoluto de la diferencia, correlación….

Similitud de Datos
Similitud de Atributos

- Variables Ordinales
  - E.g.
    - Gusto (horrible, regular, increíble)
    - Calificaciones (MB, B, C, NA)
  - Pueden representarse como una lista de número enteros sucesivos
    - (i-1)/(M-1)
    - Donde i es el i-ésimo valor y M es el número total de diferentes valores para dicho atributo
    - horrible=0/2, regular =1/2, increíble=2/2
  - Una vez transformados se manipulan como variables cuantitativas

Similitud de Datos
Similitud de Atributos

- Variables Categóricas
  - Si las variables son nominales (sin orden específico) el grado de diferencia entre parejas de valores debe darse explícitamente (e.g. Diferencia entre México y Uganda)
  - Una estrategia es sustituir la variable categoría por un vector indicador de k bits en donde k es el número de posibles valores para una categoría y sólo un bit correspondiente al valor que toma la categoría para un ejemplo esta puesto en uno
- Datos faltantes
  - Omitir del calculo de la similitud entre dos datos
  - Llenar con la media, moda o mediana (….)
  - Incluir la categoría “faltante”
  - ….

Similitud de Datos
Distancia entre Datos

- Un método común para determinar la similitud entre dos datos es mediante la distancia Euclidiana:
  - $\sqrt{(\sum(x_{i,k}-x_{j,k})^2)}$
  - La suma del cuadrado de las diferencias individuales entre atributos
  - El uso de la raíz se puede omitir pues no altera las distancias relativas de los datos
- ¿Qué otras posibilidades existen?
  - Edit distance, Manhattan distance, Mahalanobis, etc.

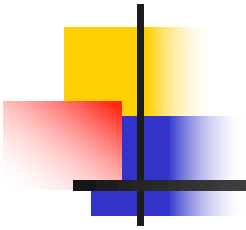
Similitud de Datos
Distancia entre Datos

- Un inconveniente es que datos cuyos rangos sean de mayor magnitud contribuirán más a la distancia
  - Eg. <distancia_al_trabajo, edad>
    - Dato 1= <2500mts, 35>
    - Dato 2=<2400mts, 15>
    - Dato 3=<2300mts, 34>
    - La distancia entre d1y d2 es 10400 mientras que entre d1 y d3 es 40001.
    - Depende de la aplicación, pero a primera vista parecería que d1 y d3 son más similares que d1 y d2
- ¿Solución?

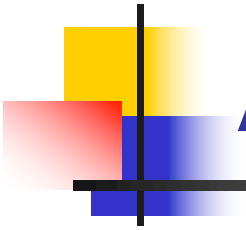
Similitud de Datos
Distancia entre Datos

- Adicionalmente se puede asignar un peso w_k a cada atributo k para establecer su importancia en determinar la diferencia entre datos
  - $\sum w_k(x_{i,k}-x_{j,k})^2$
  - Usualmente $\sum w_k=1$

Similitud de Datos
Distancia entre Datos

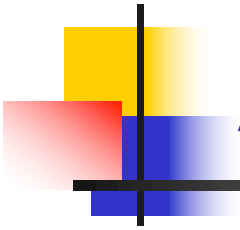


- La importancia de tener una buena medida de distancia es muchas veces mayor que el algoritmo específico de clustering a utilizar



Algoritmos de Agrupamiento

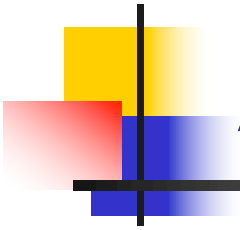
- El objetivo de estos algoritmos es dividir una serie de datos en grupos, de tal manera que las similitudes entre parejas de datos en un mismo grupo sea mayor que entre parejas de datos en diferentes grupos (o clusters)



Algoritmos de Agrupamiento

- Tres categorías importantes:
  - Métodos de Partición
    - Dados n datos, clasificarlos en k categorías disjuntas. Basados en distancias. E.g. K-medias, A-priori
  - Métodos Jerárquicos
    - Crean una jerarquía de los datos. Crea un árbol de clusters. E.g. Cobweb
  - Métodos de Densidad
    - Como partición sólo que no están basados en distancia sino en densidad. DBScan

Algoritmos de Agrupamiento

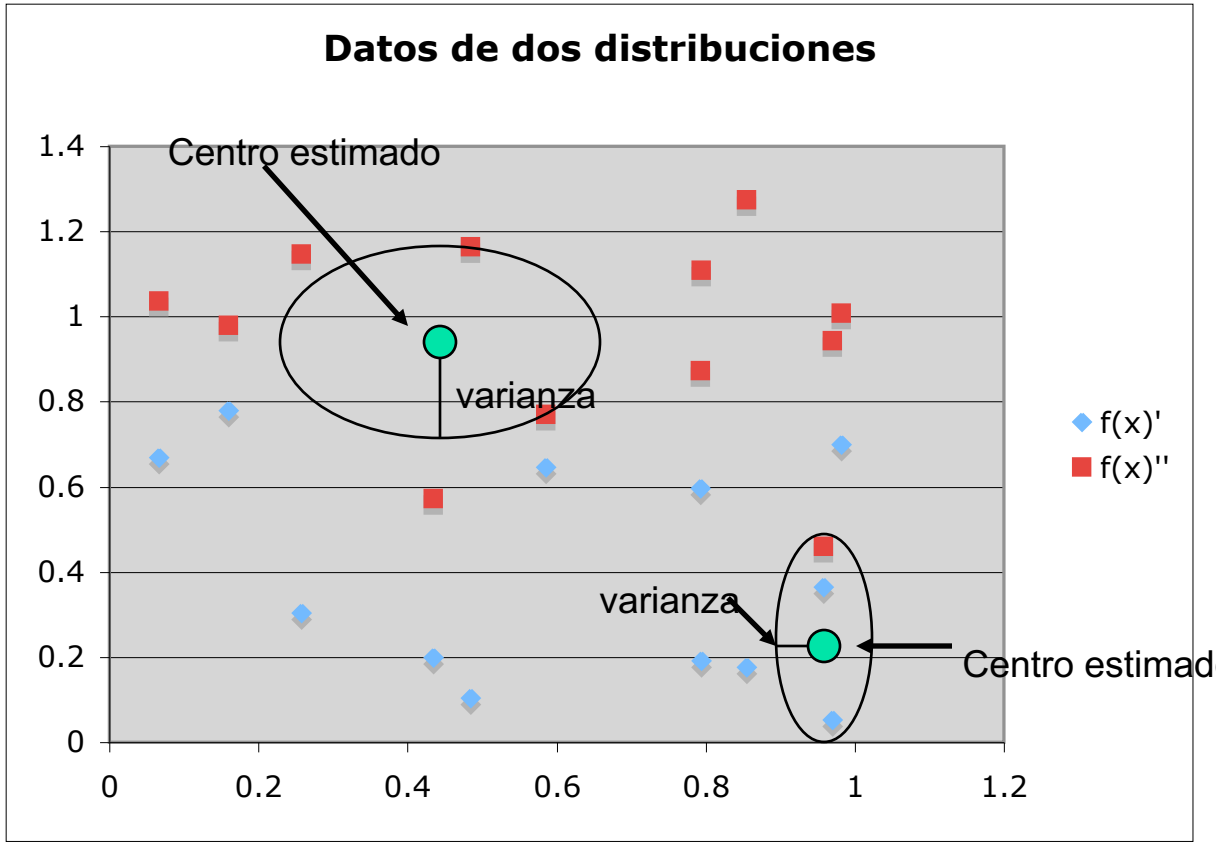


- Vamos a ver :
  - Expectation Maximization
    - k-medias (“k-means”)
  - Métodos Jerárquicos
  - DBSCAN
- Pregunta General:
  - ¿Se puede aprender algo, lo que sea, de datos sin etiquetas?
    - Depende de lo que asumas
    - La exactitud de los resultados depende de cómo la realidad y lo que asumes de ella se emparejan

Algoritmos de Agrupamiento
Expectation Maximization (EM)

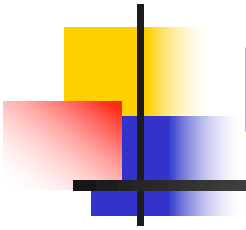
- Para su uso **asumimos** que conocemos
  - El número de categorías o grupos en los que se segmentan los datos
  - La distribución de los datos
    - Lo más común es asumir que los datos provienen de una distribución normal
- Desconocemos
  - La categoría de cada dato
    - Los ejemplos de entrenamiento no tienen el valor de la F.O.
  - Los parámetros de la distribución
    - Si asumimos que la distribución es normal, desconocemos la media y varianza

Algoritmos de Agrupamiento EM



La tarea es estimar los parámetros desconocidos de las distribuciones subyacentes (media, varianza,..)

Algoritmos de Agrupamiento k-medias “k-means”



- Las técnica que vamos a estudiar, k-medias, es miembro de la familia EM
- Es una simplificación muy efectiva, y muy usada, del algoritmo general de EM

Algoritmos de Agrupamiento k-medias "k-means"

- La simplificación consiste en que sólo vamos a estimar las medias de las distribuciones
  - La media de la distribución i (D_i) se estima:
    - $\mu_i=(1/w_i)\sum p_{i,j}\mathbf{x}_j$, donde
    - $p_{i,j}$: es la probabilidad de que el dato $\mathbf{x}_j$ haya sido generado por la distribución D_i
    - w_i: es la suma de todas las probabilidades ($p_{i,j}$) para todos los datos $\mathbf{x}_j$ en la distribución D_i
  - Nota: escribimos en negritas cuando el dato se trata de un vector
    - $\mathbf{x}_i$ consiste de varios valores, uno para cada atributo.

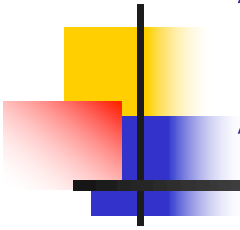
Algoritmos de Agrupamiento k-medias “k-means”

- Adicionalmente, como no pretendemos estimar la varianza podemos calcular las $p_{i,j}$ a partir del cuadrado de la distancia Euclidiana de cada punto $\mathbf{x}_j$ al centro (media $\boldsymbol{\mu}_i$) de cada distribución D_i
  - distancia($\mathbf{x}_j$,$\boldsymbol{\mu}_i$)=$||\mathbf{x}_j - \boldsymbol{\mu}_i||^2$

y aproximamos $p_{i,j}$ como

$$p_{i,j} \approx \begin{cases} 1 \text{ si } \boldsymbol{\mu}_i \text{ es la media más cercana a } \mathbf{x}_j, \\ 0 \text{ de otro modo} \end{cases}$$

Algoritmos de Agrupamiento
Algoritmo General

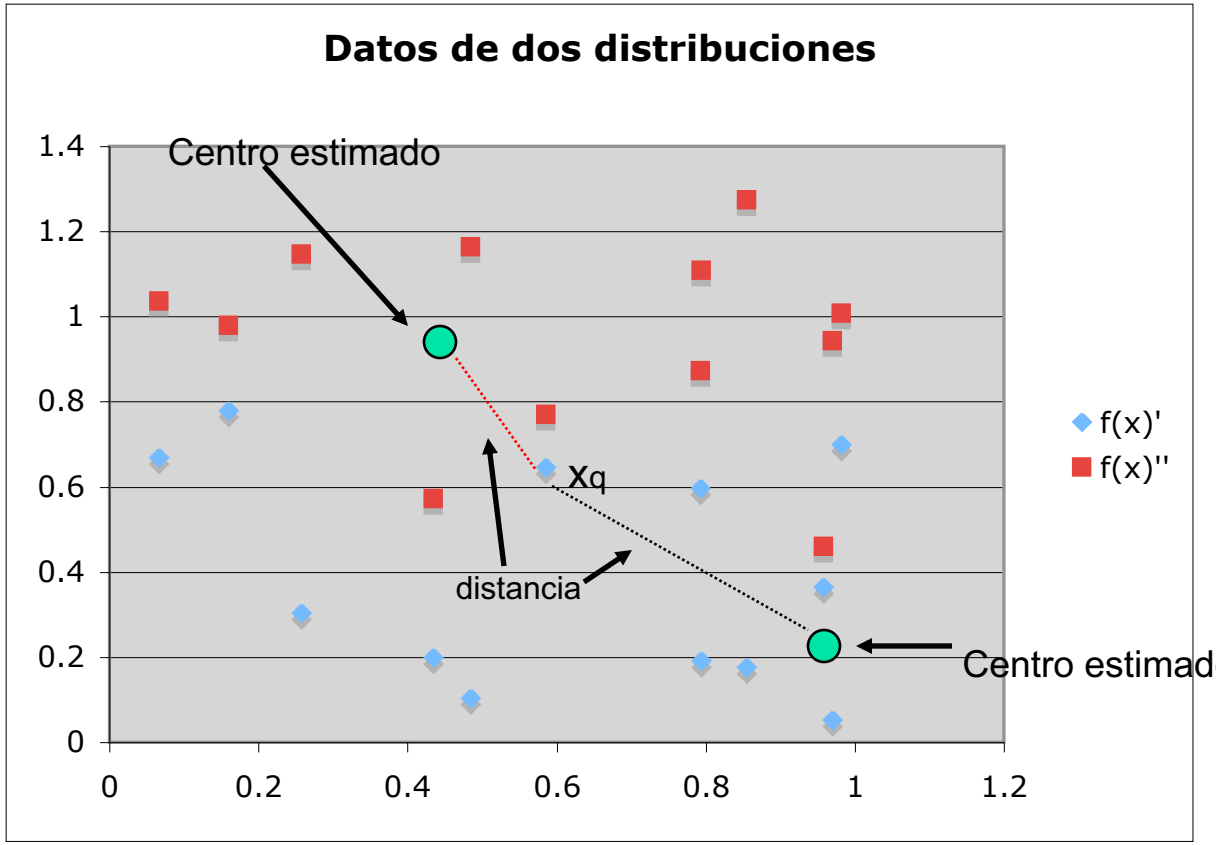


- El algoritmo, entonces, consiste de dos fases
  - Fase de estimación (fase “Expectation”)
    - Estimar la probabilidad de que un punto pertenezca a cada distribución
  - Fase de maximización (fase “Maximization”)
    - Recalcular las medias para maximizar la probabilidad de que cada dato pertenezca a alguna distribución (mueve los centros).

Algoritmos de Agrupamiento
Algoritmo k-medias

- Dados
  - n ejemplos de entrenamiento
  - Un entero k
  - Los valores iniciales para $\mu_1, \mu_2, \ldots, \mu_k$
- Do{
  - Clasificar los n ejemplos de acuerdo a la μ_i más cercana
    - Para cada $\mathbf{x}_j$, calcular cual es la μ_i más cercana
      - Mínimo$_{i=1,k}$ (distancia($\mathbf{x}_j$, μ_i))
    - Clasificar a $\mathbf{x}_j$ como miembro de esa distribución
  - Recalcular todas las μ_i
    - μ_i<---$(1/w_i)\sum p_{i,j}\mathbf{x}_j$
- }while(no haya cambio en $\mu_1, \mu_2, \ldots, \mu_k$)
- El algoritmo se repite hasta que no haya cambio en las medias estimadas, o un número de iteraciones pre-establecido

Algoritmos de Agrupamiento k-medias



Para clasificar x_q, se calcula la distancia a cada uno de los centros.
Se clasifica como perteneciente a la distribución de centro con menor distancia

Algoritmos de Agrupamiento
Ejemplo k-medias

DATOS	media 1	media 2	Pij (clase 1)	Pij (clase2)	Clase 1	Clase 2	media 1	media 2
	-0.5	1					0.144943	0.653655
	Distancia						**Distancia**	
0.67	1.17	0.33	0	1	0	0.67	0.525057	0.016345
0.19122452	0.691225	0.80878	1	0	0.19122	0	0.046282	0.46243
0.7	1.2	0.3	0	1	0	0.7	0.555057	0.046345
0.17606015	0.67606	0.82394	1	0	0.17606	0	0.031117	0.477595
0.103874	0.603874	0.89613	1	0	0.10387	0	0.041069	0.549781
0.646908	1.146908	0.35309	0	1	0	0.64691	0.501965	0.006747
0.19994854	0.699949	0.80005	1	0	0.19995	0	0.055006	0.453706
0.30341512	0.803415	0.69658	0	1	0	0.30342	0.158472	0.35024
0.0536079	0.553608	0.94639	1	0	0.05361	0	0.091335	0.600047
0.59716748	1.097167	0.40283	0	1	0	0.59717	0.452224	0.056488
0.87234622	1.372346	0.12765	0	1	0	0.87235	0.727403	0.218691
0.46032091	0.960321	0.53968	0	1	0	0.46032	0.315378	0.193334
0.97908235	1.479082	0.02092	0	1	0	0.97908	0.834139	0.325427
		Total =	5	8	0.72472	5.22924		Total =

Algoritmos de Agrupamiento
Ejemplo k-medias

DATOS	media 1	media 2	Pij (clase 1)	Pij (clase2)	Clase 1	Clase 2	media 1	media 2
	0.144943	0.653655					0.171355	0.70369
	Distancia						**Distancia**	
0.67	0.525057	0.016345	0	1	0	0.67	0.498645	0.03369
0.19122	0.046282	0.46243	1	0	0.1912	0	0.019869	0.51246
0.7	0.555057	0.046345	0	1	0	0.7	0.528645	0.00369
0.17606	0.031117	0.477595	1	0	0.1761	0	0.004705	0.52763
0.10387	0.041069	0.549781	1	0	0.1039	0	0.067481	0.59982
0.64691	0.501965	0.006747	0	1	0	0.6469	0.475553	0.05678
0.19995	0.055006	0.453706	1	0	0.1999	0	0.028594	0.50374
0.30342	0.158472	0.35024	1	0	0.3034	0	0.13206	0.40027
0.05361	0.091335	0.600047	1	0	0.0536	0	0.117747	0.65008
0.59717	0.452224	0.056488	0	1	0	0.5972	0.425812	0.10652
0.87235	0.727403	0.218691	0	1	0	0.8723	0.700991	0.16866
0.46032	0.315378	0.193334	0	1	0	0.4603	0.288966	0.24337
0.97908	0.834139	0.325427	0	1	0	0.9791	0.807727	0.27539
		Total =	6	7				

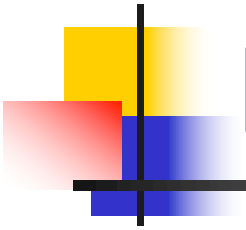
Algoritmos de Agrupamiento
Ejemplo k-medias

DATOS	media 1	media 2	Pij (clase 1)	Pij (clase2)	Clase 1	Clase 2	media 1	media 2
	0.171355	0.70369					0.17136	0.70369
	Distancia						**Distancia**	
0.67	0.498645	0.03369	0	1	0	0.67	0.49864	0.03369
0.1912	0.019869	0.51246	1	0	0.1912	0	0.01987	0.51246
0.7	0.528645	0.00369	0	1	0	0.7	0.52864	0.00369
0.1761	0.004705	0.52763	1	0	0.1761	0	0.00471	0.52763
0.1039	0.067481	0.59982	1	0	0.1039	0	0.06748	0.59982
0.6469	0.475553	0.05678	0	1	0	0.6469	0.47555	0.05678
0.1999	0.028594	0.50374	1	0	0.1999	0	0.02859	0.50374
0.3034	0.13206	0.40027	1	0	0.3034	0	0.13206	0.40027
0.0536	0.117747	0.65008	1	0	0.0536	0	0.11775	0.65008
0.5972	0.425812	0.10652	0	1	0	0.5972	0.42581	0.10652
0.8723	0.700991	0.16866	0	1	0	0.8723	0.70099	0.16866
0.4603	0.288966	0.24337	0	1	0	0.4603	0.28897	0.24337
0.9791	0.807727	0.27539	0	1	0	0.9791	0.80773	0.27539
		Total =	6	7				

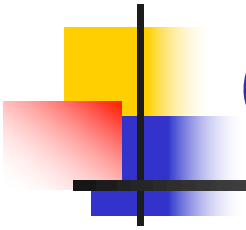
Algoritmos de Agrupamiento k-medias

- Se puede usar para datos categóricos
  - Encontrar medidas de similitud entre variables de tipo nominal y ordinal
    - Categorías binarias
      - E.g. Distancia de Hamming
    - Categorías nominales
      - E.g. La proporción de atributos iguales
    - Categorías ordinales
      - E.g. Asignar número enteros en orden ascendente. En orden de menor a mayor importancia de los valores posibles del atributo. Usar esos valores para calcular la distancia Euclidiana

Algoritmos de Agrupamiento k-medias

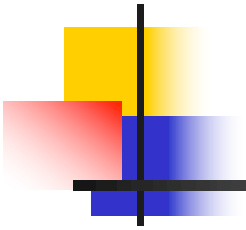


- Algunas desventajas
  - Hay que estimar el número, k, de clusters
  - Es sensible al valor inicial de cada media
  - El orden de los datos importa,…



Otros Algoritmos

- X-means: es una extensión de k-medias que ayuda a estimar el número de clusters
- Fuzzy k-means
- K-medioids: Parecido a k-medias pero utilizando medianas
- Métodos Jerárquicos
  - Vecinos cercanos
  - Vecinos lejanos
- Métodos de Densidad
  - DBSCAN



Ejercicio

- Baje datos de UCI
  - Sugiero Abalone
- Ejecute k-medias
- Pruebe, para cada grupo
  - Un árbol de decisión
  - Un random forest
- Compare los resultados
  - Aumentó alguna métrica segmentando los datos utilizando k-medias como preproceso?