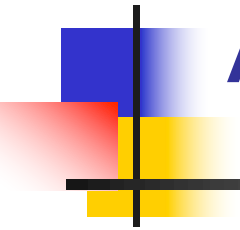
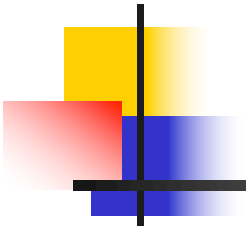
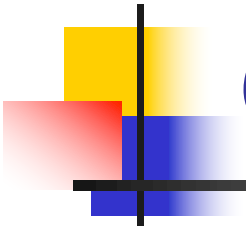


Aprendizaje de Máquina



Menú

- Aprendizaje no-supervisado
- “Clustering”: Agrupamiento y segmentación de datos
- Técnicas
  - Métodos de Partición
    - EM: k-medias
  - Métodos Jerárquicos
  - Métodos de Densidad

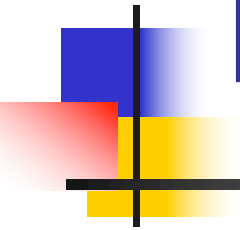


Objetivo

- Agrupar los datos en categorías o “clusters” de manera que los datos que estén más estrechamente relacionados pertenezcan al mismo grupo

Aprendizaje No-Supervisado Usos

- 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
  - …..

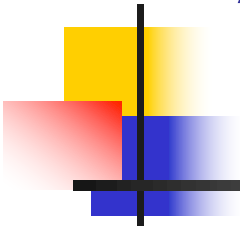


Métodos Jerárquicos

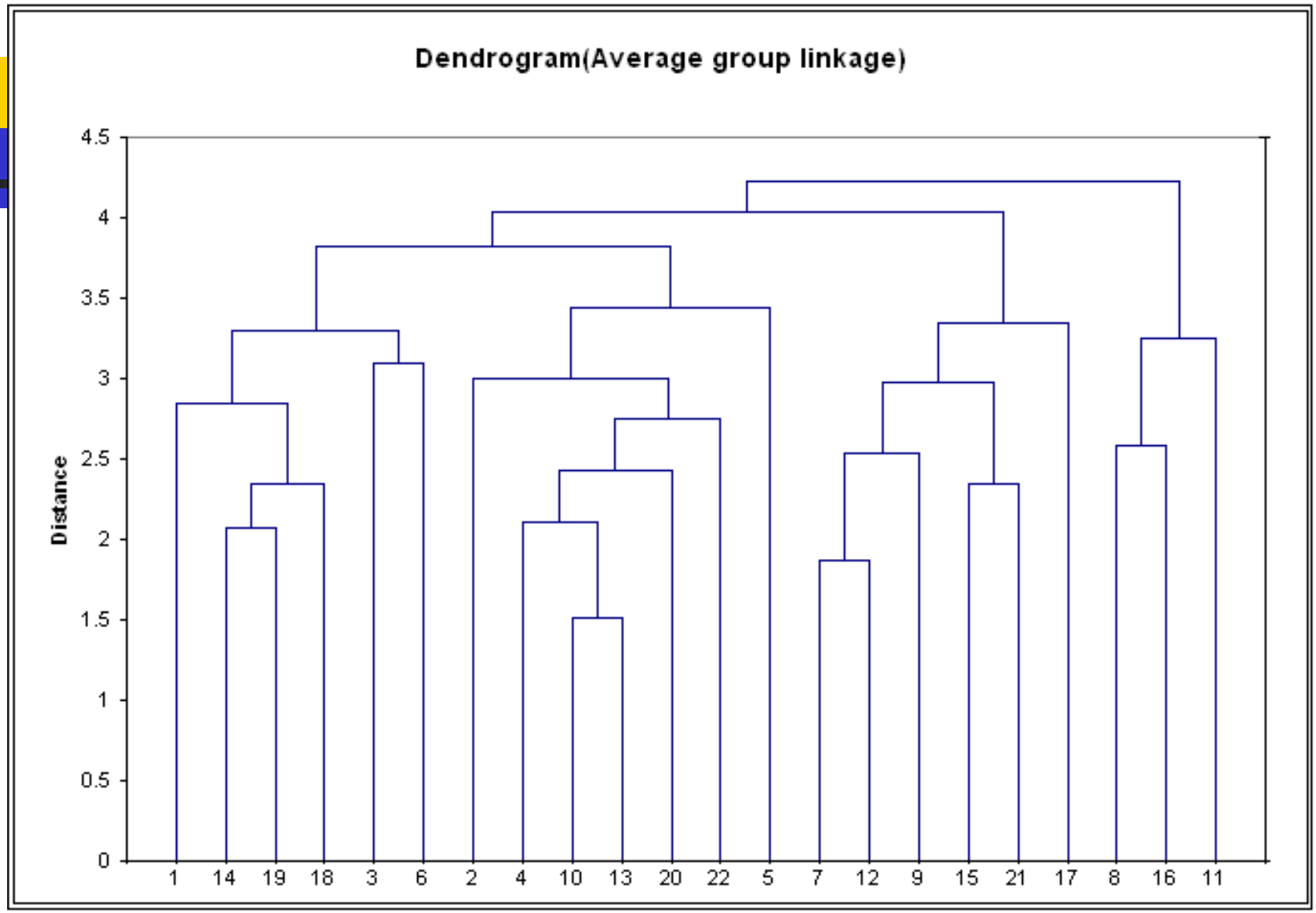
Algoritmos de Agrupamiento Jerárquico

- Estos algoritmos crean una jerarquía en la que cada nivel se forman grupos a partir de los grupos de los niveles inferiores
  - El nivel mas bajo esta formado de los datos individuales (grupos con un solo dato)
  - El nivel más alto tiene todos los datos
- La estructura resultante puede representarse como un árbol con el nodo raíz siendo el grupo de todos los datos y las hojas los datos individuales

Algoritmos de Agrupamiento Jerárquico



- El árbol suele dibujarse de manera que la altura de cada nodo es proporcional a la disimilitud entre los sub-grupos que lo componen
- La figura resultante se conoce como dendrograma y provee una imagen descriptiva de los datos



Por ejemplo 1=Seattle 14=San Francisco, 19=LA, 18= Phoenix

Dendrograma

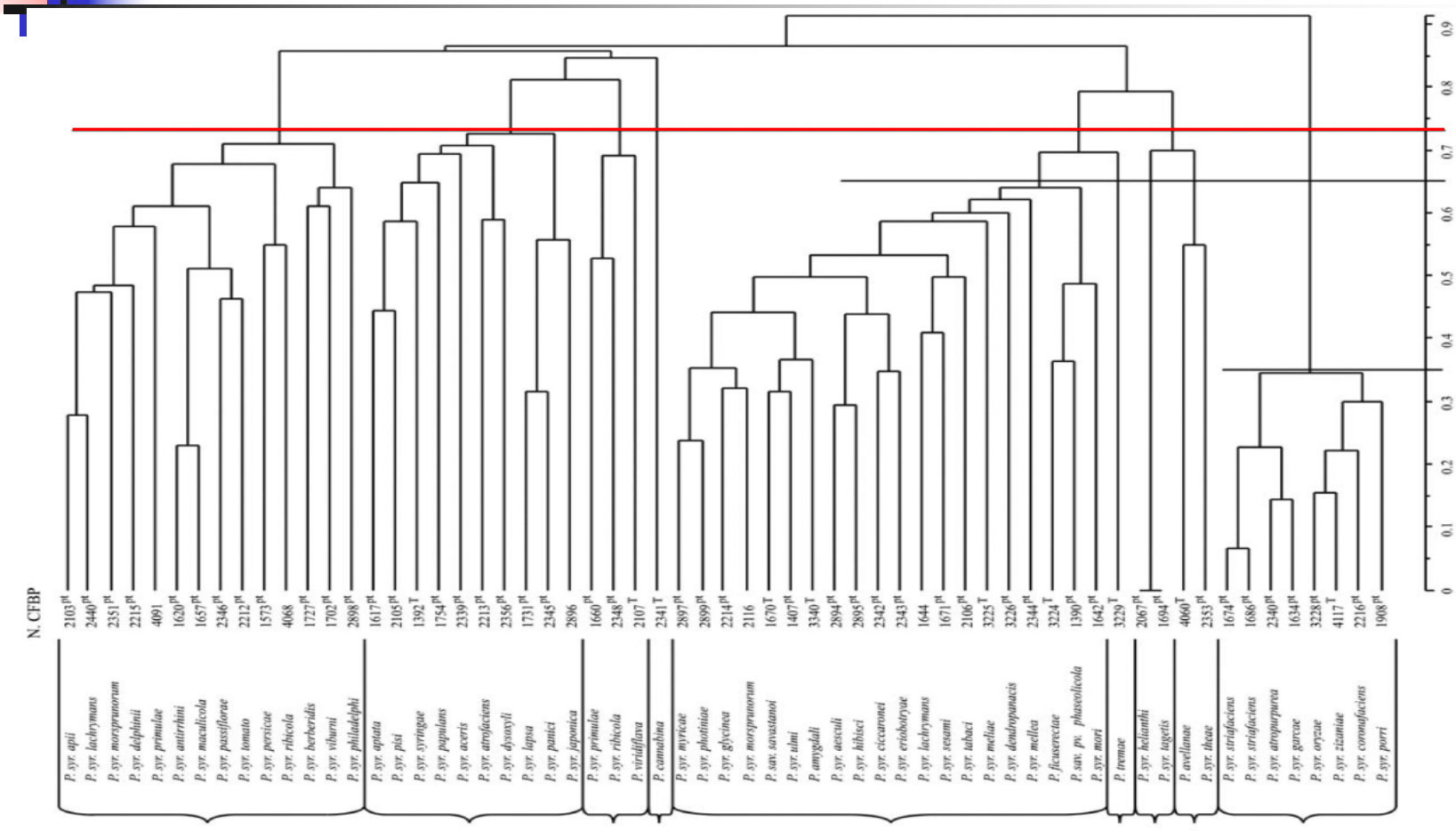
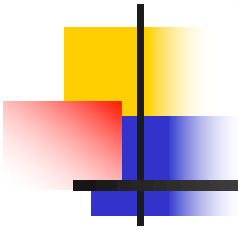


Figure 3 - Dendrogram obtained by comparison of BOX-PCR fingerprinting patterns from 61 bacterial type strains belonging to *Pseudomonas syringae* - *Pseudomonas viridiflava* large group (UPGMA analysis, Jaccard coefficient). Isolates obtained from the "Collection Française des Bactéries Phytopathogènes" (CFBP, Angers, France). T: species type strain, PM: pathotype strain.

Algoritmos de Agrupamiento Jerárquico



- A diferencia de k-medias, aquí no se especifica en número de grupos de antemano
- Es labor del analista determinar a que nivel de la jerarquía se encuentra el agrupamiento más natural de los datos
- Nota:
  - Estos algoritmos siempre van a encontrar una jerarquía, aunque ninguna exista en los datos
  - Diferentes algoritmos encuentran diferentes jerarquías
- La formación de grupos se basa en alguna medida de similitud entre sub-grupos

Algoritmos de Agrupamiento Jerárquico

- Existen dos estrategias para crear dicha estructuras
  - Agregación
    - Comienzan en el nivel más bajo (datos individuales) y recursivamente forman grupos más y más grandes hasta terminar con uno solo
  - Disgregación
    - Comienzan con el grupo de todos los datos y lo dividen en dos luego en cuatro, etc., hasta tener cada dato en un grupo

Algoritmos de Agrupamiento Jerárquico

- Algoritmos de Agregación
  - Tomar los dos grupos más similares y unir en un solo grupo
  - Repetir hasta que sólo quede un grupo
- Necesitamos una manera de calcular la similitud entre grupos

Algoritmos de Agrupamiento Jerárquico

- “Single Linkage” o Vecino Cercano
  - La distancia entre los grupos A y B es la mínima distancia entre sus datos
  - Ejemplo (usando la distancia Euclidiana al cuadrado):
    - A={<1,2>,<3,2>}, B={<5,6>,<5,7>,<6,6>}
    - d(<1,2>,<5,6>)=32, d(<1,2>,<5,7>)=41, d(<1,2>,<6,6>)=41
    - d(<3,2>,<5,6>)=20, d(<3,2>,<5,7>)=29, d(<3,2>,<6,6>)=25
    - La distancia entre A y B es 20

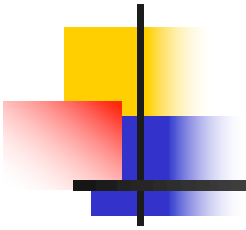
Algoritmos de Agrupamiento Jerárquico

- "Complete Linkage" o Vecino Lejano
  - La distancia (o similitud) entre dos grupos es el valor de la pareja más disímil
  - Del ejemplo anterior
    - La distancia entre A y B es 41
- Promedio de grupo
  - La distancia entre dos grupos se toma como el promedio de la distancia entre todas las parejas de datos

$$\frac{1}{N_k N_L} \sum_{i \in C_k} \sum_{C_L} d(\mathbf{X_i}, \mathbf{X}_j)$$

Algoritmos de Agrupamiento
Comentarios

- “Single Linkage”
  - Grupos poco compactos (chaining)
  - El diámetro de un grupo es la máxima distancia entre dos de sus elementos
  - Este método produce grupos con diámetros grandes
- “Complete Linkage” o Vecino Lejano
  - Lo opuesto al anterior: crea grupos con diámetro chico
  - Es más probable que datos ruidosos provoquen agrupaciones inadecuadas
- Promedio de grupo
  - Compromiso entre los dos anteriores

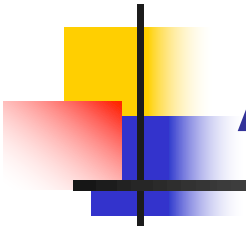


Nota

- Una ves que se encuentra la similitud entre todos los grupos usando alguna de las medidas anteriores, se juntan los dos grupos más similares y se repite el proceso

Algoritmos de Agrupamiento

- Algoritmos de Disgregación
  - Tomar uno de los grupos con más de un elemento y dividir en dos grupos
  - Repetir hasta que todos los grupos tengan un solo elemento
- La división en grupos debe buscar que los elementos en cada subgrupo sean más similares entre los elementos del otro sub grupo
- ¿Ideas?



Algoritmos de Agrupamiento

- Ejercicio en clase

Métodos de Densidad

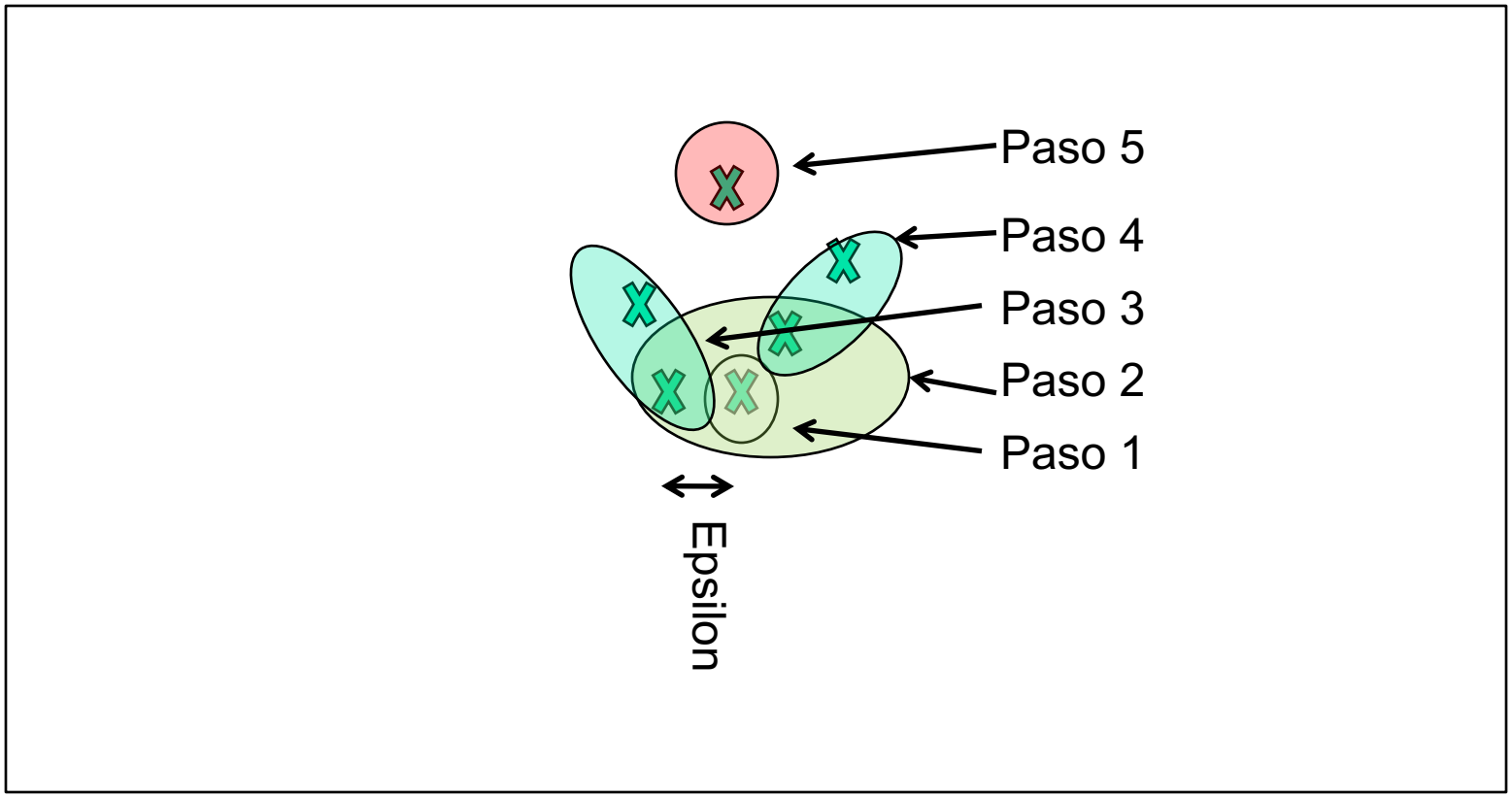
Algoritmos de Agrupamiento
Densidad (DBSCAN)

- Genera grupos que tengan cierta densidad
- Recibe dos parámetros: Epsilon y Min
  - La distancia de un dato a su vecino más cercano dentro del mismo grupo es a lo más épsilon
  - Min especifica el número de datos mínimo para formar un grupo

Algoritmos de Agrupamiento
DBSCAN Algoritmo

- D<-Todos los datos, i=0
- While (D <>{})
  - Sacar un dato p de D (D<--D-{p})
  - V<--{p}, Ci<--{}
  - While(V<>{})
    - Sacar un dato d de V (V<--V-{d})
    - O<--Todos los datos de D a Epsilon de distancia o menos de d
    - If |O|+1<Min
      - si Ci={}, d es clasificado como ruido
    - Else
      - V<--V U O, D<--D-V
      - Insertar V y d en Ci: Ci<--Ci U V U {d}
  - If (Ci<>{})
    - i<--i+1

DBSCAN



Algoritmos de Agrupamiento
Ejemplo DBSCAN

- Epsilon= 1
- Min =2

Datos
23
8
12
13
4
22
14
3

Algoritmos de Agrupamiento
Densidad

- No requiere que se especifique el número de grupos
- Encuentra datos ruidosos
- Los grupos tienen “formas” arbitrarias
- Es un arte definir Epsilon. Epsilon se define una vez por corrida por lo que se le dificulta al algoritmo encontrar grupos con diferentes densidades
  - ¿Posible solución?

Algoritmos de Agrupamiento
Densidad

- Aplicaciones
  - Apareo de proteínas
  - Detección satelital de uso de tierra
  - ….