Text Mining
Introducción a Minería de Datos

Laura Alonso i Alemany

Facultad de Matemática, Astronomía y Física
UNC, Córdoba (Argentina)
http://www.cs.famaf.unc.edu.ar/~laura

SADIO
12 de Marzo de 2008

qué es la minería de datos?

A technique using software tools geared for the user who typically does not know exactly what he's searching for, but is looking for particular patterns or trends.
`http://www.dmreview.com/rg/resources/glossary.cfm`
Discovery mode of data analysis, or analyzing detail data to unearth unsuspected or unknown relationships, patterns and associations that might be of value to the organization. Most often associated with predictive analysis.
`http://it.csumb.edu/departments/data/glossary.html`
áreas relacionadas:

- exploratory data analysis
- machine learning

qué es la minería de datos?

A technique using software tools geared for the user who typically does not know exactly what he's searching for, but is looking for particular patterns or trends.
`http://www.dmreview.com/rg/resources/glossary.cfm`
Discovery mode of data analysis, or analyzing detail data to unearth unsuspected or unknown relationships, patterns and associations that might be of value to the organization. Most often associated with predictive analysis.
`http://it.csumb.edu/departments/data/glossary.html`
áreas relacionadas:

- exploratory data analysis looking at data to form hypotheses worth testing
- machine learning

qué es la minería de datos?

A technique using software tools geared for the user who typically does not know exactly what he's searching for, but is looking for particular patterns or trends.
`http://www.dmreview.com/rg/resources/glossary.cfm`
Discovery mode of data analysis, or analyzing detail data to unearth unsuspected or unknown relationships, patterns and associations that might be of value to the organization. Most often associated with predictive analysis.
`http://it.csumb.edu/departments/data/glossary.html`
áreas relacionadas:

- exploratory data analysis
- machine learning discover patterns in the data that relate data attributes with a target (class) attribute. These patterns are then utilized to predict the values of the target attribute in future data instances. ← The data have no target attribute. We want to explore the data to find some intrinsic structures

diferencias entre minería de datos y minería de texto

en la minería de texto encontramos

- input **no estructurado**, es necesario
  - identificar entidades
  - identificar relaciones
- falta de abstracción, porque los ejemplos no están clasificados (etiquetados) o las clases son muy chicas
  - $\longrightarrow$ alta dimensionalidad
  - $\longrightarrow$ difícil de interpretar
  - $\longrightarrow$ no está conectado con otros tipos de conocimiento
- rango de resultados **totalmente desconocido**

algunas técnicas de minería de datos

- técnicas **no supervisadas**
  - *reglas de asociación* vs. reglas orientadas a clases
  - *minado de secuencias*
  - *clustering* vs. clasificación

reglas de asociación

las reglas de asociación son el ejemplo prototípico de minería de datos, y su aplicación típica, el análisis de la canasta de mercado

- los datos son categóricos (y los contínuos?)
- $I = \{i_1, i_2, ..., i_n\}$ un conjunto de items
- la transacción t es un conjunto de items $t \subseteq I$
- la base de datos es un conjunto de transacciones $T = \{t_1, t_2, ..., t_n\}$

¿ qué serían items, transacciones, etc. en una base de datos de un hipermercado?

¿ qué serían items, transacciones, etc. en una base de datos de un e-comerciante?

¿ qué serían items, transacciones, etc. en una base de datos de documentos?

el modelo: las reglas

- un **itemset** X es un conjunto de items en I, $X \subseteq I$
  el concepto de *frequent itemset* es básico para varios algoritmos
- una transacción t contiene un itemset X si $X \subseteq t$
- una **regla** es una implicación de forma $X \rightarrow Y$, donde $X, I \subset I$ y $X \cap Y = \emptyset$
- en un conjunto de transacciones T,
  - el **soporte** de una regla es el porcentaje de transacciones T que contienen $X \cup Y$ ($s = Pr(X \cup Y)$)
  - la **confianza** de una regla es el porcentaje de transacciones de T que contienen X y también Y ($c = Pr(Y|X)$)

  $\rightarrow$ el usuario especifica un soporte y confianza mínimos

  $\rightarrow$ puede haber diferentes valores (especialmente de soporte) para diferentes items, condicionados por la probabilidad de ocurrencia de cada item (así recuperamos reglas con items que ocurren poco)

cuestiones sobre las reglas

- se encuentran todas las reglas (pueden ser muchas)
- se representa a los items de forma atómica, simplificando (sin atributos como precio, cantidad, etc.)
- muchos algoritmos: **Apriori** es uno de los más eficientes, se basa en *frequent itemsets*

  `Rakesh Agrawal, Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules. Proc. 20th Int. Conf. on Very Large Data Bases, VLDB.`

  ← no se puede aplicar si se usan diferentes valores de soporte
- no hay objetivos predefinidos en la parte derecha de la regla

  → por eso es **no supervisado**

← si el usuario está interesado en ciertos items usa **reglas de asociación de clases (CARs)**

ejemplos de clases que nos podrían interesar en un hipermercado?
en una colección de documentos?

reglas de asociación vs. reglas de asociación de clase (CARs)

- en las CARs, hay que encontrar los **ruleitems** que superan el soporte mínimo especificado
  - un *ruleitem* es una tupla *(conjunto_de_condiciones, clase)*
  - un *conjunto_de_condiciones* es un *itemset*
  - una CAR tiene la forma *conjunto_de_condiciones* $\rightarrow$ *clase*
- se pueden usar diferentes soportes y confianzas mínimos para diferentes clases
- si no queremos que se generen reglas para una clase, podemos determinar que su soporte mínimo sea de 100%

minado en secuencias

asunción básica: el **orden** importa (vs. reglas simples, sin orden)
objetivo: encontrar **patrones** en secuencias, por ejemplo:

- orden de compra de productos (para eso las tarjetas de fidelización!)
- comportamiento navegacional de usuarios
- uso de disco, red, etc.
- adn, proteínas
- el texto es una secuencia!!!!

el modelo en minado de secuencias

- una **secuencia** es una lista ordenada de *itemsets*, donde cada itemset es un *elemento* de la secuencia
- el **tamaño** de una secuencia es su cantidad de elementos (itemsets)
- la **longitud** de una secuencia es su cantidad de items
- el **soporte** de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S
- las **secuencias frecuentes** (o **patrones secuenciales**) son las subsecuencias de una secuencia que tienen un soporte mínimo

qué es clustering?

Clustering is a process of partitioning a set of data into subsets or clusters such that a data element belonging to a cluster is more similar to data elements belonging to the same cluster than the data elements belonging to other clusters.
`http://www.dmreview.com/rg/resources/glossary.cfm`

- cómo se representan los objetos?
- cómo se calcula la semejanza?
- cómo se agrupan los elementos?
- cómo sabemos si una solución de clustering es buena?

qué es clustering?

Clustering is a process of partitioning a set of data into subsets or clusters such that a data element belonging to a cluster is more similar to data elements belonging to the same cluster than the data elements belonging to other clusters.
`http://www.dmreview.com/rg/resources/glossary.cfm`

- cómo se representan los objetos?
  - los objetos se representan como vectores, donde cada característica es una dimensión
  - se requiere buen conocimiento del dominio
  - y también de cómo funcionan las técnicas de clustering
- cómo se calcula la semejanza?
- cómo se agrupan los elementos?
- cómo sabemos si una solución de clustering es buena?

qué es clustering?

Clustering is a process of partitioning a set of data into subsets or clusters such that a data element belonging to a cluster is more similar to data elements belonging to the same cluster than the data elements belonging to other clusters.
`http://www.dmreview.com/rg/resources/glossary.cfm`

- cómo se representan los objetos?
- cómo se calcula la semejanza?
  - medidas de distancia geométrica en un espacio vectorial: distancia euclídea, coseno de los vectores, etc.
  - medidas de diferencia de distribuciones de probabilidad: divergencia de Kullback-Leibler
- cómo se agrupan los elementos?
- cómo sabemos si una solución de clustering es buena?

qué es clustering?

Clustering is a process of partitioning a set of data into subsets or clusters such that a data element belonging to a cluster is more similar to data elements belonging to the same cluster than the data elements belonging to other clusters.
`http://www.dmreview.com/rg/resources/glossary.cfm`

- cómo se representan los objetos?
- cómo se calcula la semejanza?
- cómo se agrupan los elementos?
  - métodos jerárquicos vs. planos
  - métodos partitivos vs. aglomerativos
  - k-means
  - diferentes criterios (algoritmos) para crear o dividir clusters: el cluster con mayor semejanza, con mayor disparidad, media
- cómo sabemos si una solución de clustering es buena?

qué es clustering?

Clustering is a process of partitioning a set of data into subsets or clusters such that a data element belonging to a cluster is more similar to data elements belonging to the same cluster than the data elements belonging to other clusters.
`http://www.dmreview.com/rg/resources/glossary.cfm`

- cómo se representan los objetos?
- cómo se calcula la semejanza?
- cómo se agrupan los elementos?
- cómo sabemos si una solución de clustering es buena?
  - un experto inspecciona el contenido de los clusters
  - se ofrece contenido representativo: el medoide, las características más frecuentes
  - pureza: mayor semejanza entre los elementos de un cluster, mayor disparidad entre los elementos de distintos clusters
  - poder clasificador del modelo resultante
  - comparación con algunas clases ← parcialmente supervisado

para qué sirve el clustering?

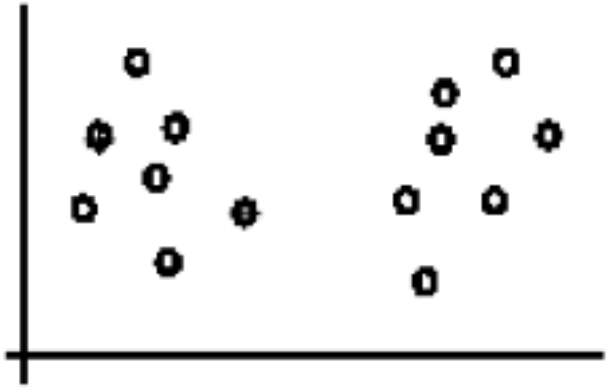
- agrupar clientes con el mismo comportamiento → con los mismos gustos, para recomendar a unos lo que gustó a los otros (p.ej., Netflix Prize)
- detectar clientes con un alto riesgo para una compañía aseguradora
- detectar pacientes con un alto riesgo de desarrollar cáncer
- predecir el tiempo? ;)

en tratamiento del lenguaje natural

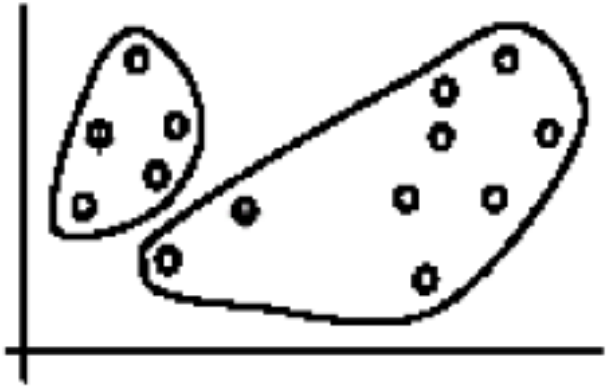
- agrupar palabras que se comportan igual
  - co-ocurren con las mismas palabras
  - ocurren en las mismas estructuras sintácticas
  - ocurren en los mismos contextos de documentos
- identificar grupos de significados parecidos
- identificar temas (*topics*), posiblemente armando una jerarquía de temas (deportes → futbol → Bundesliga...)

k-means: el método más simple

1. el usuario determina k = número de clusters
2. se eligen como centros de clusters (centroides) iniciales k puntos en el espacio de objetos (al azar o con algún criterio)
3. cada objeto se asigna al centroide más cercano
4. se re-calculan los centroides en base a la población de cada cluster
5. si no se llegó al criterio de convergencia, volver a 2.
   criterio de convergencia: no hay ningún cambio en los centroides o en la assignación de objetos a clusters

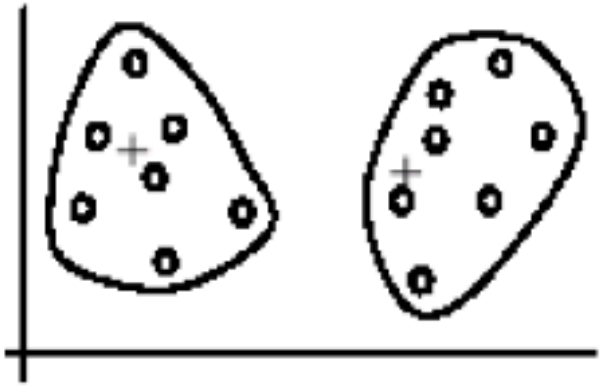


(A). Random selection of k centers

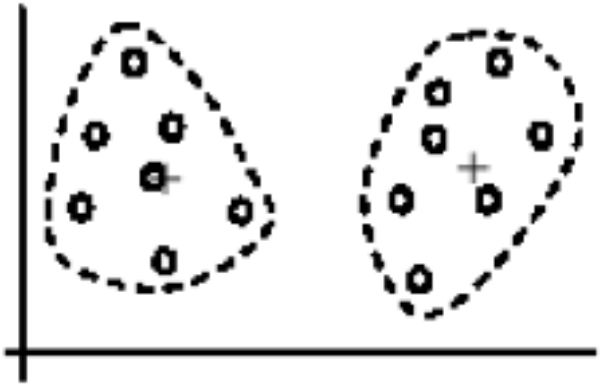


Iteration 1: (B). Cluster assignment

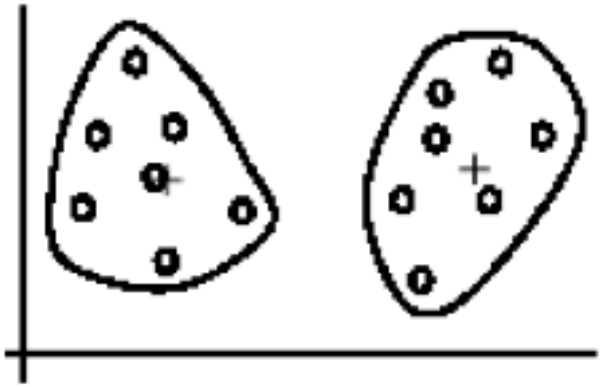
(C). Re-compute centroids



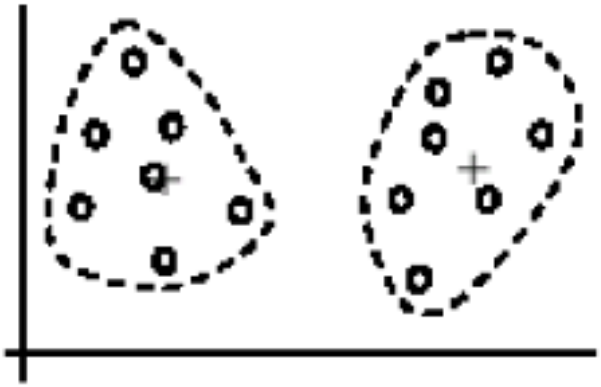
Iteration 2: (D). Cluster assignment



(E). Re-compute centroids



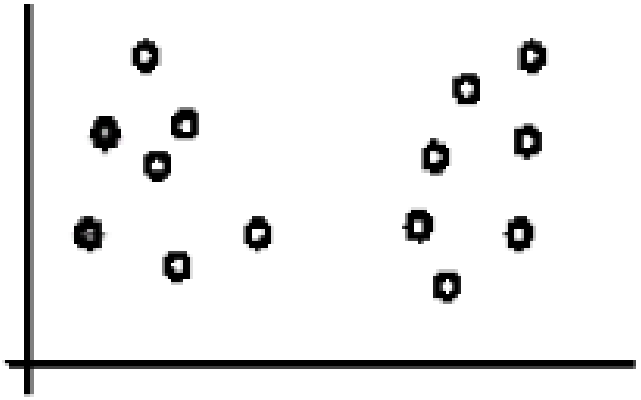
Iteration 3: (F). Cluster assignment



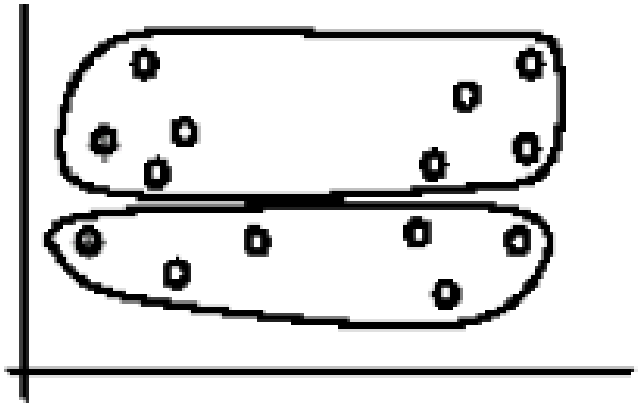
(G). Re-compute centroids

propiedades de k-means

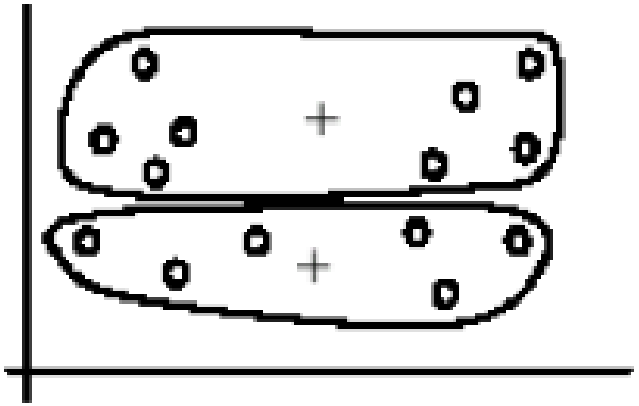
- se puede usar con datos en disco, no necesariamente en memoria
- muy simple, fácil de entender e implementar
- eficiente
- converge a un óptimo local (el global es muy difícil de encontrar, mucha complejidad)
- es sensible a las semillas iniciales, principal razón para converger a un óptimo global
- sólo se puede aplicar si se puede definir una *media* de los valores de las características (dimensiones de los vectores) $\rightarrow$ difícil para valores categóricos
- k es especificado por el usuario
- es sensible a *outliers* $\leftarrow$ se puede usar sampling
- no es capaz de reconocer formas distintas a gaussianos (hiper-esferas)



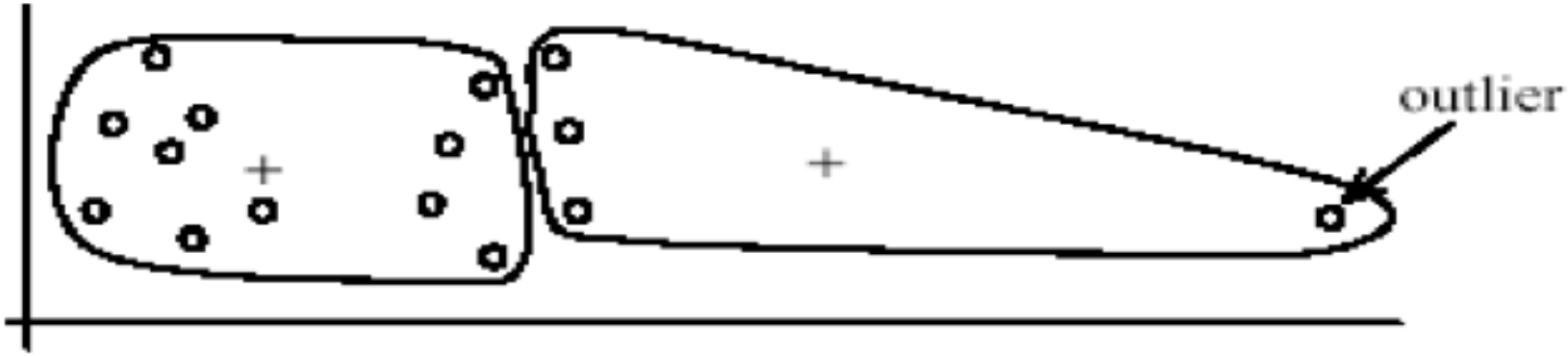
(A). Random selection of seeds (centroids)



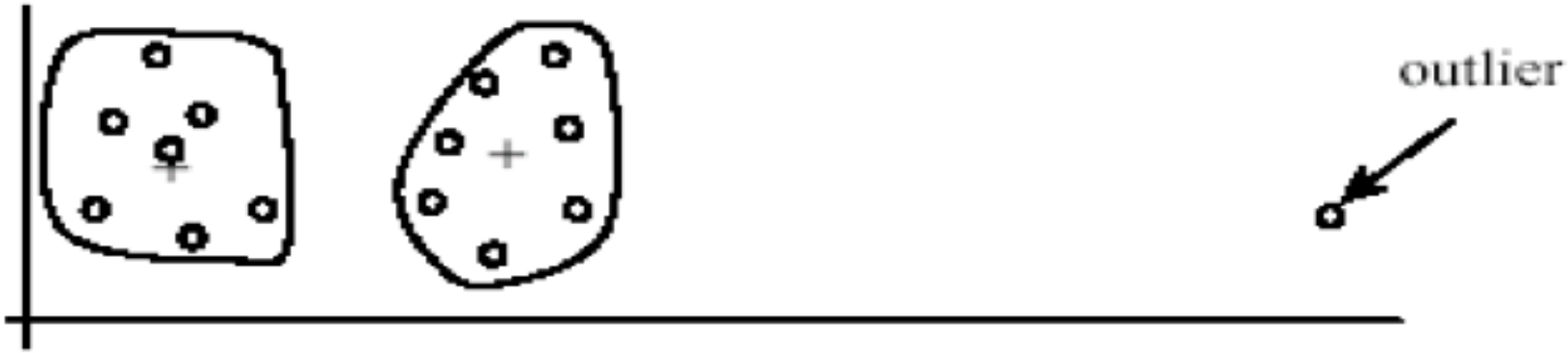
(B). Iteration 1



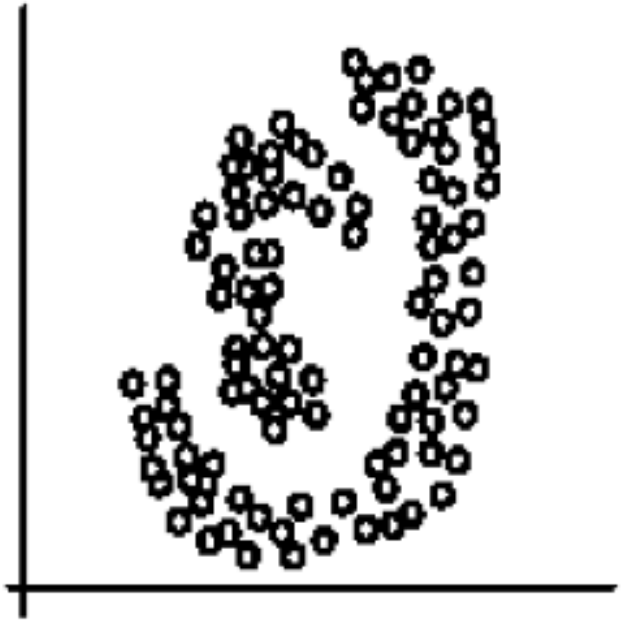
(C). Iteration 2



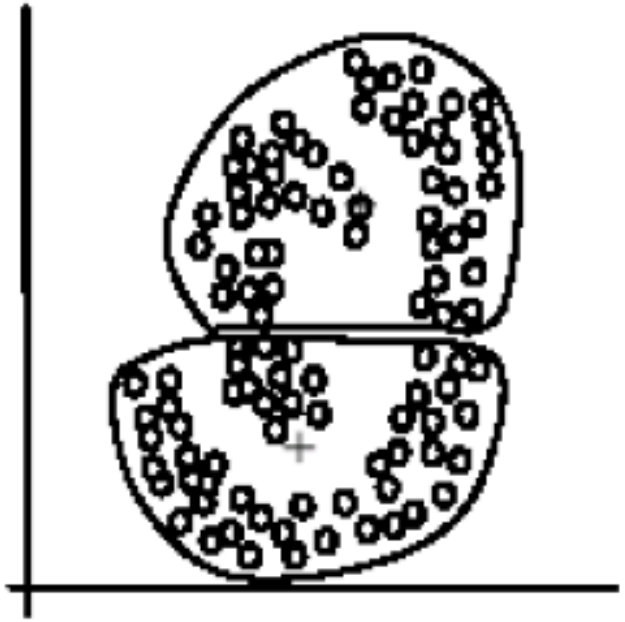
(A): Undesirable clusters



(B): Ideal clusters



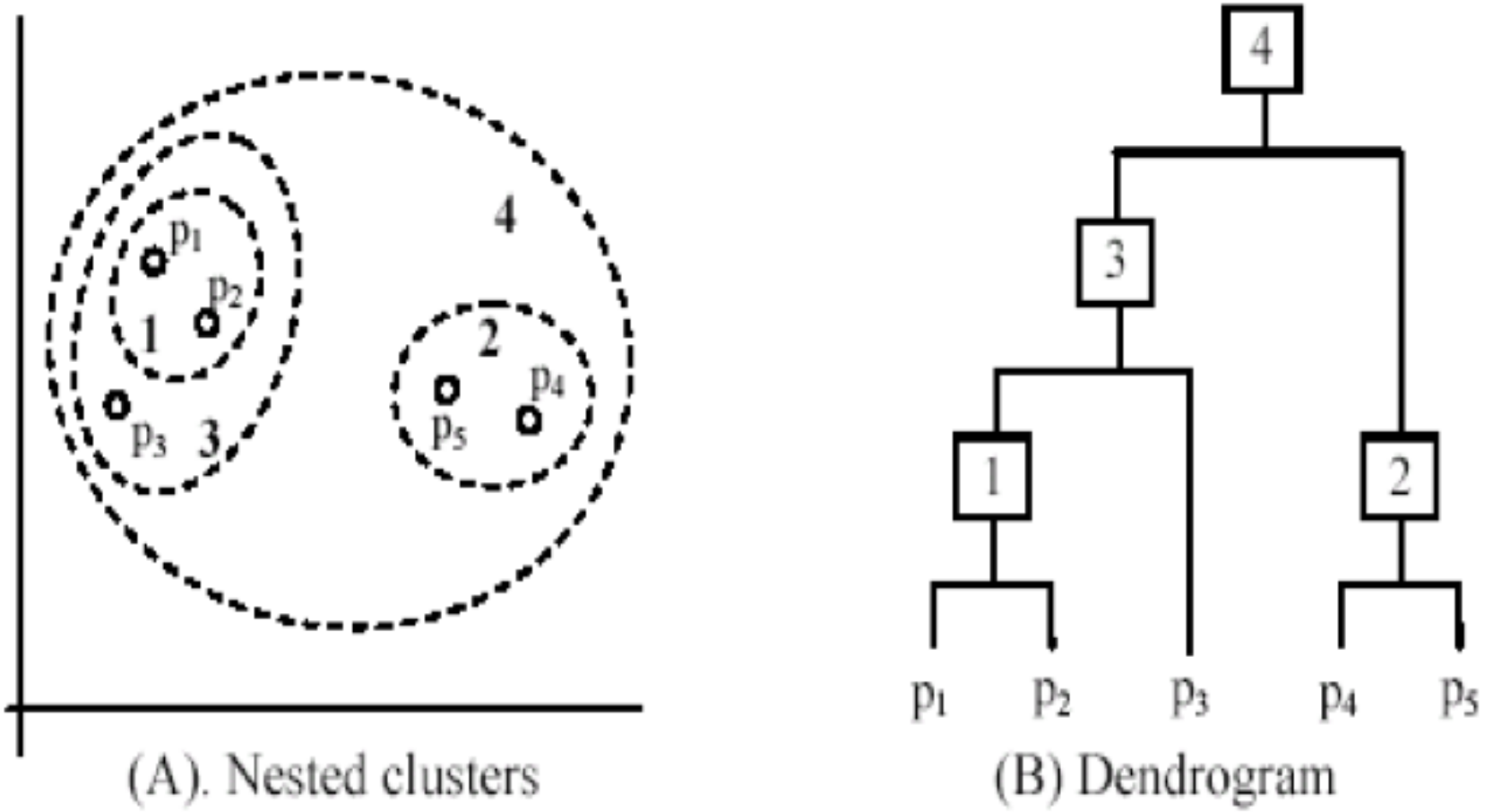
(A): Two natural clusters



(B): k-means clusters

clustering jerárquico

produce una secuencia de clusters anidados, un árbol (*dendrograma*)



tipos de clustering jerárquico

- aglomerativo (*bottom-up*): une los dos elementos (o clusters) más semejantes
- partitivo (*top-down*): divide el cluster que contiene elementos más distintos

diferentes criterios para medir la semejanza entre clusters:

- single link: distancia entre los dos puntos más cercanos, uno de cada cluster
  $\rightarrow$ produce clusters con forma alargada $\rightarrow O(n^2)$
- complete link: distancia entre los dos puntos más lejanos, uno de cada cluster
  $\rightarrow$ sensible a outliers $\rightarrow O(n^2 logn)$
- average link: la distancia entre dos clusters es la media de las distancias entre todos los pares de puntos, uno de cada cluster
  $\rightarrow O(n^2 logn)$
- centroides: la distancia entre dos clusters es la distancia entre