Clustering

Clasificación no supervisada

Javier G. Sogo

10 de marzo de 2015

Introducción

Objetivos

Conjunto de casos u objetos, caracterizados por varias variables:

	Var_1	...	Var_i	...	Var_n
x_1	x_1^1	...	x_i^1	...	x_n^1
...	...	...	...	...	...
x_j	x_1^j	...	x_i^j	...	x_n^j
...	...	...	...	...	...
x_N	x_1^N	...	x_i^N	...	x_n^N

- Encontrar **conglomerados**/grupos/clusters que emerjan naturalmente de los datos.
- Los elementos de un grupo son **similares** y pueden ser descritos por características comunes.
- **Homogeneidad** del grupo y **heterogeneidad** entre grupos.
- Puede ser considerado un **arte** (¡cuidado!).

Ejemplos

Marketing

- Segmentación de clientes.
- Sistemas de recomendación: clientes similares... pero también productos similares (clustering en columnas)
- Publicidad dirigida, ofertas.

Textos

- Clasificación en temáticas: *topic modelling*.
- Recuperación de información.

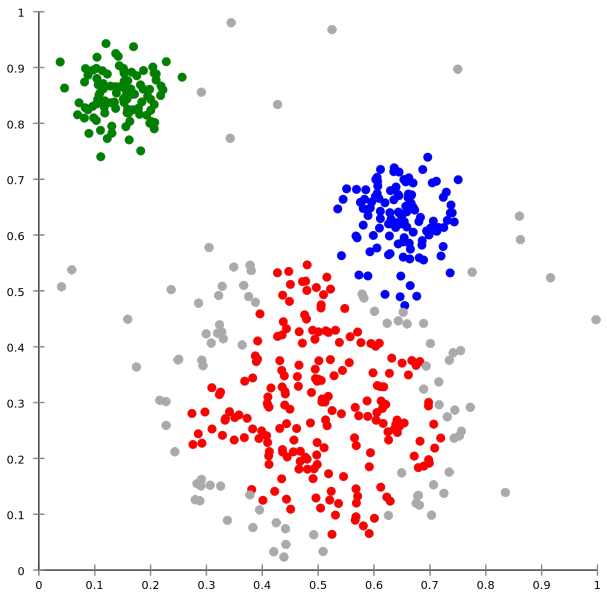
Internet

- Patrones de comportamiento de visitas en web.
- Clasificación de logs de acceso.

Tipos de clustering

- **Jerárquico**: los datos se van agrupando en conjuntos cada vez más numerosos hasta que sólo queda uno de ellos que reune a todos los elementos.
- No jerárquico:
  - **Particional**: los elementos se dividen en un número determinado de grupos (prefijado de antemano)
  - **Probabilista**: un elemento puede pertenecer a varios grupos simultáneamente con distintas probabilidades (mixtura de Gaussianas).

Interpretación geométrica

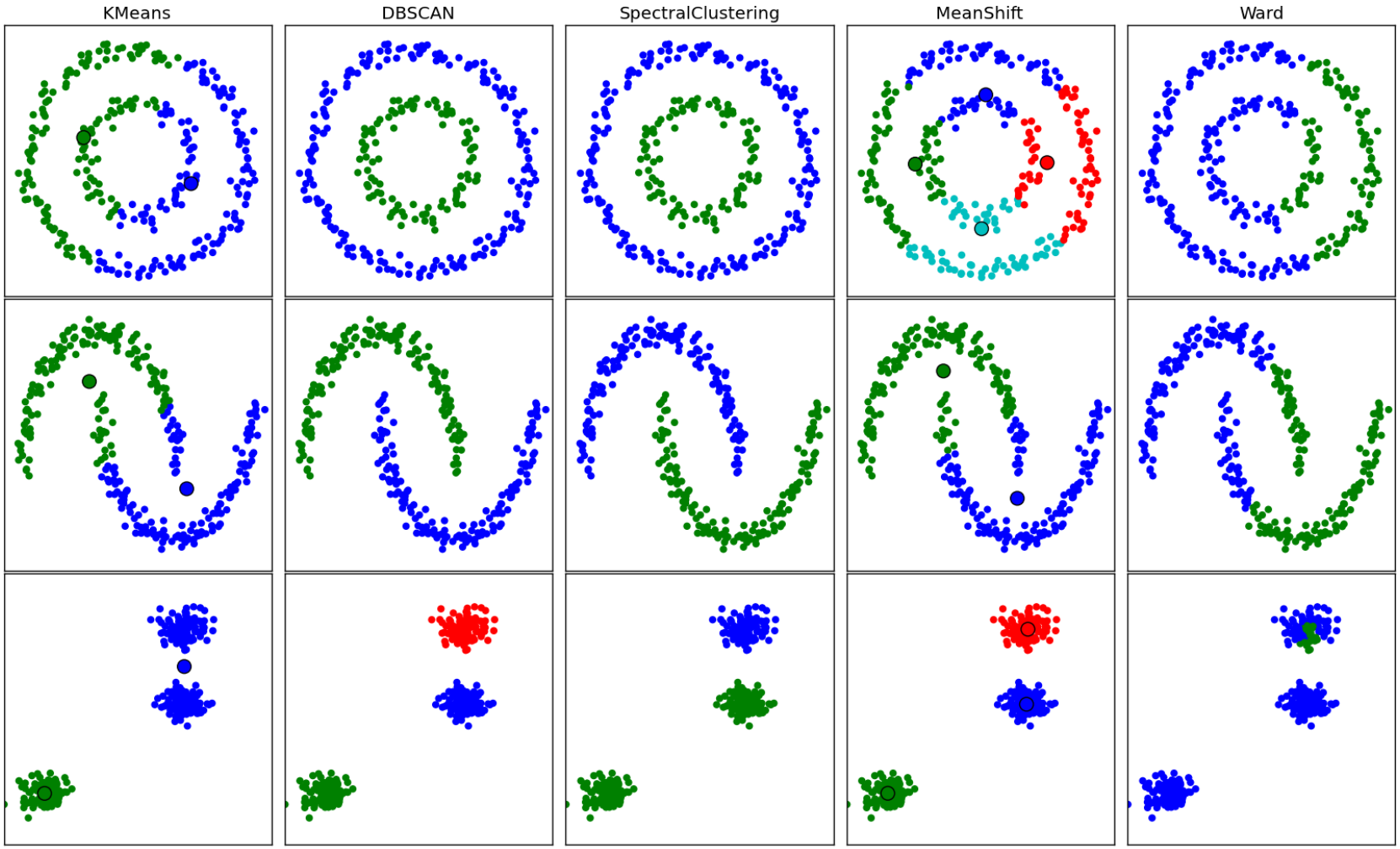


Clustering en 2 dimensiones: `http://en.wikipedia.org/wiki/Cluster_analysis`

. . . pero

- En general cada observación será un punto en $\Re^n$
- Visualmente resulta poco claro para $n > 2$
- Se considera una herramienta **exploratoria** para generar hipótesis.

Resultados



Comparación visual de algunos algoritmos de clustering:
`http://jaquesgrobler.github.io/Online-Scikit-Learn-stat-tut/modules/clustering.html`

Procedimiento

Pasos

1. Seleccionar una medida de **distancia**/similaridad adecuada.
2. Elegir la **técnica** de clustering: jerárquico, no jerárquico.
3. Elegir el **método/algoritmo** dentro de la técnica.
4. (Decidir el número de clusters)
5. **Interpretar** los resultados (en base a qué atributos se ha generado la división)

Clustering jerárquico

Tipos de clustering jerárquico

Ascendente o aglomerativo

- Paso 1: *N* clusters (cada elementos es un cluster).
- Paso 2: *N* - *1* clusters (se unen los dos elementos más próximos).
- ...
- Paso N: *1* cluster con todos los puntos.

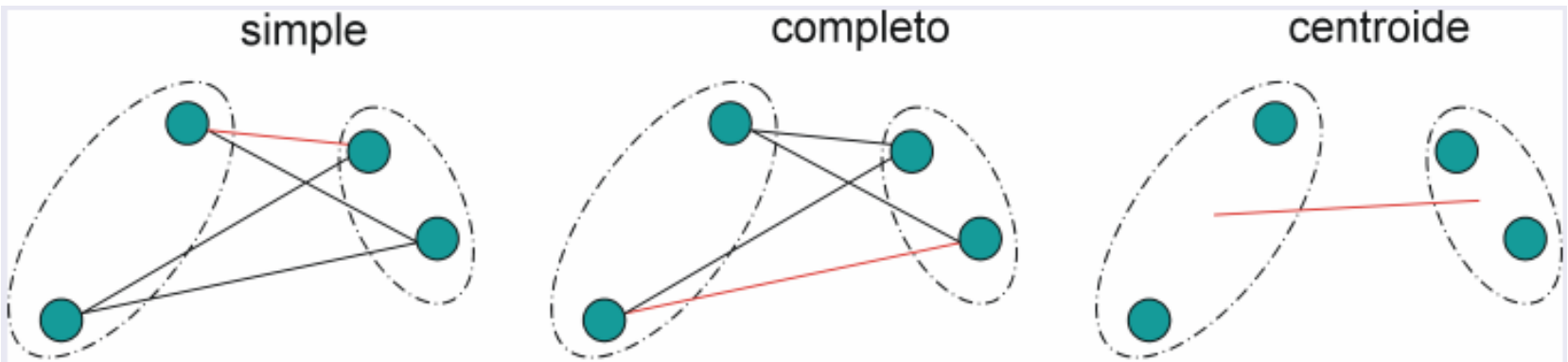
Descendente o divisivo

Es el proceso inverso al *clustering jerárquico ascendente*.

Distancia entre un elemento y un cluster

El cálculo de la distancia entre un punto y un cluster puede realizarse de diferentes formas:

- **Enlace simple** o vecino más próximo: mínima distancia entre todos los posibles pares de objetos en ambos clusters:
  $D(C, C') = min_{x \in C, x' \in C'} d(\mathbf{x}, \mathbf{x}')$
- **Enlace completo** o vecino más lejano: máxima distancia entre todos los posibles pares. $D(C, C') = max_{x \in C, x' \in C'} d(\mathbf{x}, \mathbf{x}')$
- **Enlace medio**: media de las distancias de todos los pares.
  $D(C, C') = \frac{1}{|C||C'|} \sum_{x \in C, x' \in C'} d(\mathbf{x}, \mathbf{x}')$



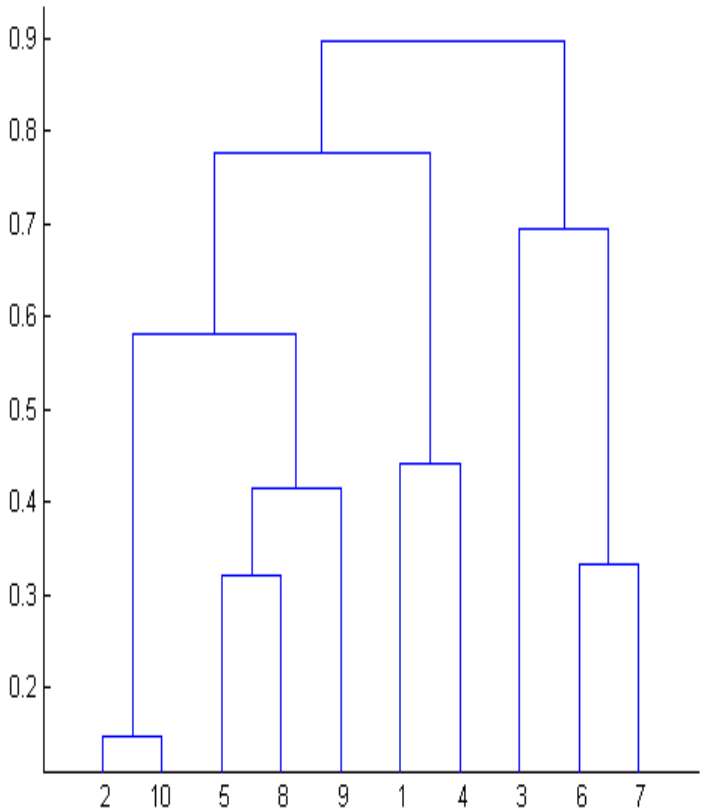
Distancia entre un elemento y un cluster

- **Centroide**: reemplazar cada cluster por su centroide (unitario) y calcular la distancia entre centroides. $c^j = \frac{1}{|C|} \sum_{x \in C} x_r^j, \quad r = 1, \ldots, n$
  $D(C, C') = d(c, c')$
- **Ward**: se calcula la suma total de desviaciones de la media de un cluster y trata de minimizarla. **No es una medida de distancia**.

Nota

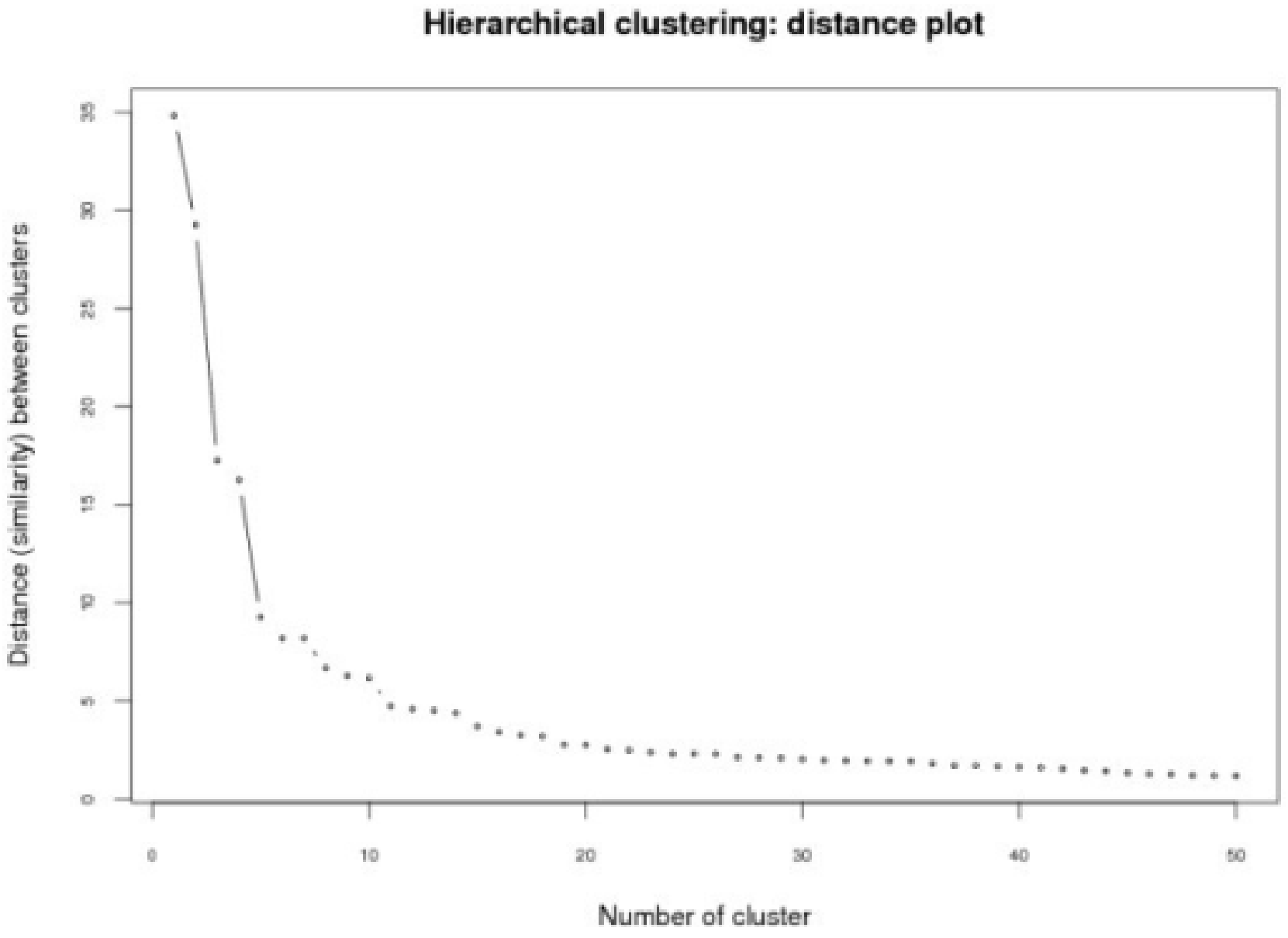
Problema: en algunos conjuntos de datos no es posible calcular el centroide (variables categóricas), entonces habrá que utilizar **prototipos**, elementos más próximos al *valor medio*.

Dendograma o árbol jerárquico



- 2, 10, 5, 8, 9, 1, 4, 3, 6, 7
- (2, 10), 5, 8, 9, 1, 4, 3, 6, 7
- (2, 10), (5, 8), 9, 1, 4, 3, 6, 7
- (2, 10), (5, 8), 9, 1, 4, 3, (6, 7)
- (2, 10), (5, 8, 9), 1, 4, 3, (6, 7)
- (2, 10), (5, 8, 9), (1, 4), 3, (6, 7)
- (2, 10, 5, 8, 9), (1, 4), 3, (6, 7)
- (2, 10, 5, 8, 9), (1, 4), (3, 6, 7)
- (2, 10, 5, 8, 9, 1, 4), (3, 6, 7)
- (2, 10, 5, 8, 9, 1, 4, 3, 6, 7)

¿Dónde cortar?



Clustering particional

Definición del problema

Problema de optimización

- **Objetivo**: agrupar N objetos en k clusters disjuntos.
- . . . pero no se conoce el valor de k de antemano.
- El número de posibles agrupaciones con N objectos en k grupos es (número de Stirling de segunda especie):

$$S(N,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\binom{k}{i}i^N$$

Número de Stirling de segunda especie: `http://en.wikipedia.org/wiki/Stirling_numbers_of_the_second_kind`

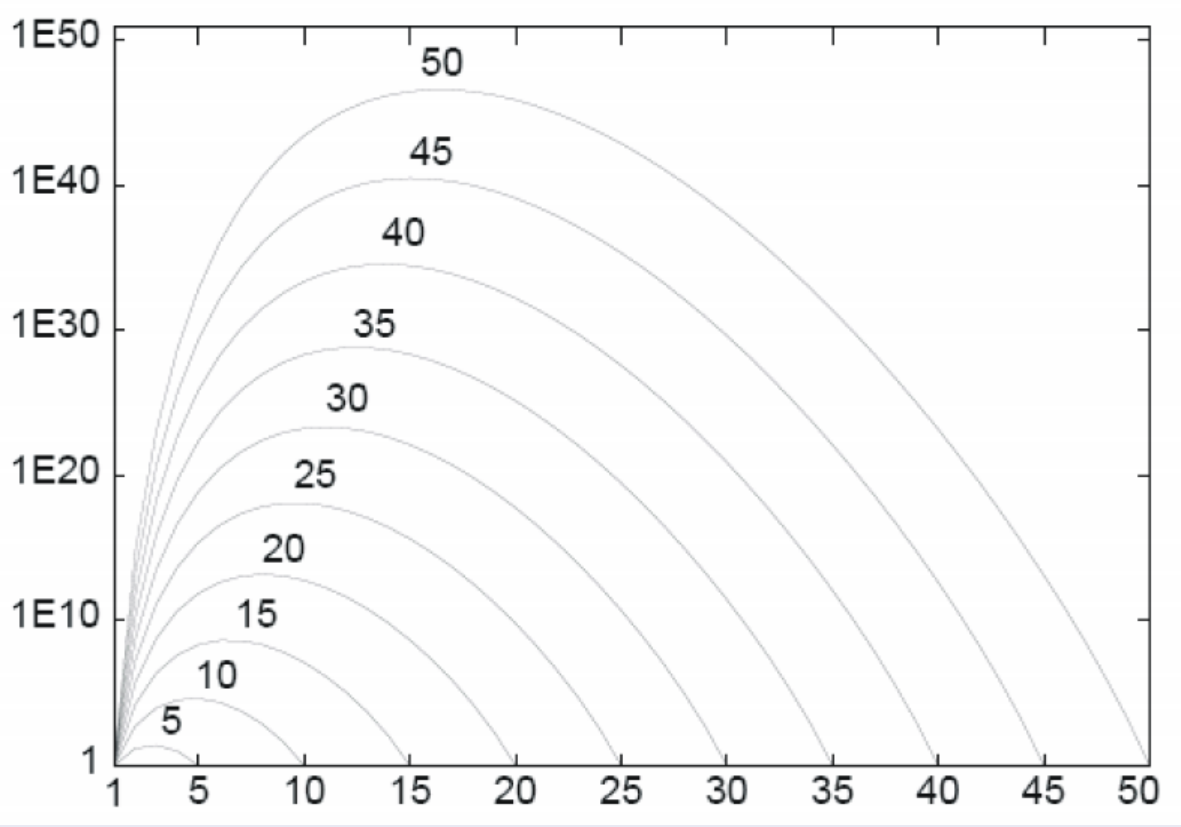
- El problema del clustering será minimizar la función J donde μ_i representa al prototipo de cada cluster y r_{ij} es una variable que vale 1 si x_j es asignado al cluster i y vale 0 en otro caso:

$$min_{r_{ij},\mu_i} J = \sum_{j=1}^{N}\sum_{i=1}^{k} r_{ij}\|x_j - \mu_i\|^2$$

Definición del problema

Complejidad de $S(N, k)$

Posibilidades (eje Y) con $N = 5, 10, 15, \ldots, 50$ objetos (sobre la curva) en $k = 1, 2, \ldots$ clusters (eje X):



Procedimiento

Pasos

1. Seleccionar ***k*** semillas iniciales (centroides/prototipos).
2. Asignar cada objeto al cluster más cercano.
3. Actualizar los prototipos.
4. Repetir hasta convergencia.

Criterios

- Cómo seleccionar las semillas iniciales
- Cómo actualizar los prototipos

Procedimiento

Paso 1: Selección de las semillas iniciales

- Al azar
- Los primero k objetos
- Basadas en conocimiento previo (experto)
- Heurística que busque elementos lo más alejados posible
- Primera semilla al azar, la segunda debe estar alejada una distancia data, etc. . .

Paso 2: Asignación al cluster más cercano

- Elección de la medida de distancia

Procedimiento

Paso 3: Actualización de los prototipos

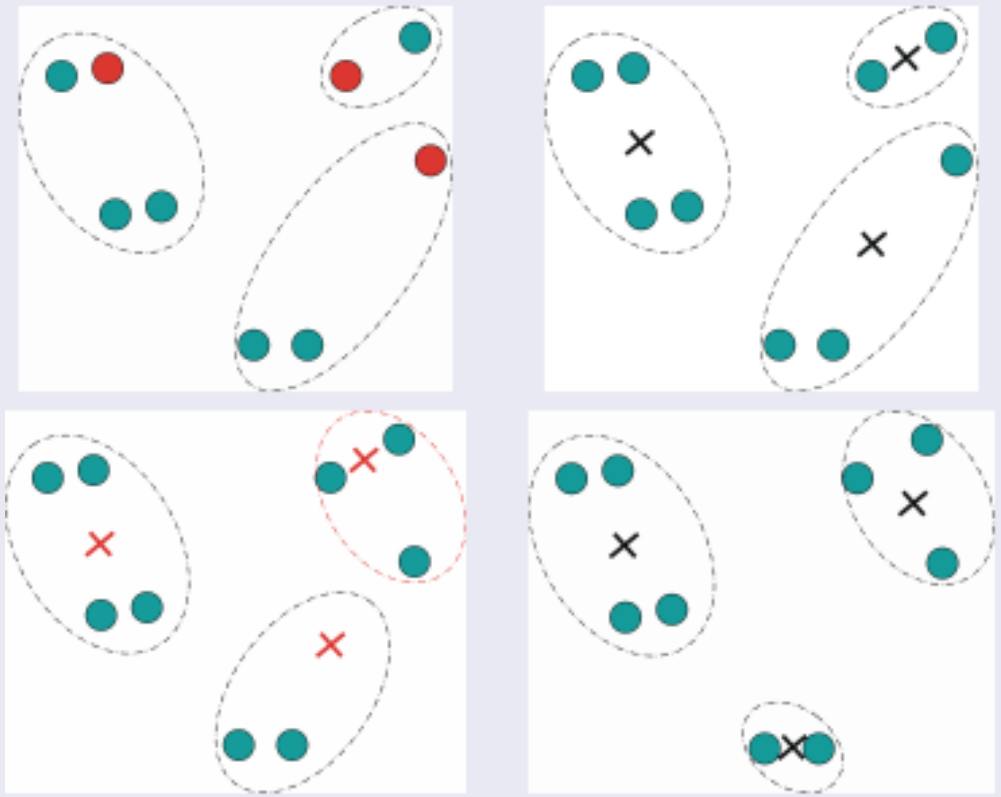
- **Forgy (1965)**: una vez que se han asignado todos los objetos a un cluster, actualizar los centroides y recalcular.
- **MacQueen (1967)**: recalcular los centroides cada vez que se asigna un objeto (del que provenía y al que va ahora)
- Otros: asignar los objetos intentando minimizar algún criterio estadístico (varianza dentro del cluster).

Paso 4: Convergencia

Continuar hasta que los centroides no se muevan

Ejemplos

Forgy



Ejemplo

k-medias - Segmentación de imágenes (Bishop, 2006)

Clustering probabilista

Introducción

Hipótesis

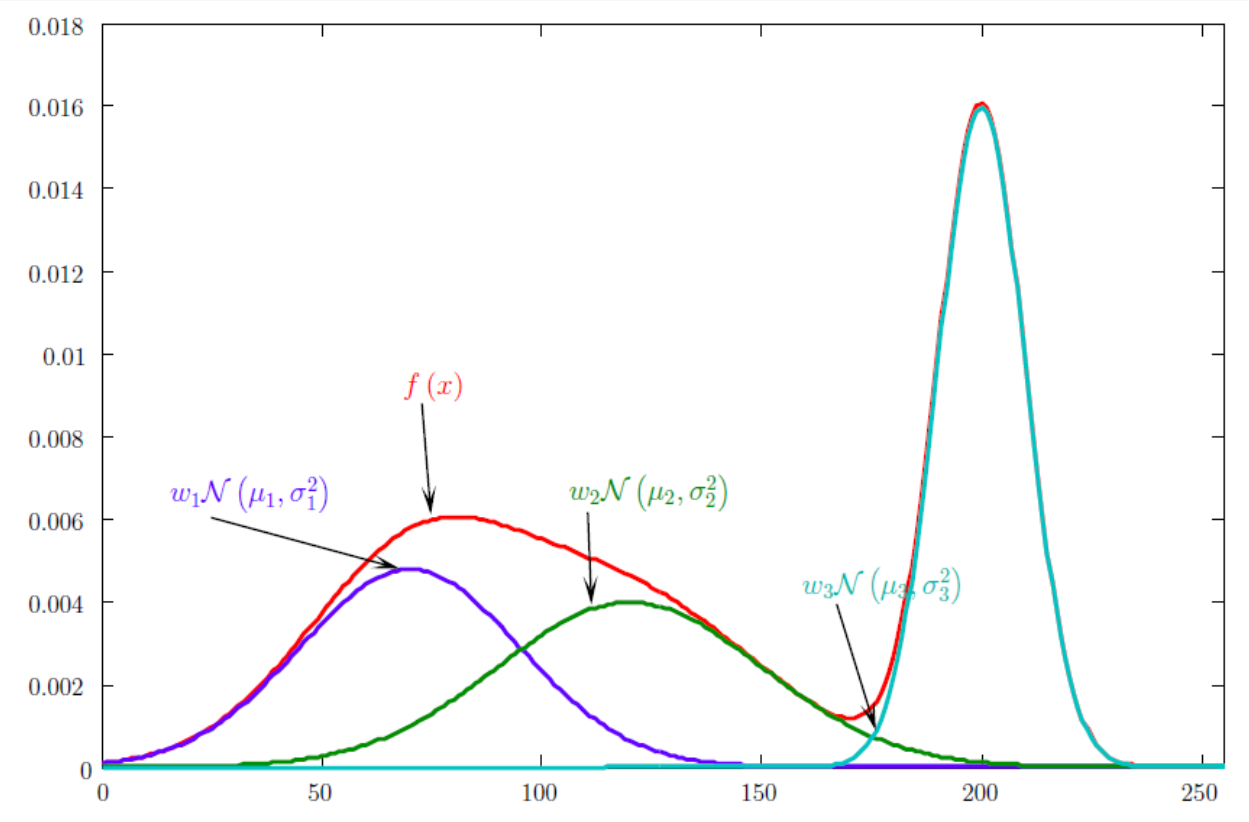
- Los datos provienen de una **mixtura de k distribuciones** de probabilidad, una para cada cluster.
- Cada distribución da la probabilidad de que un objeto pertenezca a un cluster (**soft clustering**).
- Obviamente cada objeto en realidad pertenece a un único cluster, pero **no podemos saber a cual**.

Objetivo

- Encontrar el conjunto de distribuciones más probable dados los datos.

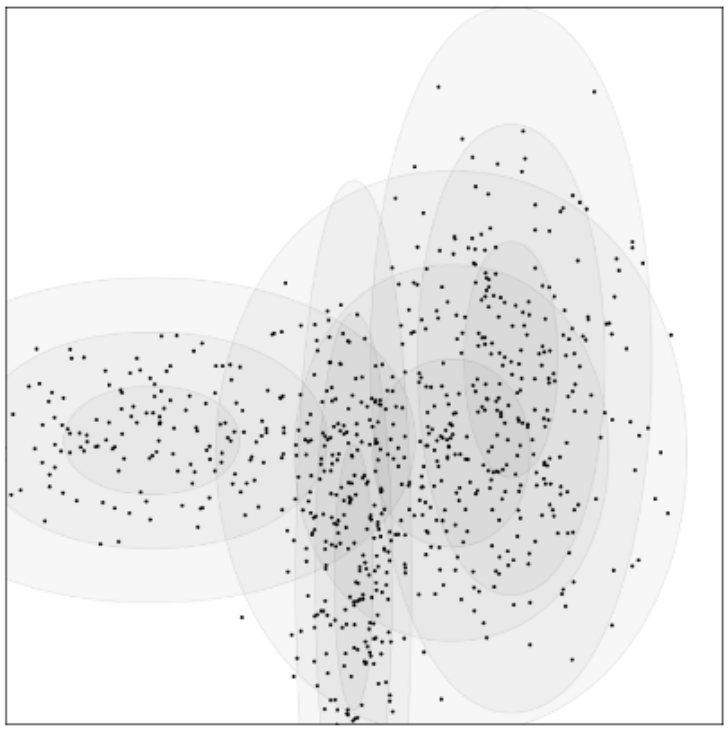
Introducción

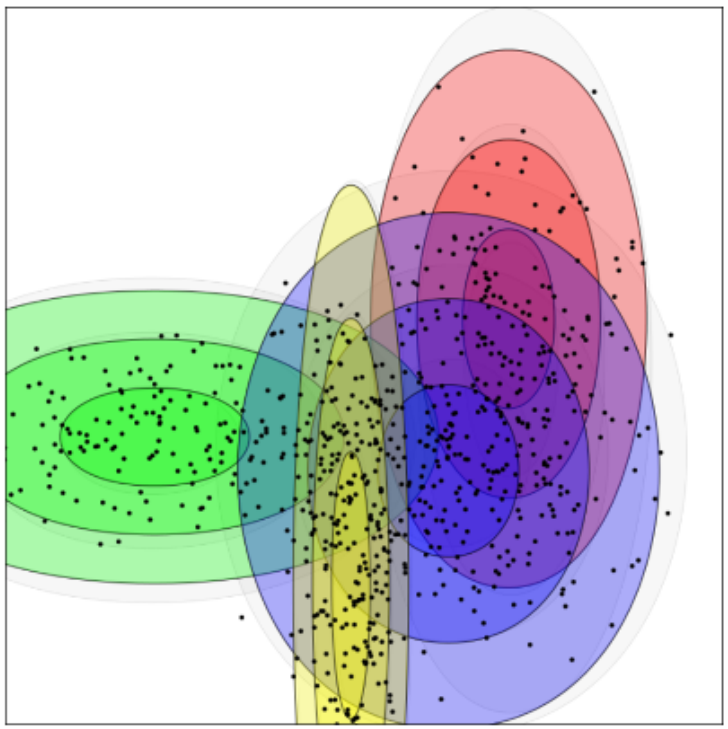
Ejemplo: Mixtura de Gaussianas (una variable)



Introducción

Ejemplo: Mixtura de Gaussianas (dos variables)





. . . pero en el caso general tenemos n variables (dimensiones) y podemos utilizar k clusters.

Modelo

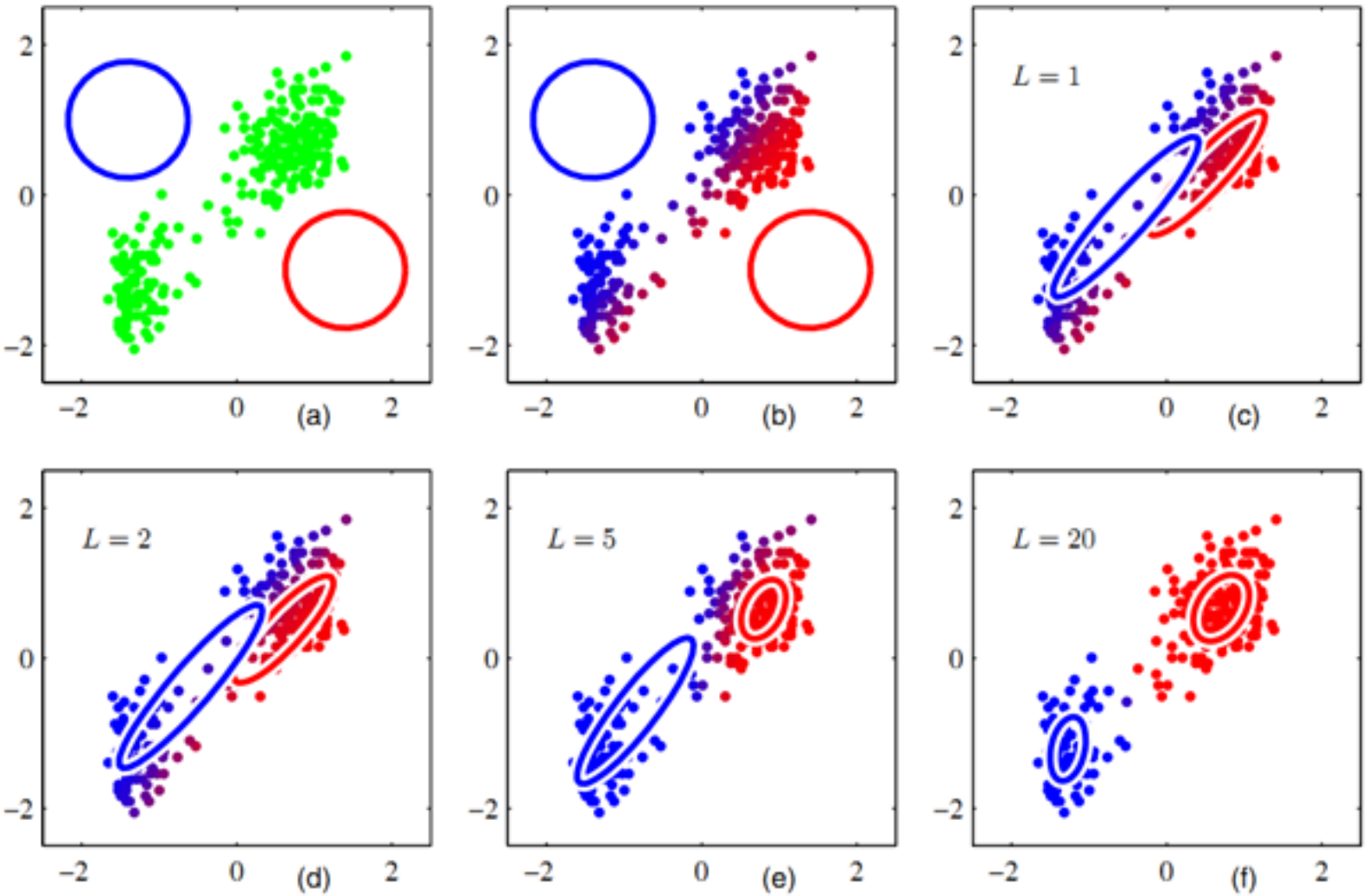
Mixtura de Gaussianas

- **Combinación lineal de k gaussianas**:

$$f(x) = \sum_{i=1}^{k} \pi_i \mathcal{N}(\mathbf{x}|\mu_\mathbf{i}, \ ^\circ_\mathbf{i})$$

- Los parámetros del modelo serán π_i, μ_i, $\Sigma_i \quad i = 1, \ldots, k$, y los podemos estimar por **máxima verosimilitud** $==>$ métodos numéricos iterativos o **algoritmo EM**.
- El algoritmo EM no garantiza encontrar el óptimo global, así que habrá que repetir con distintas inicializaciones y tomar la mejor.

Mixtura de Gaussianas



Conclusiones

Técnica: ¿jerárquico o particional?

Jerárquico

- No requiere la elección de semillas iniciales.
- El dendograma puede ser muy útil.
- No permite reasignaciones.
- Sensible a la medida de distancia elegida.

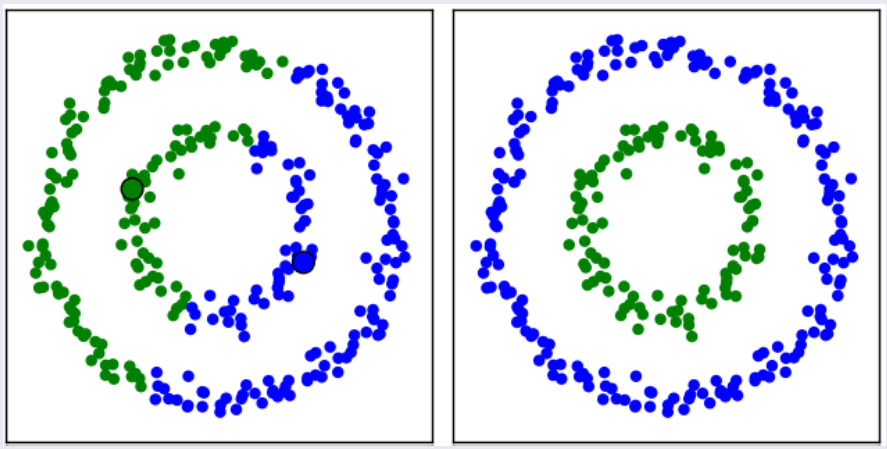
No jerárquico

- **Muy sensible** a las semillas iniciales.
- Partitional: tomar como entrada la solución del jerárquico.
- Probabilista: inicializar con los clusters del partitional.

Método

Consideraciones

- Ruido en los datos.
- Usar varios métodos y comparar
- El cluster jerárquico tiene **efecto encadenamiento** (meterse en clusters que ya existen, más que agruparse en nuevos)
  - grave si ocurre al principio
  - el enlace simple es muy susceptible a este ¿problema?

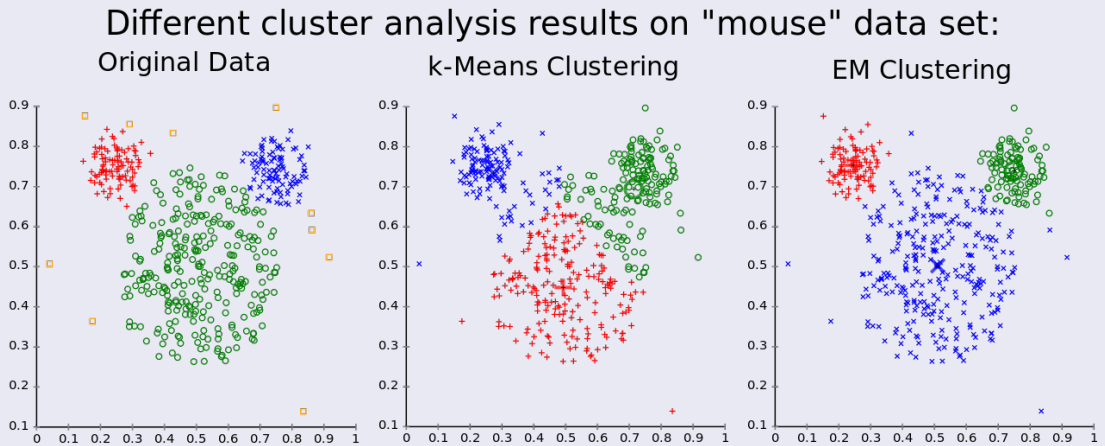


- el enlace completo (y Ward) identifica clusters compactos y está menos afectado por outliers que el enlace simple.

Método

k-medias

- Funciona bien cuando los clusters son compactos y bien separados (hiperesferas).
- Es fácil asignar nuevos objectos a los clusters existentes (el espacio queda particionado en polígonos de Thiessen).
- Sensible a la **escala**, ruido y **outliers**.
- Utilizar una **estrategia** multicomienzo y escoger la que minimiza J.



Medida de distancia

Distancia de Minkowski

$$d(\mathbf{x}, \mathbf{x}') = (\sum_{r=1}^{n} |x_r - x'_r|^p)^{1/p}$$

- Distancia **euclídea** para $p = 2$, **Manhattan** para $p = 1$, Tchebychev o norma del supremo para $p = \infty$

Problema

- Las **unidades** de medida influyen en el resultado del clustering (la variable con mayores unidades es dominante)
- Soluciones:
  - Medir cada variable en escalas comparables
  - Estandarizar los datos
  - Usar una medida de distancia que elimina este problema: **Mahalanobis**

$$d(\mathbf{x}, \mathbf{x}') = ((\mathbf{x} - \mathbf{x}')^T \Sigma^{-1} (\mathbf{x} - \mathbf{x}'))^{1/2}$$ || Σ es la matriz varianza-covarianza

Medida de distancia

Datos cualitativos

- **Datos cualitativos**: sustituir el concepto de distancia por el de **similitud**.
  - ¿Se puede calcular un centroide? Hablar de prototipos
  - Edit-distance, correlación, coseno del ángulo, información mutua, *ad-hoc* para el problema,...

Resultados

Evaluación

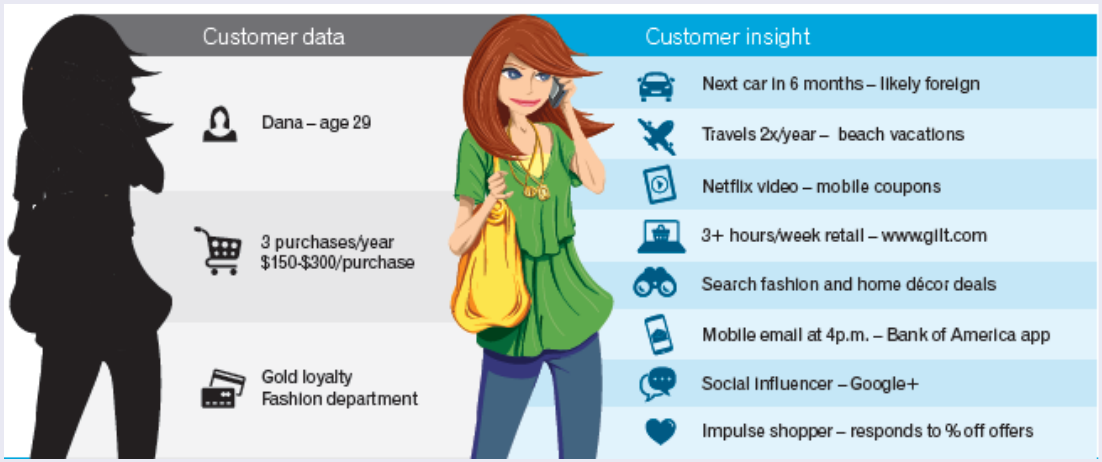
- Valor de la función J que tratábamos de minimizar.
- Distancias grandes entre centroides.
- Validación externa (auditor, analista,...)
- Evaluar los clusters respecto a cada variable. Eliminar alguna variable y repetir el análisis.

Interpretación de la solución

- Etiquetar los clusters utilizando sus centroides.
- Si los centroides son muy distintos en cierta variable, se puede utilizar ésta para etiquetar.

Resultados

¿Se parece la realidad a lo que obtenemos?



¡Muchas gracias!

@jgsogo

https://github.com/jgsogo/talks