Aprendizaje Automático

Aprendizaje No Supervisado

Viviana Cotik

1er cuatrimestre 2019

Avance

1er parte

- Introducción, Datos, Sesgos de Datos, Aprendizaje de conceptos, Sesgo Inductivo, Árboles de decisión, Naive Bayes, Evaluación de algoritmos

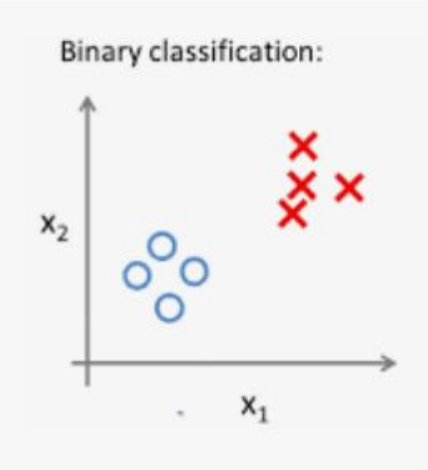
2da parte

- Aprendizaje no supervisado
- Ensambles
- Aprendizaje por refuerzo
- Redes Neuronales
- Algunas aplicaciones

Tipos de aprendizaje automático

Aprendizaje automático:

- **supervisado**:
  - requiere **instancias etiquetadas** para entrenamiento
  - **regresión, clasificación**



Tipos de aprendizaje automático

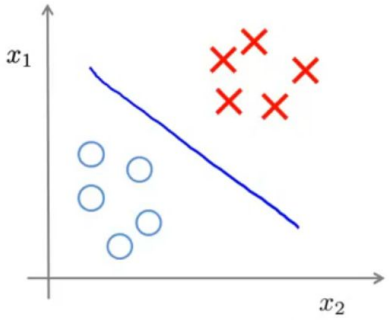
Aprendizaje automático:

- **supervisado:**
  - requiere **instancias etiquetadas** para entrenamiento
  - regresión, clasificación
- **no supervisado:**
  - las **instancias no** están **etiquetadas**
  - se usa para **visualizar** los datos, **entenderlos**, **resumirlos**
  - **clustering**, **reducción de la dimensión** (PCA, T-SNE, MDS, ISOMAP)

Otros:

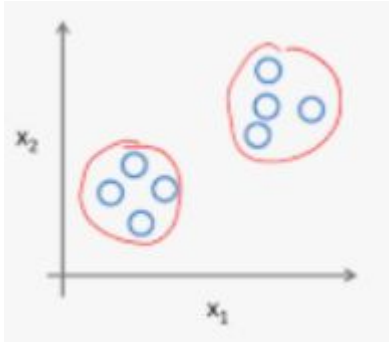
- **aprendizaje por refuerzos**

Aprendizaje supervisado vs. no supervisado



Supervisado

- **$\{(x^{(1)}, c(x^{(1)})), (x^{(2)}, c(x^{(2)})), \dots, (x^{(m)}, c(x^{(m)}))\}$**
- **Objetivo:** encontrar una hipótesis que satisfaga los datos



No supervisado

- **$\{x^{(1)}, x^{(2)}, \dots x^{(m)}\}$**
- **Objetivo:** que el algoritmo encuentre cierta estructura

Fuente: curso ML Stanford. Coursera. Andrew Ng

Algoritmos de Aprendizaje Supervisado

Datos etiquetados

- **Árboles de decisión**
- **Naive Bayes**
- LDA (Linear Discriminant Analysis) (AID)
- SVM (Support Vector Machines) (AID)
- Regresión logística (Enfoque estadístico)
- KNN (k - nearest neighbors)
- NN (RN, Redes Neuronales Artificiales) (también hay no supervisadas)
- ...
- Ensambles (combinación de modelos)

Aprendizaje No Supervisado

Datos no etiquetados

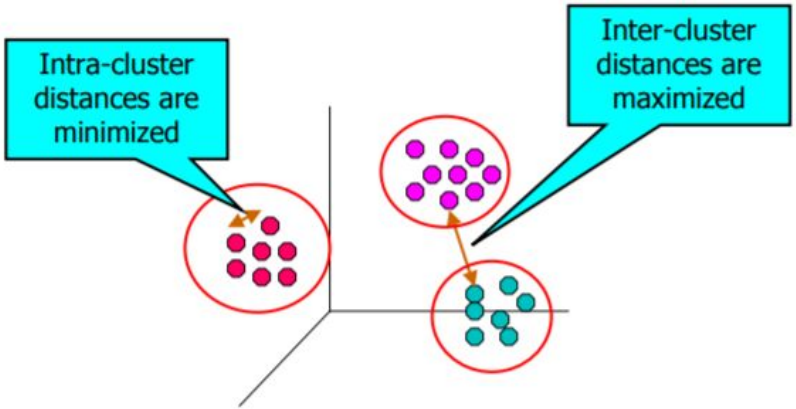
- Clustering: para encontrar patrones ocultos. Entender, resumir.
- Reducción de la dimensionalidad (PCA -principal component analysis- y otros)

Clustering

Encontrar **grupos de instancias (clusters)** a partir de **información en los datos** que **describan objetos** y **sus relaciones**.

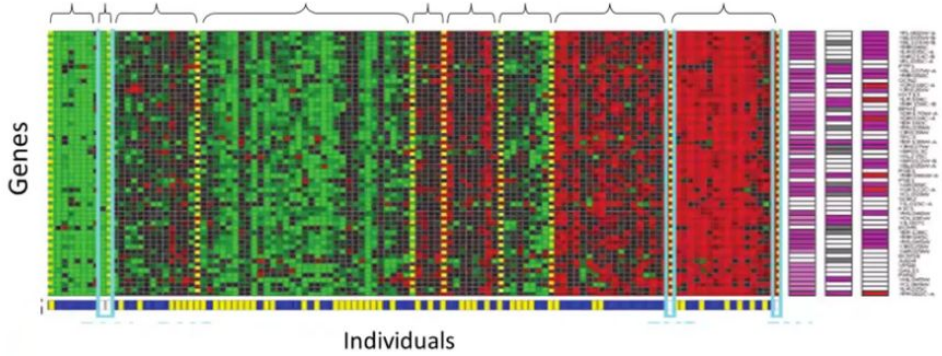
Instancias de un cluster tienen que ser:

- similares entre sí y
- diferentes a las de otros clusters



Tan, Steinbach & Kumar, Introduction to Data Mining
https://www-users.cs.umn.edu/~kumar001/dmbook/dmslides/chap8_basic_cluster_analysis.pdf

Clustering: aplicaciones



Fuente: curso ML Stanford

Análisis de redes sociales

uente: Wikimedia commons

Segmentación del mercado.
Fuente: internet

Algoritmos de clustering

Tipos de clustering:

- partición / jerárquicos
- exclusivos / no exclusivos

Algoritmos de clustering

- **De partición:** se clasifican **n datos** en **k clusters**. Cada cluster satisface requerimientos de una partición:
  - cada dato está en un y sólo un cluster
  - cada cluster debe tener al menos un dato
- **Jerárquicos**
  - **Aglomerativos (bottom up):** empiezan con n clusters y se combinan grupos hasta terminar en un cluster con n observaciones.
  - **Divisorios (top down):** comienzan con un cluster de n observaciones y en cada paso se dividen un cluster en dos hasta tener n clusters.

K-means (K-medias)

Un método muy popular. Es de **partición**.

Entrada:

datos no etiquetados ($\mathbf{x^{(1)}, x^{(2)}, \dots x^{(m)}}$), $x^{(i)}$ es un vector $\in R^n$

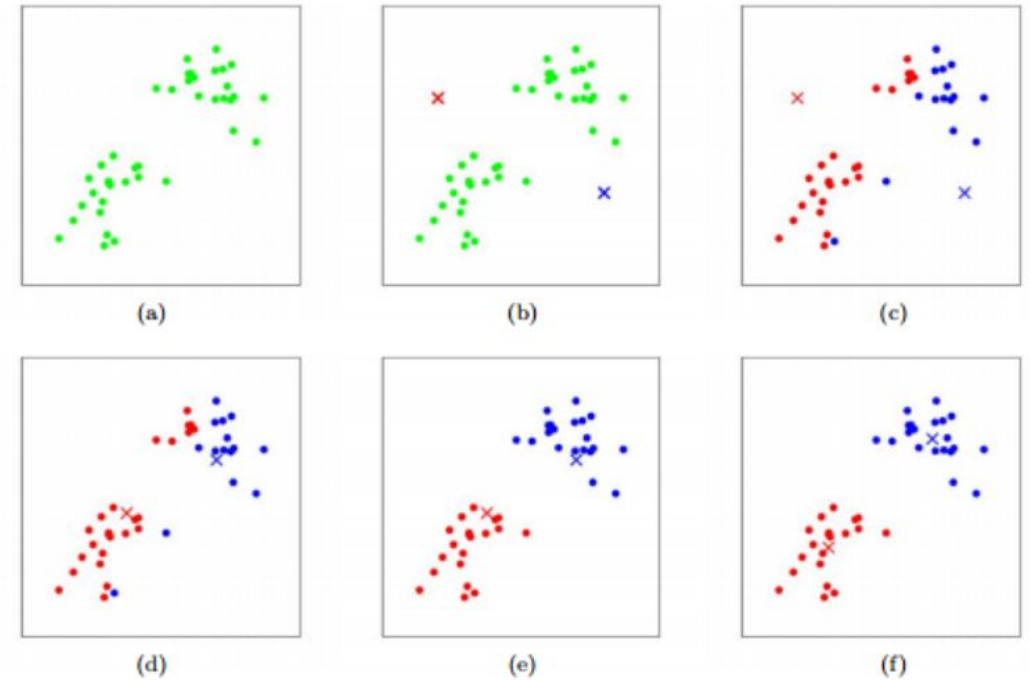
K: cantidad de clusters

Algoritmo:

- Inicializar aleatoriamente K centroides de los clusters μ1, ...μK $\in R^n$
- **Repetir**
  a. **asignación de cluster:** para cada dato se fija su distancia a cada centroide y es asignado al más cercano
  b. **movida de centroide:** tomar los centroides y moverlos a la posición promedio de los puntos de cada color

  hasta que los centroides no se muevan

K-means



Puntos: datos de entrenamiento
Cruces: centroides de los clusters

(a) Conjunto de datos original
(b) Asignación aleatoria de centroides
(c-f) dos iteraciones de k-means:

- **asignación de cluster** -datos pintados del mismo color del centroide-,
- **movida de centroides** -a la media de los puntos asignados a este-

Stanford CS 221
https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

K-means

Si un cluster no tiene puntos:

- se elimina el centroide o
- se ubica nuevamente el centroide al azar

K-means

Optimización, función de costo a minimizar. Función de **distorsión**.

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} a_{ik} \cdot ||x^{(i)} - \mu_k||^2$$

K: nro de clusters, m: cant. datos

μ_k: centroide de cluster k

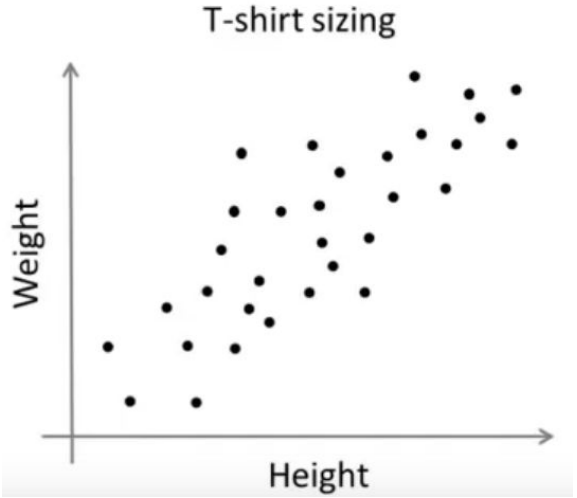
$x^{(i)}$: dato nro i

a_{ik} { 1 si $x^{(i)}$ está asignado al cluster k
{ 0 en otro caso

K-means intenta encontrar μ_k y a_{ik} que minimicen J

- En **asignación de cluster**: minimiza J con respecto a a_{ik} (asignando los puntos al centroide más cercano, que está fijo)
- En **movida de centroide**: minimiza J con respecto a μ_k

K-means



Fuente: curso ML Stanford

Los problemas **no necesariamente están bien separados en clusters**.

Además, muchas veces la **dimensión > 3** (técnicas de reducción de la dimensionalidad)

K-means - Inicialización

K: cantidad de clusters
m: cantidad de datos

Requisito: K<m

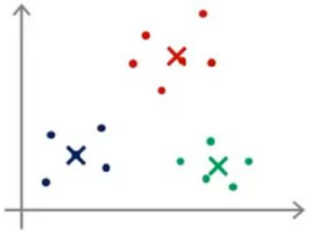
Inicialización de centroides:

- al azar
- **elegir K datos cualesquiera**

K-means - Inicialización

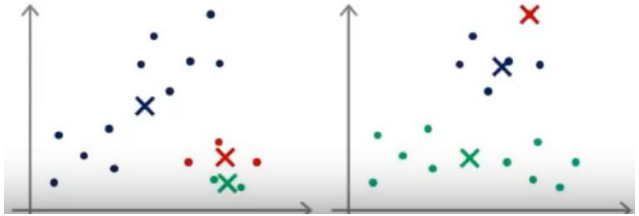
K: cantidad de clusters
m: cantidad de datos

Requisito: K<m



Inicialización de centroides:

- al azar
- **elegir K datos cualesquiera**



Puede converger a distintas soluciones dependiendo de cómo lo inicializo

K-means - Inicialización

Inicialización de centroides:

- al azar
- elegir K datos cualesquiera
- **múltiples inicializaciones al azar (para evitar óptimos locales)**

  Hacer entre 50 y 1000 veces
  - **inicializar centroides**
  - **correr k-means y calcular la función de costo**

  elegir el clustering que tuvo la menor función de costo

Útil en casos con K chicos (<10)

- usar k clusters de un método jerárquico e inicializar con sus centroides
- ...

Distancias

Desde un punto de vista formal, para un conjunto de elementos X se define **distancia** o **métrica** como cualquier función matemática o aplicación $d(a,b)$ de $X \times X$ en $\mathbb{R}$ que verifique las siguientes condiciones:

- No negatividad: $d(a,b) \geq 0 \; \forall a,b \in X$
- Simetría: $d(a,b) = d(b,a) \; \forall a,b \in X$
- Desigualdad triangular: $d(a,b) \leq d(a,c) + d(c,b) \; \forall a,b,c \in X$
- $\forall x \in X : d(x,x) = 0$.
- Si $x,y \in X$ son tales que $d(x,y) = 0$, entonces $x = y$.

Distancias

Atributos numéricos

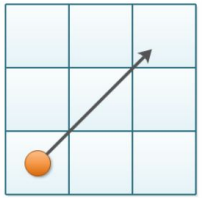
- distancia euclídea
- distancia de Manhattan
- distancia de Chebychev

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$\sum_{i=1}^{n}|x_i - y_i|$$

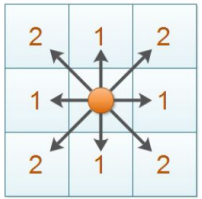
$$max_{i=1..n}|x_i - y_i|$$

Euclidean Distance



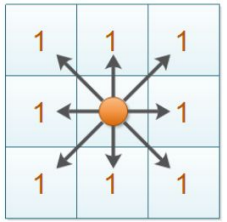
$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Manhattan Distance



$$|x_1 - x_2| + |y_1 - y_2|$$

Chebyshev Distance



$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

https://lyfat.wordpress.com/2012/05/22/euclidean-vs-chebyshev-vs-manhattan-distance/

Distancias

Atributos discretos:

Value Difference Metric (VDM)

Otras distancias:

Similaridad de coseno, Jackard distance (ambas para documentos), Hamming distance, Levensthein Distance (ambas para cadenas de caracteres)

H		O	N	D	A	
H	Y	U	N	D	A	I

https://dzone.com/articles/the-levenshtein-algorithm-1

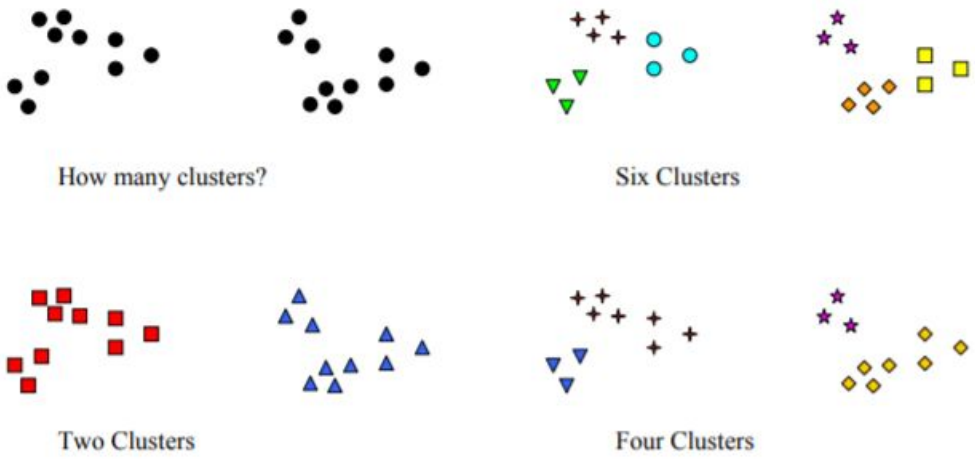
Distancia de Levenshtein: mínimo nro de ediciones de caracteres (agregado, borrado, sustitución) requeridos para cambiar una palabra por otra. Por ej. **para ADN**

DistanciaL (Honda, Hyundai) =3

K-means - Elección del K

¿Cómo elegimos el K? ¿Manualmente?

Ambigüedad



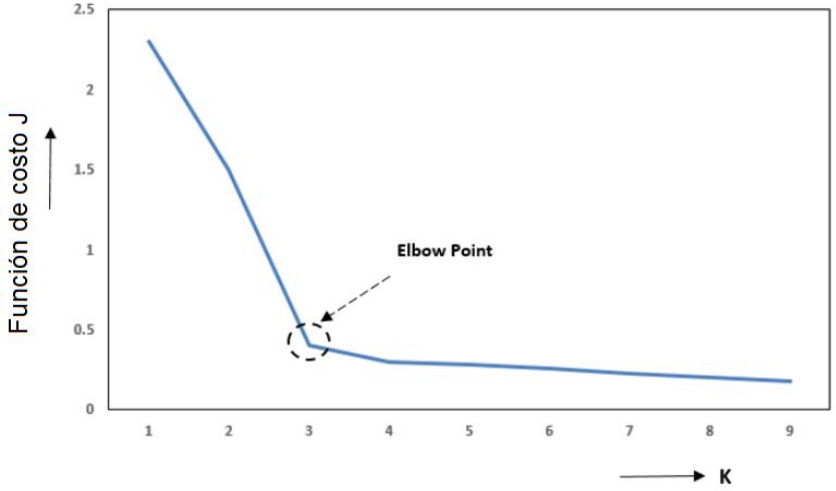
Tan, Steinbach & Kumar, Introduction to Data Mining. Cap 8

Elección del K

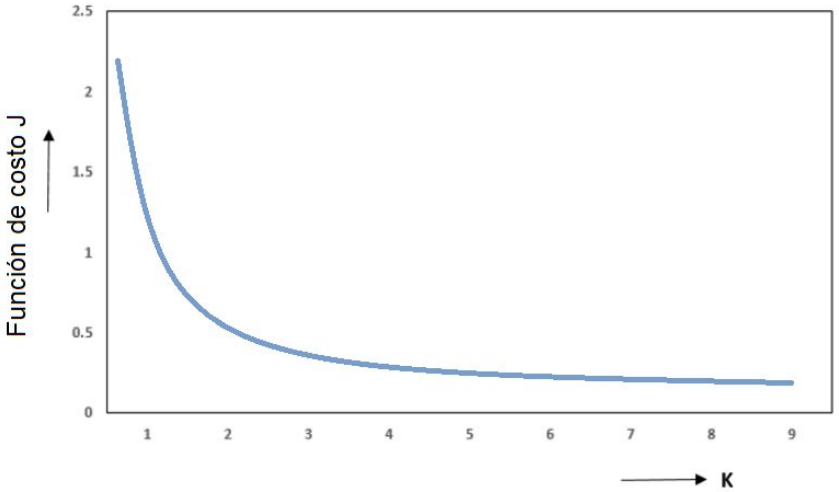
- **manualmente**. Problema Ambiguedad.
- **elbow method**
- **evaluar con una métrica** y ver cuán bien funciona para propósito posterior
  - (venta de remeras, K:3-5)
  - compresión de imágenes (cuán bien se ve, cuán comprimida está)

K-means - Elección del K

Elbow method



no siempre es útil...



K-means

Ventajas:

- algoritmo simple
- eficiente

Desventajas:

- sensible a la elección de los centroides iniciales
- sensible al ruido y a outliers
- hay que especificar el K

Expectation Maximization y Mezcla de Gaussianas

Gaussian Mixture Models

Asumimos un **overlap (soft clustering)**. Los elementos tienen una probabilidad de estar en distintos clusters.

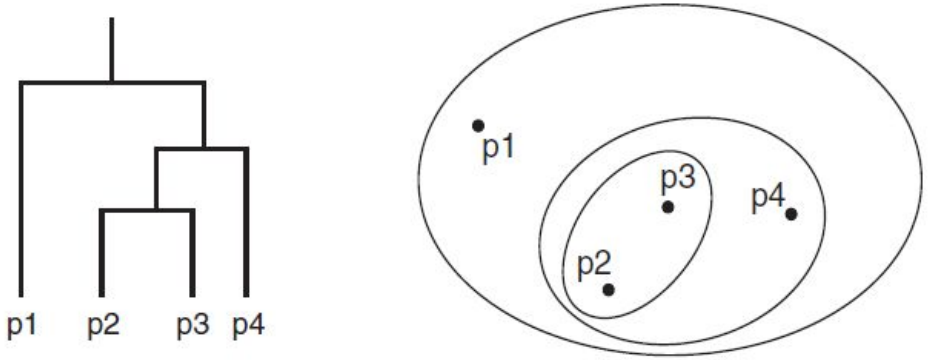
Cada cluster corresponde a una **distribución de probabilidades** (normal o Gaussiana) . Se quieren descubrir los parámetros: media y varianza

A diferencia de k-means computa la **probabilidad de que un elemento esté en distintos clusters**

Ej. de aplicaciones: reconocimiento del hablante.

Clustering jerárquico

Se suele mostrar en un **diagrama en forma de árbol**, llamado **dendograma**. Muestra clusters, subclusters y el orden en que fueron unidos.



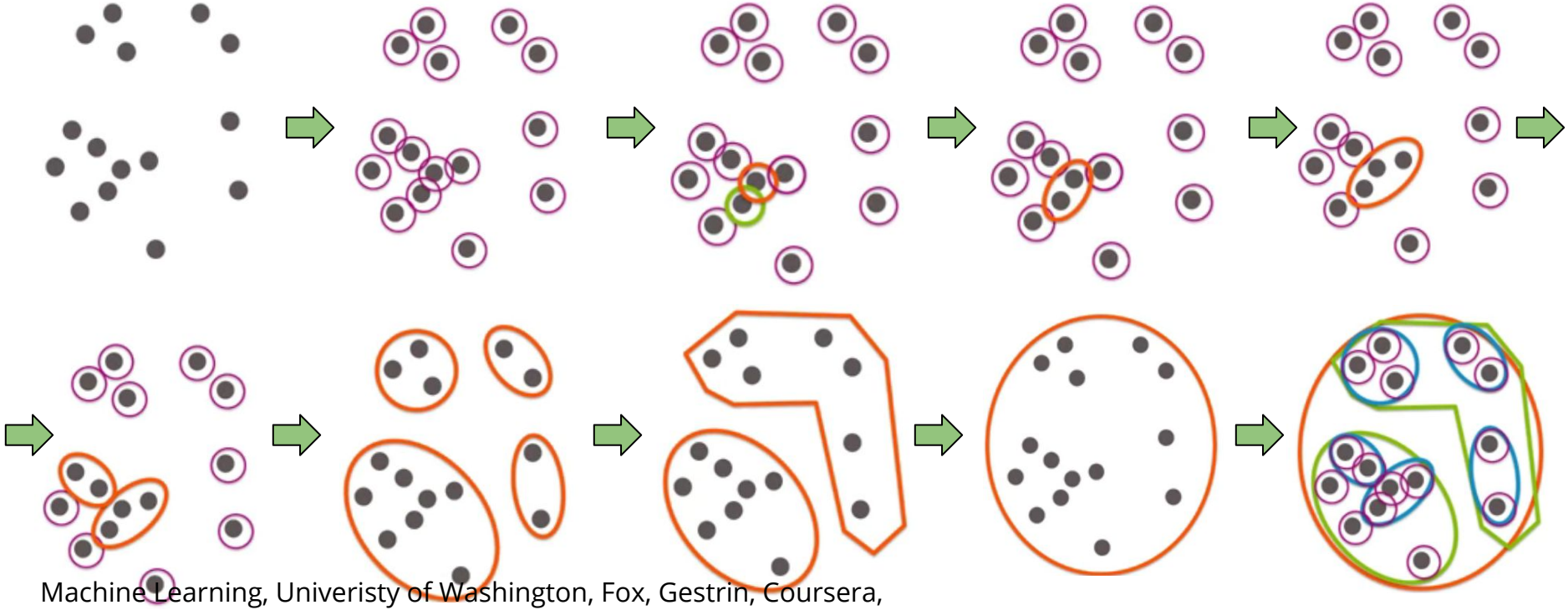
(a) Dendrogram. (b) Nested cluster diagram.

Tan, Steinbach & Kumar, Introduction to Data Mining (Cap. 8)

Clustering jerárquico

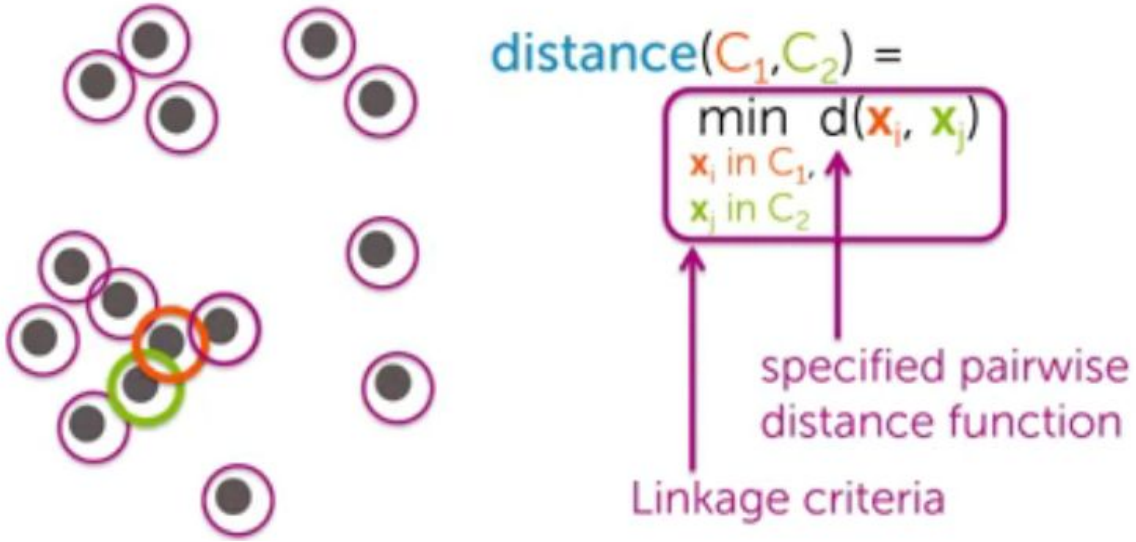
- **Tipos de clustering jerárquico**
  - **Aglomerativos (bottom up):** empiezan con **n clusters de un elemento** y se combinan grupos de a uno hasta terminar en un cluster con n observaciones.
  - **Divisorios (top down):** comienzan con **un cluster de n observaciones** y en cada paso se divide un cluster en dos hasta obtener n clusters de un elemento cada uno.

Aglomerativo: single linkage



Machine Learning, Univeristy of Washington, Fox, Gestrin, Coursera,
https://www.coursera.org/lecture/ml-clustering-and-retrieval/agglomerative-clustering-bsFBT

Aglomerativo: single linkage



Single linkage:

uno clusters de menor distancia.

distancia: distancia entre cualesquiera dos puntos más cercanos (pertenecientes a distintos clusters).

Machine Learning, Univeristy of Washington, Fox, Gestrin, Coursera,
https://www.coursera.org/lecture/ml-clustering-and-retrieval/agglomerative-clustering-bsFBT

Clustering Aglomerativo

1. Cada punto forma un cluster
2. Computar matriz de proximidad
3. Repetir:
    a. Buscar el par de clusters más similar y hacer un merge
    b. Actualizar la matriz de proximidad
4. hasta que haya un solo cluster

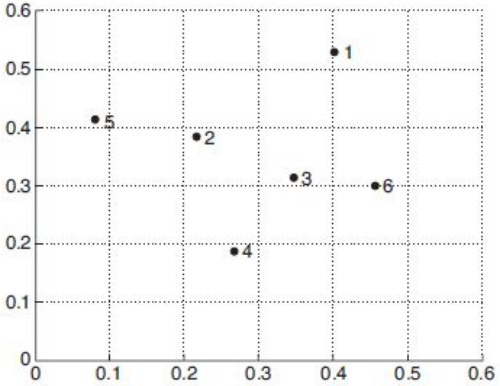


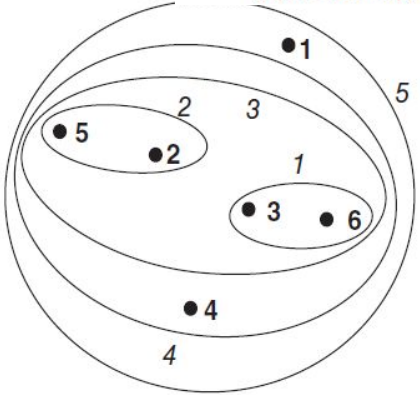
Figure 8.15. Set of 6 two-dimensional points.

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

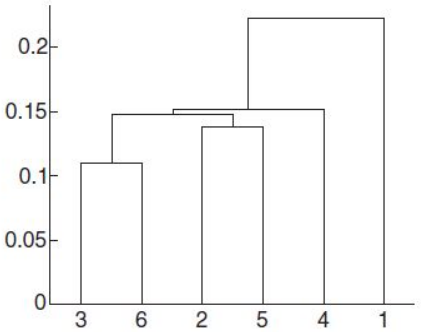
Table 8.4. Euclidean distance matrix for 6 points.

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Table 8.3. xy coordinates of 6 points.



(a) Single link clustering.

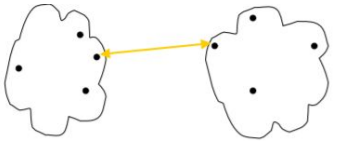


(b) Single link dendrogram.

$$
\begin{aligned}
dist(\{3,6\},\{2,5\}) &= \min(dist(3,2), dist(6,2), dist(3,5), dist(6,5)) \\
&= \min(0.15, 0.25, 0.28, 0.39) \\
&= 0.15.
\end{aligned}
$$

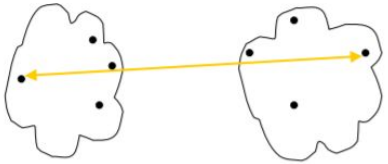
Medición de similaridad entre clusters

Cómo definimos similaridad entre clusters?



- **MIN (single linkage)**

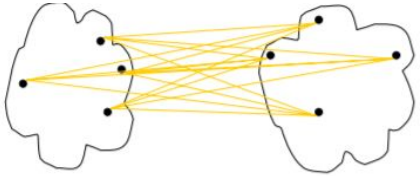
distancia mínima entre dos puntos de los dos distintos clusters.



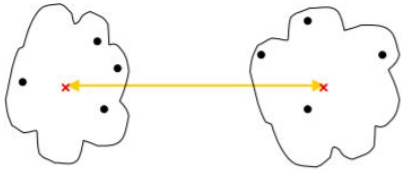
- **MAX (complete link)**

distancia máxima entre dos puntos de los distintos clusters

- **AVG** promedio de la distancia entre los puntos de los clusters



- distancia entre centroides



Tan, Steinbach & Kumar, Introduction to Data Mining
https://www-users.cs.umn.edu/~kumar001/dmbook/dmslides/chap8_basic_cluster_analysis.pdf

Clustering jerárquico

Ventajas:

- no asume ningún número de clusters (se pueden obtener cortando el dendograma en el nivel deseado)
- pueden corresponder a taxonomías (ej. reino animal)

Desventajas:

- Sensible a ruido y outliers
- Computacionalmente más caro en tiempo y en espacio

Resumen

- Aprendizaje supervisado vs. no supervisado
- Clustering y aplicaciones
- Algoritmos
  - k-means
  - EM
  - Aglomerativo: single linkage

Bibliografía

Capítulos de libros:

Tan, Steinbach & Kumar, Introduction to Data Mining. Cap 8

Otros:

https://towardsdatascience.com/supervised-machine-learning-classification-5e685fe18a6d
https://stanford.edu/~cpiech/cs221/handouts/kmeans.html
https://www.cs.princeton.edu/courses/archive/spring19/cos324/files/kmeans.pdf
http://axon.cs.byu.edu/~randy/jair/wilson2.html (distancias)