AGRUPAMIENTO (CLUSTERING)

Dr. Jorge Hermosillo
Laboratorio de Semántica Computacional

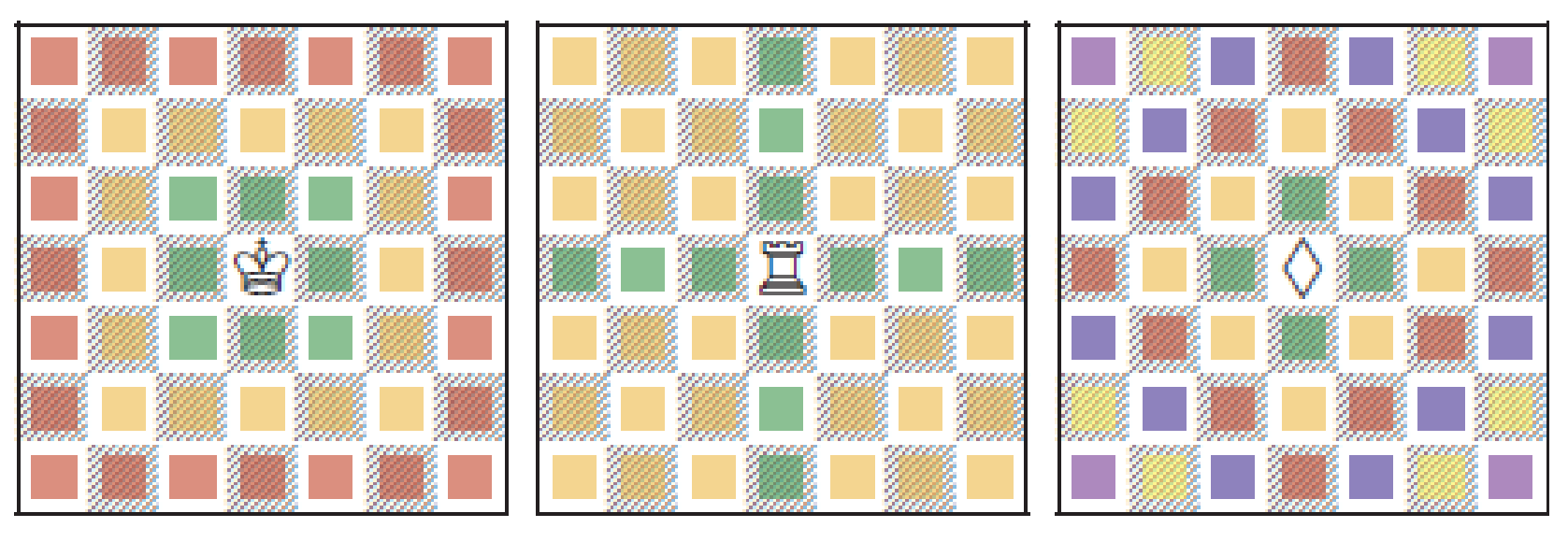
Centro de Investigación en Ciencias

Universidad Autónoma del Estado de Morelos

MEDIDAS DE DISTANCIA

DISTANCIA DE MINKOWSKI

- ¿De cuántas formas podemos medir la distancia entre dos puntos?



R T RT

DISTANCIA DE MINKOWSKI

- Si $\mathcal{X} = \mathbb{R}^d$, la distancia de Minkowski de orden $p > 0$ se define como:

$$\mathrm{Dis}_{\mathrm{p}}(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^{d} \left| x_j - y_j \right|^p \right)^{1/p} = \|\mathbf{x} - \mathbf{y}\|_p$$

donde $\|\mathbf{z}\|_p = \left(\sum_{j=1}^{d} \left| z_j \right|^p \right)^{1/p}$ es la norma-p (a veces denominada norma L_p) del vector $\mathbf{z}$.

- La norma-2 se refiere a la distancia Euclidiana:

$$\mathrm{Dis}_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{d} \left(x_j - y_j \right)^2} = \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})}$$

DISTANCIAS CON P VARIABLE

- Norma-1: distancia de Manhattan (experimentada por RT)

$$\mathrm{Dis}_1(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{d} |x_j - y_j|$$

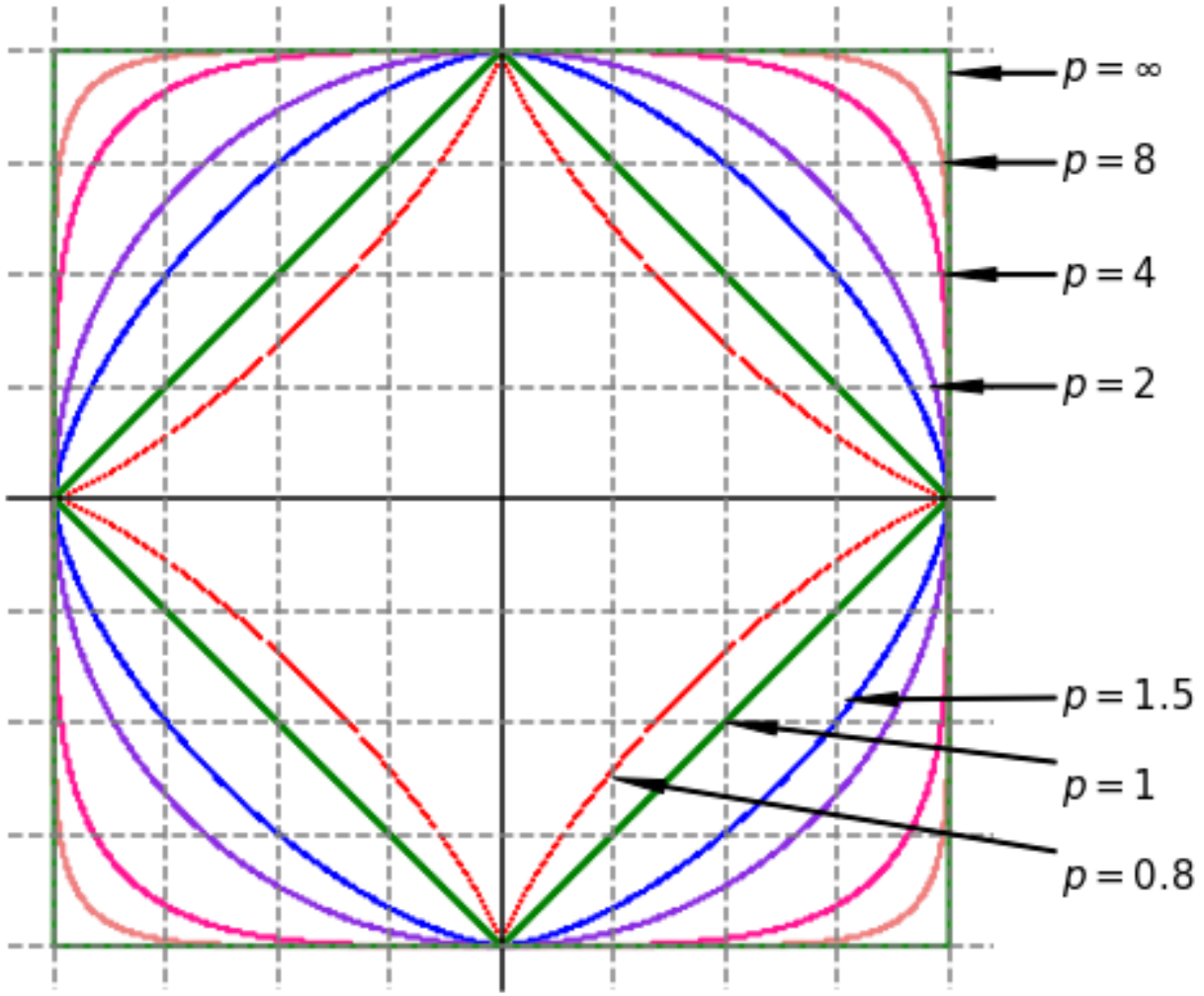
- Norma-∞ : Distancia de Chebyshev

$$\mathrm{Dis}_\infty(\mathbf{x},\mathbf{y}) = \max_j |x_j - y_j|$$

PUNTOS EQUIDISTANTES SEGÚN *p*

$$\mathrm{Dis}_{\mathrm{p}}(\mathbf{x},\mathbf{y}) = \left(\sum_{j=1}^{d} |x_j - y_j|^p\right)^{1/p}$$

- Líneas que conectan puntos a una distancia 1 de Minkowski de orden p respecto al origen.
- La única medida de distancia invariante a la rotación es la Euclidiana (p=2)



LA NORMA-0

- La ***norma-0*** (L_0) ***cuenta el número de elementos no nulos en un vector***.
- La ***distancia*** correspondiente ***cuenta el número de posiciones*** en las que dos vectores $\boldsymbol{x}$ y $\mathbf{y}$ difieren.
- No es una distancia de Minkowski, pero se puede definir como:

$$\mathrm{Dis}_0(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{d}(x_j - y_j)^0 = \sum_{j=1}^{d} \boldsymbol{I}\left[x_j \neq y_j\right]$$

entendiendo que $x^0 = 0$ para $x = 0$ y 1 de otra forma.

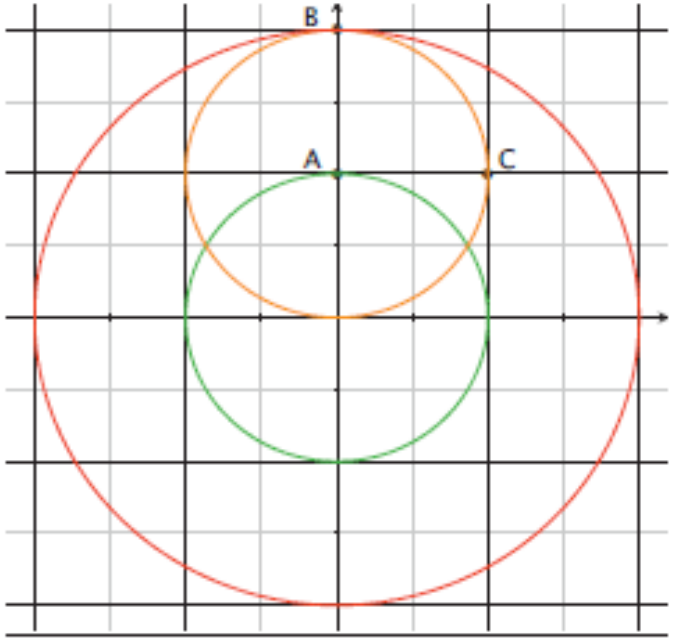
- Si $\boldsymbol{x}$ y $\mathbf{y}$ son binarias, se le llama *distancia de Hamming*.
- Para cadenas no binarias de longitud desigual esta distancia se puede generalizar en la *distancia de edición* o *distancia de Levenshtein*.

MÉTRICA DE DISTANCIA

- Dado un espacio de instancias $\mathcal{X}$, una métrica de distancia es una función Dis: $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ tal que para cualquier $x, y, z \in \mathcal{X}$:

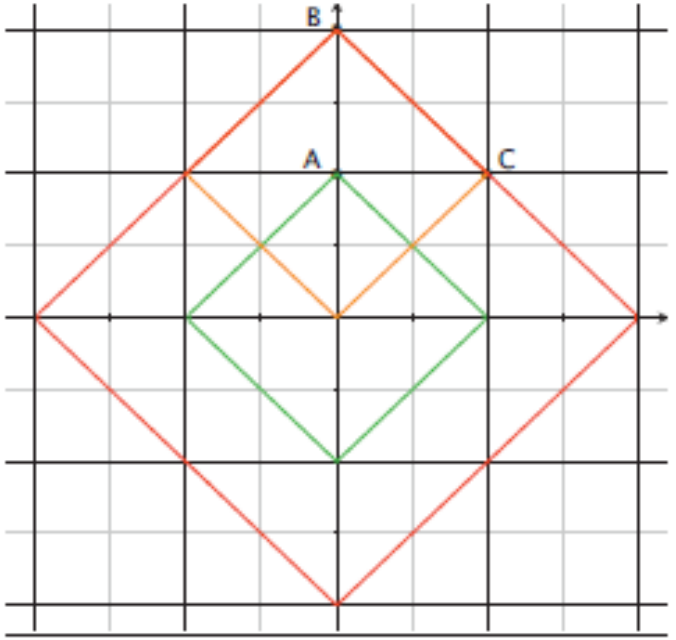
1. *Las distancias entre un punto y él mismo son 0:* $\mathrm{Dis}(x, x) = 0$
2. *Cualquier otra distancia es mayor a 0:* si $x \neq y$ entonces $\mathrm{Dis}(x, y) > 0$
3. *Las distancias son simétricas:* $\mathrm{Dis}(y, x) = \mathrm{Dis}(x, y)$
4. *Las desviaciones no pueden acortar la distancia:* $\mathrm{Dis}(x, z) \leq \mathrm{Dis}(x, y) + \mathrm{Dis}(y, z)$

MÉTRICAS VÁLIDAS E INVÁLIDAS SEGÚN p



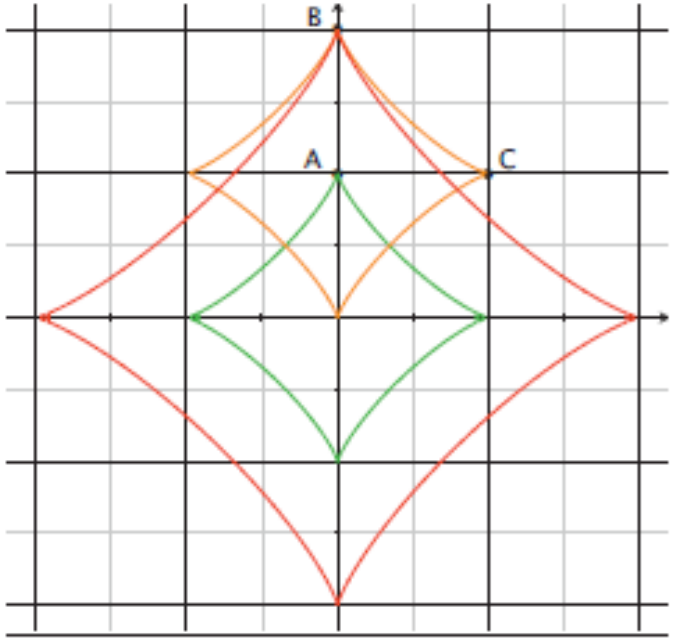
$$p = 2$$
$$\overline{AB} = \overline{AC}$$
$$\overline{OB} = \overline{OA} + \overline{AB} > \overline{OC}$$
$$\overline{OA} + \overline{AC} > \overline{OC}$$

✓



$$p = 1$$
$$\overline{AB} = \overline{AC}$$
$$\overline{OB} = \overline{OC}$$
$$\overline{OA} + \overline{AC} = \overline{OC}$$

✓

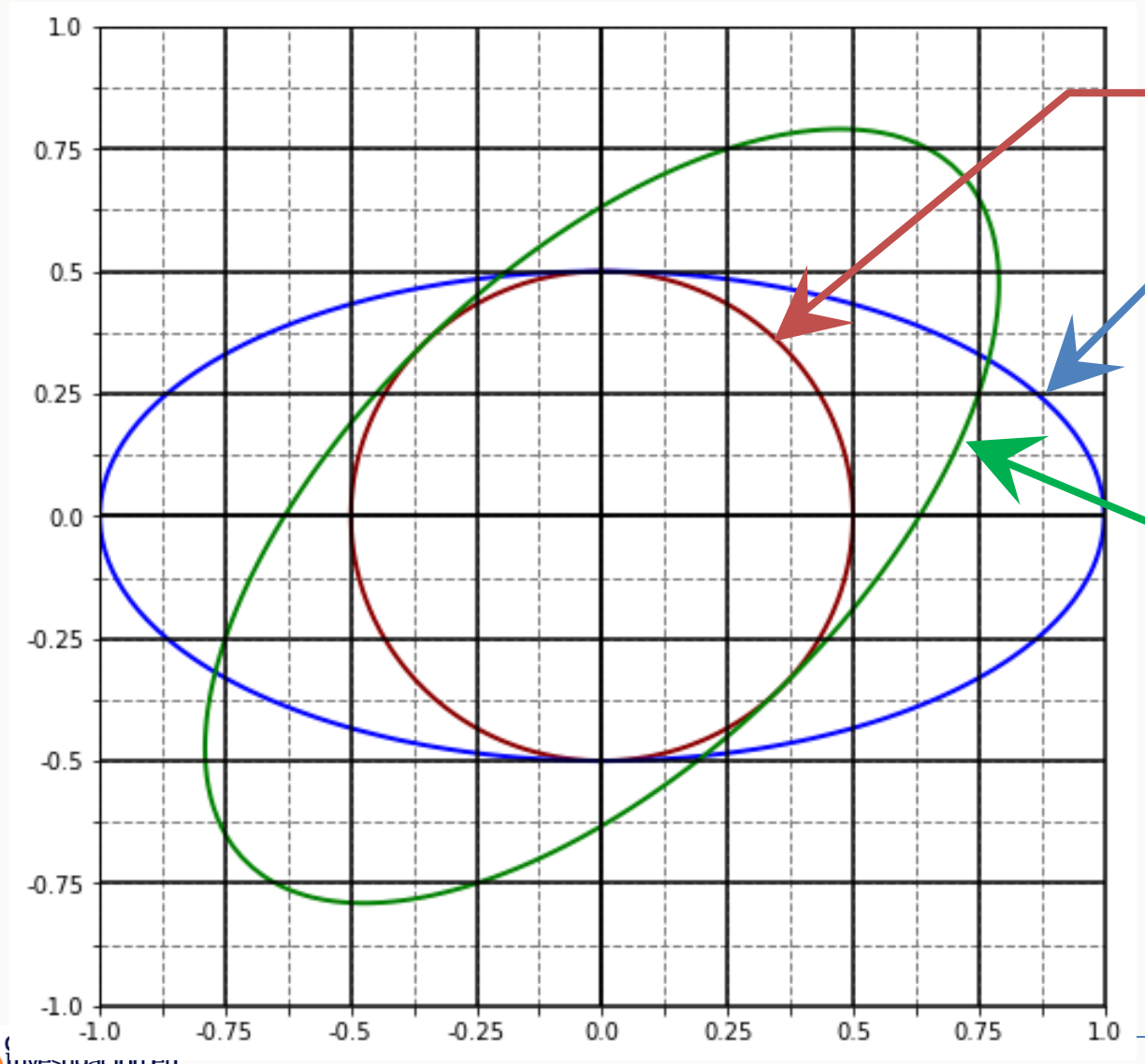


$$p = 0.8$$
$$\overline{AB} = \overline{AC}$$
$$\overline{OB} < \overline{OC}$$
$$\overline{OA} + \overline{AC} < \overline{OC}$$

X

DISTANCIA DE MAHALANOBIS

EJES COORDENADOS A DISTINTA ESCALA



$$\mathbf{x}^{\mathrm{T}}\mathbf{x} = 1/4$$

$$\mathbf{x}^{\mathrm{T}}\mathbf{S}^2\mathbf{x} = 1/4 \qquad \mathbf{S} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\text{Si } \mathbf{x} = (x, y),\ \mathbf{x}^{\mathrm{T}}\mathbf{S}^2\mathbf{x} = \frac{x^2}{(2)^2} + \frac{y^2}{(1)^2}$$

$$\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x} = 1/4$$

$$\mathbf{M} = (\mathbf{SR})^{\mathrm{T}}(\mathbf{SR}) = \mathbf{R}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{SR} = \mathbf{R}^{\mathrm{T}}\mathbf{S}^2\mathbf{R}$$

$$\mathbf{R} = \begin{bmatrix} \cos(\pi/4) & \sin(\pi/4) \\ -\sin(\pi/4) & \cos(\pi/4) \end{bmatrix}$$

$$\mathbf{M} = \begin{bmatrix} 5/8 & -3/8 \\ -3/8 & 5/8 \end{bmatrix}$$

MATRIZ DE COVARIANZA

- Típicamente la forma de la elipse se obtiene de los datos y corresponde a la matriz de covarianza $\boldsymbol{\Sigma}$:

$$\mathbf{M} = \boldsymbol{\Sigma}^{-1}$$

- Para el caso bidimensional la matriz de covarianza es:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\det(\boldsymbol{\Sigma})} \begin{bmatrix} \sigma_2^2 & -\rho_{12}\sigma_1\sigma_2 \\ -\rho_{12}\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

DISTANCIA DE MAHALANOBIS

- Nota que para la Distribución Normal Univariada podemos escribir el exponente como:

$$-\frac{(x-\mu)^2}{2\sigma^2} = -\frac{1}{2}(x-\mu)(\sigma^{-1})^2(x-\mu)$$

- Distribución Normal multivariada sobre vectores de dimensión d: $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$

$$\mathrm{P}(\mathbf{x}|\boldsymbol{\mu}\,\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\underbrace{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}_{\text{distancia de Mahalanobis}}\right)$$

- Los parámetros son el vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}}$ y la matriz de covarianza $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}^{-1}$ es la inversa de la matriz de covarianza y $|\boldsymbol{\Sigma}|$ es el determinante de la matriz.

MATRIZ DE COVARIANZA

- En nuestro contexto la matriz $\mathbf{X}$ representa un conjunto de datos con N ejemplos, cada uno con d atributos ($features$).

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} \text{ donde } \mathbf{x}_i = [x_{i1}, \dots, x_{id}] \text{ para } i = 1,2, \dots, N$$

- El promedio por **columna** es $\mu_j = \sum_{i=1}^{N} x_{ij}$; $\boldsymbol{\mu}^{\mathrm{T}}$ es un vector fila que contiene todos los promedios por columna (i.e. por atributo).

- Si **1** es un vector de dimensión N con puros unos:

$$\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \text{ entonces } \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} = \begin{bmatrix} \mu_1 & \cdots & \mu_d \\ \vdots & \ddots & \vdots \\ \mu_1 & \cdots & \mu_d \end{bmatrix}$$

MATRIZ DE COVARIANZA

- La matriz

$$\mathbf{X_c} = \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}}$$

es la matriz de datos *centrada en cero*.

- La *matriz de dispersión* (*scatter matrix*) es la matriz dxd:

$$\mathbf{S} = \mathbf{X_c}^{\mathrm{T}}\mathbf{X_c} = \left(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}}\right)$$

MATRIZ DE COVARIANZA

- La *matriz de covarianza* de $\mathbf{X}$ es:

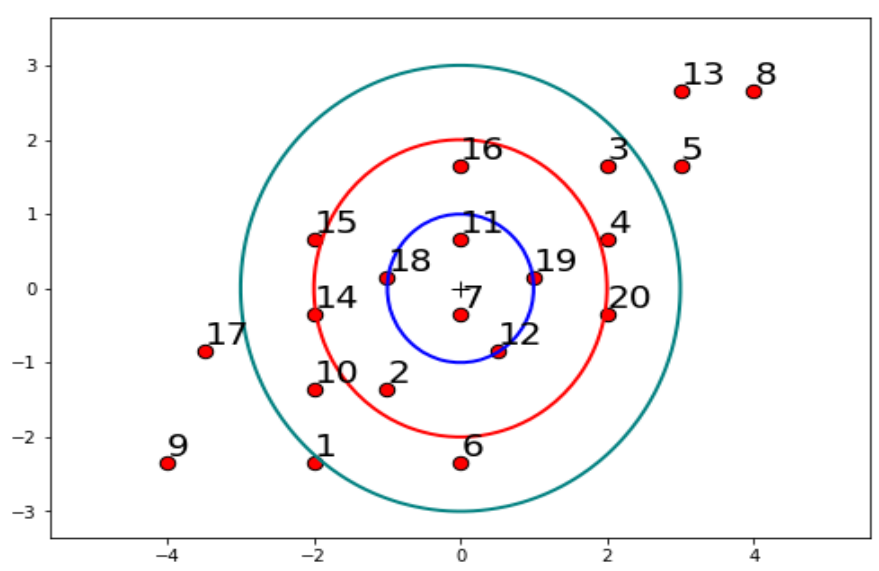
$$\boldsymbol{\Sigma} = \frac{1}{N-1}\mathbf{S}$$

- En la diagonal de $\boldsymbol{\Sigma}$ tenemos las *varianzas* σ_{jj} o σ_j^2:

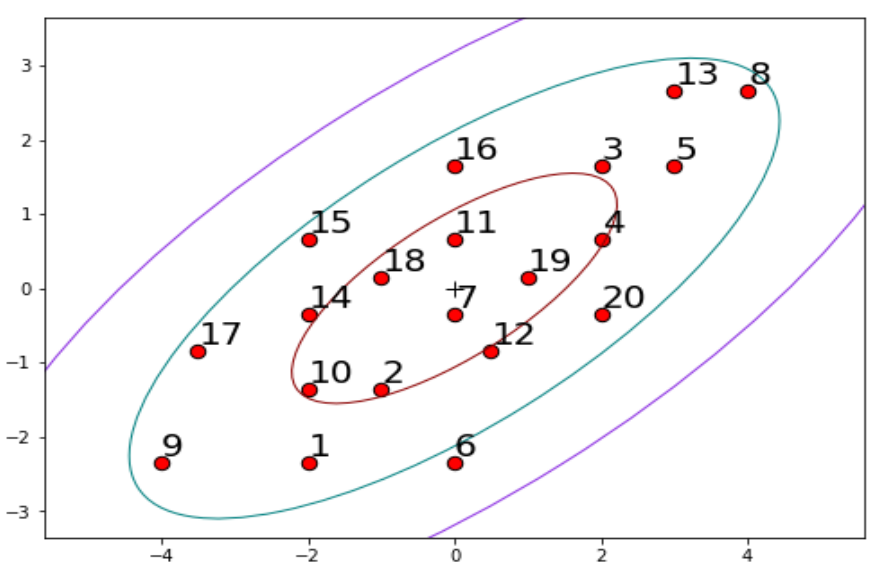
$$\sigma_j^2 = \sigma_{jj} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij} - \mu_j\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_{ij}^2 - \mu_j^2$$

DISTANCIA DE MAHALANOBIS

- Tutorial disponible en:
- R. De Maesschalck, D. Jouan-Rimbaud and D. L. Massart. *Tutorial The Mahalanobis distance*. Chemometrics and Intelligent Laboratory Systems. 50 (1) pp 1 – 18. Elsevier (2000).



Los círculos representan puntos equidistantes del centro bajo una distancia Euclidiana.



Las elipses representan puntos equidistantes del centro bajo una distancia de Mahalanobis.

DISTANCIA DE MAHALANOBIS

$$DE(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})\mathbf{I}(\mathbf{x}-\mathbf{y})^T}$$

$$DM(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})^T}$$

- Para el caso bidimensional:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\det(\boldsymbol{\Sigma})}\begin{bmatrix} \sigma_1^2 & -\rho_{12}\sigma_1\sigma_2 \\ -\rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

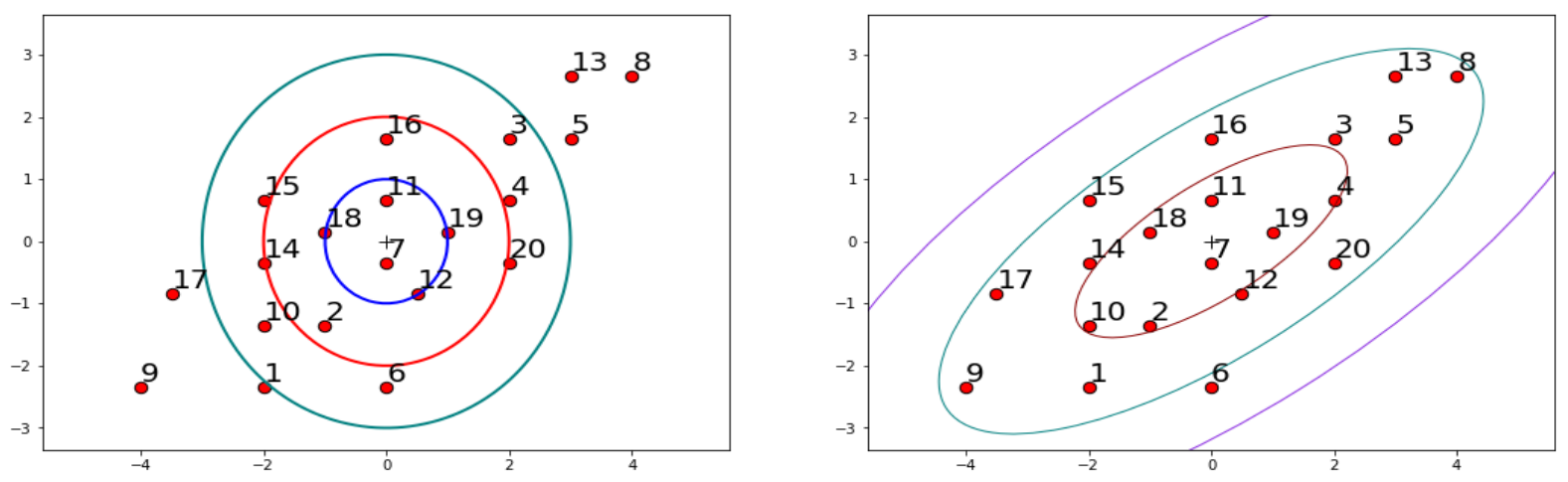
objeto	ED	MD
1	3.0859	1.5464
2	1.6800	0.9122
3	2.5928	1.0814
4	2.1030	0.9652
5	3.4238	1.3576
6	2.3500	2.2133
7	0.3500	0.3296
8	4.7982	1.8947
9	4.6392	1.8278
10	2.4130	0.9549
11	0.6500	0.6122
12	0.9862	1.0640
13	4.0028	1.7186
14	2.0304	1.0983
15	2.1030	1.8097
16	1.6500	1.5540
17	3.6017	1.8037
18	1.0112	0.7664
19	1.0112	0.5629
20	2.0304	1.5710

DISTANCIA DE MAHALANOBIS

- Para el ejemplo del artículo:
- $\mathrm{DM_i} = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{C}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})}$ con $\mathbf{C} = \begin{bmatrix} 4.921 & 2.500 \\ 2.500 & 2.397 \end{bmatrix}$
- Esta expresión se puede re-escribir para $\mathbf{x} = (x_1, x_2)$:
- $$\mathrm{DM_i} = \sqrt{\left(\frac{x_{i1}-\bar{x}_1}{\sigma_1}\right)^2 + \left[\left\{\left(\frac{x_{i2}-\bar{x}_2}{\sigma_2}\right) - \rho_{12}\left(\frac{x_{i1}-\bar{x}_1}{\sigma_1}\right)\right\}\frac{1}{\sqrt{1-\rho_{12}^2}}\right]^2}$$
- Esta ecuación muestra:
  - *La parte de la segunda variable que ya fue explicada por la primera se substrae. En otras palabras la DM corrige la correlación entre los datos (de-correlaciona los datos).*
  - *Cuando los datos no están correlacionados* $(\rho_{12} = 0)$ *la ecuación se reduce a la fórmula de la DE.*

CONSECUENCIAS DE LA DM

- Los puntos 6 y 10 están respectivamente a 2.21 y menos de 1 DM del centro, mientras que sus DE son casi iguales.
- Este ejemplo ilustra el efecto de tomar en cuenta la matriz de varianza-covarianza de los puntos. Puesto que el punto 6 se encuentra en una dirección donde hay menos correlación, la probabilidad de que una nueva medición esté en ese lado de la nube es menor que en la posición del punto 10.
- La DM toma en cuenta esta probabilidad debido a la correlación entre las variables y le atribuye al punto 6 una mayor distancia que al punto 10, mientras que la DE no hace esto.



ALGORITMO K-MEANS

AGRUPAMIENTO BASADO EN DISTANCIA

Recordemos que la matriz de dispersión es:

$$\mathbf{S} = \mathbf{X_c^T X_c} = \left(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}}\right) = \sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}(\mathbf{x}_i - \boldsymbol{\mu})$$

donde **$\boldsymbol{\mu}$ es un vector**

La dispersión de $\mathbf{X}$ es $\mathrm{Disp}(\mathbf{X}) = \sum_{i=1}^{N}\|\mathbf{x}_i - \boldsymbol{\mu}\|^2$, que es igual a la Traza de la matriz de dispersión (es, decir, la suma de los elementos de la diagonal)

Dispersión intra-cluster e inter-cluster

Imaginemos que hacemos una partición de los datos $\mathbf{D}$ en K grupos (clusters)

$$\mathbf{D} = \mathbf{D}_1 \uplus \mathbf{D}_2 \cdots \uplus \mathbf{D}_K$$ y sea $\boldsymbol{\mu}_j$ el promedio de $\mathbf{D}_j$, y $\mathbf{S}_j$ la matriz de dispersión de $\mathbf{D}_j$.

- Las matrices de dispersión tienen entonces esta relación: $\mathbf{S} = \sum_{j=1}^{K} \mathbf{S}_j + \mathbf{B}$
- $\mathbf{B}$ es la matriz de dispersión inter-cluster que resulta de remplazar cada punto de $\mathbf{D}$ con su centroide correspondiente $\boldsymbol{\mu}_j$: describe la dispersión (varianza) de los centroides.
- A cada $\mathbf{S}_j$ se le llama una matriz de dispersión intra-cluster y describe cuan compacto es un grupo j.
- Las trazas de estas matrices se descomponen de manera similar:

$$\mathrm{Disp}(\mathbf{D}) = \sum_{j=1}^{K} \mathrm{Disp}(\mathbf{D}_j) + \sum_{j=1}^{K} |\mathbf{D}_j| \|\boldsymbol{\mu}_j - \boldsymbol{\mu}\|^2$$

- El **problema de los K-promedios (K-means)** es encontrar una partición que **minimice la dispersión intra-cluster.**

EJEMPLO

Consider the following five points centred around $(0,0)$: $(0,3)$, $(3,3)$, $(3,0)$, $(-2,-4)$ and $(-4,-2)$. The scatter matrix is

$$\mathbf{S} = \begin{pmatrix} 0 & 3 & 3 & -2 & -4 \\ 3 & 3 & 0 & -4 & -2 \end{pmatrix} \begin{pmatrix} 0 & 3 \\ 3 & 3 \\ 3 & 0 \\ -2 & -4 \\ -4 & -2 \end{pmatrix} = \begin{pmatrix} 38 & 25 \\ 25 & 38 \end{pmatrix}$$

with trace $\text{Scat}(D) = 76$. If we cluster the first two points together in one cluster and the remaining three in another, then we obtain cluster means $\mu_1 = (1.5,3)$ and $\mu_2 = (-1,-2)$ and within-cluster scatter matrices

$$\mathbf{S}_1 = \begin{pmatrix} 0-1.5 & 3-1.5 \\ 3-3 & 3-3 \end{pmatrix} \begin{pmatrix} 0-1.5 & 3-3 \\ 3-1.5 & 3-3 \end{pmatrix} = \begin{pmatrix} 4.5 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\mathbf{S}_2 = \begin{pmatrix} 3-(-1) & -2-(-1) & -4-(-1) \\ 0-(-2) & -4-(-2) & -2-(-2) \end{pmatrix} \begin{pmatrix} 3-(-1) & 0-(-2) \\ -2-(-1) & -4-(-2) \\ -4-(-1) & -2-(-2) \end{pmatrix} = \begin{pmatrix} 26 & 10 \\ 10 & 8 \end{pmatrix}$$

with traces $\text{Scat}(D_1) = 4.5$ and $\text{Scat}(D_2) = 34$.

EJEMPLO

Two copies of μ_1 and three copies of μ_2 have, by definition, the same centre of gravity as the complete data set: $(0,0)$ in this case. We thus calculate the between-cluster scatter matrix as

$$\mathbf{B} = \begin{pmatrix} 1.5 & 1.5 & -1 & -1 & -1 \\ 3 & 3 & -2 & -2 & -2 \end{pmatrix} \begin{pmatrix} 1.5 & 3 \\ 1.5 & 3 \\ -1 & -2 \\ -1 & -2 \\ -1 & -2 \end{pmatrix} = \begin{pmatrix} 7.5 & 15 \\ 15 & 30 \end{pmatrix}$$

with trace 37.5. Alternatively, if we treat the first three points as a cluster and put the other two in a second cluster, then we obtain cluster means $\mu_1' = (2,2)$ and $\mu_2' = (-3,-3)$, and within-cluster scatter matrices

$$\mathbf{S}_1' = \begin{pmatrix} 0-2 & 3-2 & 3-2 \\ 3-2 & 3-2 & 0-2 \end{pmatrix} \begin{pmatrix} 0-2 & 3-2 \\ 3-2 & 3-2 \\ 3-2 & 0-2 \end{pmatrix} = \begin{pmatrix} 6 & -3 \\ -3 & 6 \end{pmatrix}$$

$$\mathbf{S}_2' = \begin{pmatrix} -2-(-3) & -4-(-3) \\ -4-(-3) & -2-(-3) \end{pmatrix} \begin{pmatrix} -2-(-3) & -4-(-3) \\ -4-(-3) & -2-(-3) \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$$

EJEMPLO

with traces $\text{Scat}(D'_1) = 12$ and $\text{Scat}(D'_2) = 4$. The between-cluster scatter matrix is

$$\mathbf{B}' = \begin{pmatrix} 2 & 2 & 2 & -3 & -3 \\ 2 & 2 & 2 & -3 & -3 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 2 & 2 \\ 2 & 2 \\ -3 & -3 \\ -3 & -3 \end{pmatrix} = \begin{pmatrix} 30 & 30 \\ 30 & 30 \end{pmatrix}$$

with trace 60. Clearly, the second clustering produces tighter clusters whose centroids are further apart.

ALGORITMO K-MEANS

- K-means es un problema NP-completo
  - *No hay una solución eficiente para encontrar el mínimo global.*
  - *Hay que recurrir a un algoritmo heurístico.*
- El mejor algoritmo conocido se llama también K-means.
  - *El algoritmo itera entre la partición de los datos mediante la regla de decisión del centroide más cercano, y el recálculo de los centroides de la partición.*
  - *Se puede demostrar que una iteración de K-means no puede jamás incrementar la dispersión intra-cluster, de lo cuál se infiere que el algoritmo llegará a un punto estacionario: un punto en el cuál ya no puede mejorar.*

ALGORITMO K-MEANS

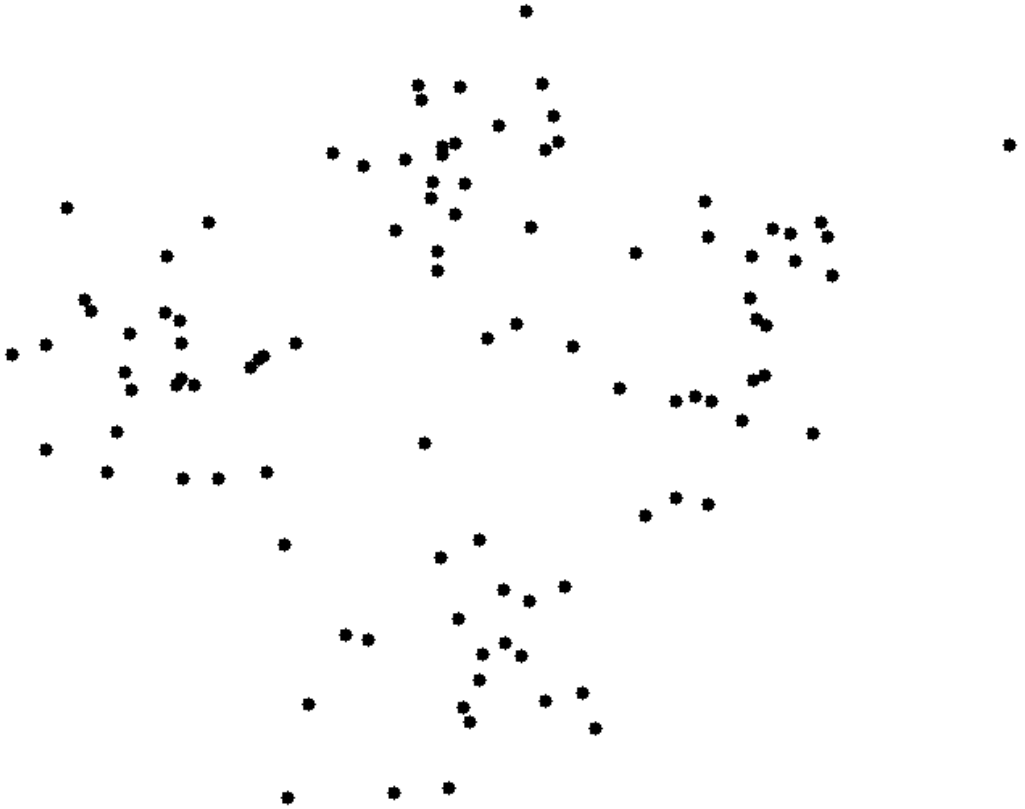
Algorithm KMeans(D, K)

Input : data $D \subseteq \mathbb{R}^d$; number of clusters $K \in \mathbb{N}$.
Output : K cluster means $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$.

randomly initialise K vectors $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$;
repeat
  assign each $\mathbf{x} \in D$ to $\arg\min_j \mathrm{Dis}_2(\mathbf{x}, \mu_j)$;
  for $j = 1$ to K **do**
    $D_j \leftarrow \{\mathbf{x} \in D | \mathbf{x}$ assigned to cluster $j\}$;
    $\mu_j = \frac{1}{|D_j|} \sum_{\mathbf{x} \in D_j} \mathbf{x}$;
  end
until no change in $\mu_1, \ldots, \mu_K$;
return $\mu_1, \ldots, \mu_K$;

SIMULACIÓN K-MEANS

Créditos: Andrey A. Shabalin, Ph.D. (http://shabal.in/visuals.html)
K-means clustering. K-means++ (http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf)



MÉTODOS DE SELECCIÓN DE K

MÉTODO DEL CODO (ELBOW METHOD)

- Usualmente se utiliza el criterio de minimización de la inercia o de la suma-de-cuadrados intra-cluster (within-cluster sum-of-squares – WSS):

$$\mathrm{WSS} = \sum_{i=0}^{n} \min_{\mu_j \in \mathcal{C}} \left(\left\| x_i - \mu_j \right\|^2 \right)$$

1. Para cada K, calcula WSS.
2. Traza WSS en función de K.
3. La ubicación del la rodilla (codo) es considerado un indicador del número apropiado de cúmulos.

SILUETAS

- Para cualquier punto $\mathbf{x}_i$, sea $\mathrm{d}(\mathbf{x}_i, \mathbf{D}_j)$ la distancia promedio de $\mathbf{x}_i$ a los puntos del cluster $\mathbf{D}_j$, y sea $j(i)$ el índice al cual $\mathbf{x}_i$ pertenece.

- Sea además, $a(\mathbf{x}_i) = \mathrm{d}(\mathbf{x}_i, \mathbf{D}_{j(i)})$ la distancia promedio de $\mathbf{x}_i$ a los puntos de su propio cluster $\mathbf{D}_{j(i)}$, y sea $b(\mathbf{x}_i) = \min_{\mathrm{k} \neq j(i)} \mathrm{d}(\mathbf{x}_i, \mathbf{D}_k)$ la distancia promedio a los puntos de su cluster vecino.

- Esperaríamos que $a(\mathbf{x}_i)$ sea considerablemente más pequeño que $b(\mathbf{x}_i)$, pero esto no se puede garantizar.

- Así que podemos tomar la diferencia $b(\mathbf{x}_i) - a(\mathbf{x}_i)$ como un indicador de "qué tan bien" están agrupados los $\mathbf{x}_i$, y dividir esto por $b(\mathbf{x}_i)$ para normalizar entre 0 y 1.

SILUETAS

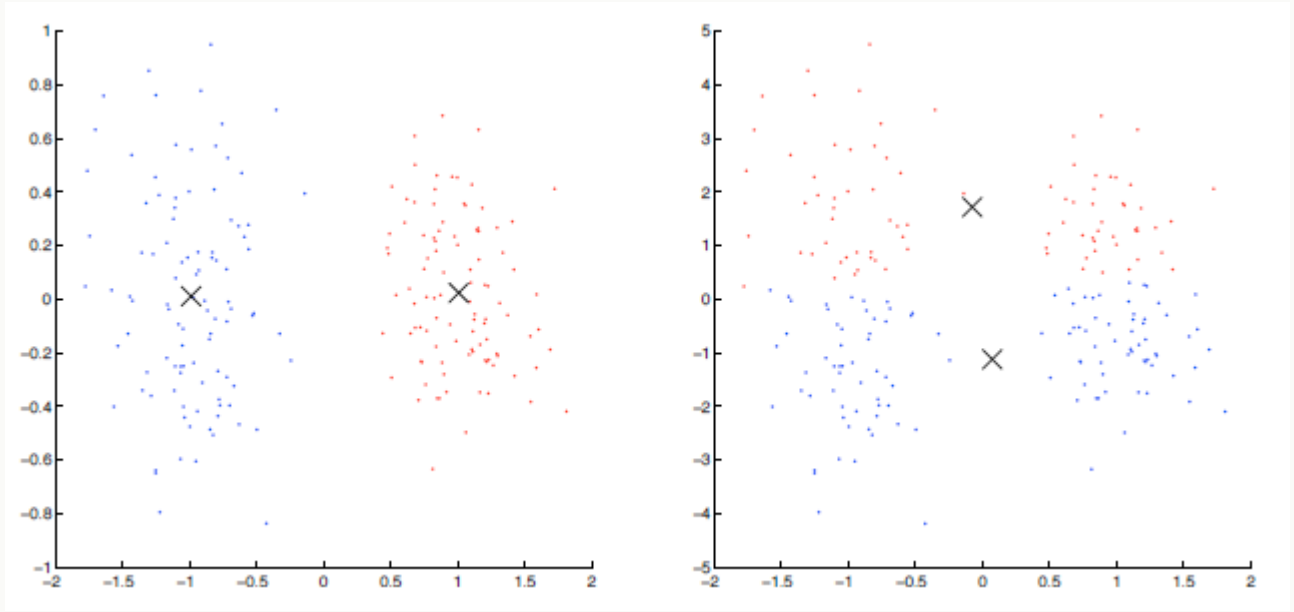
- Es concebible sin embargo que $a(\mathbf{x}_i) > b(\mathbf{x}_i)$, en cuyo caso la diferencia sería negativa.

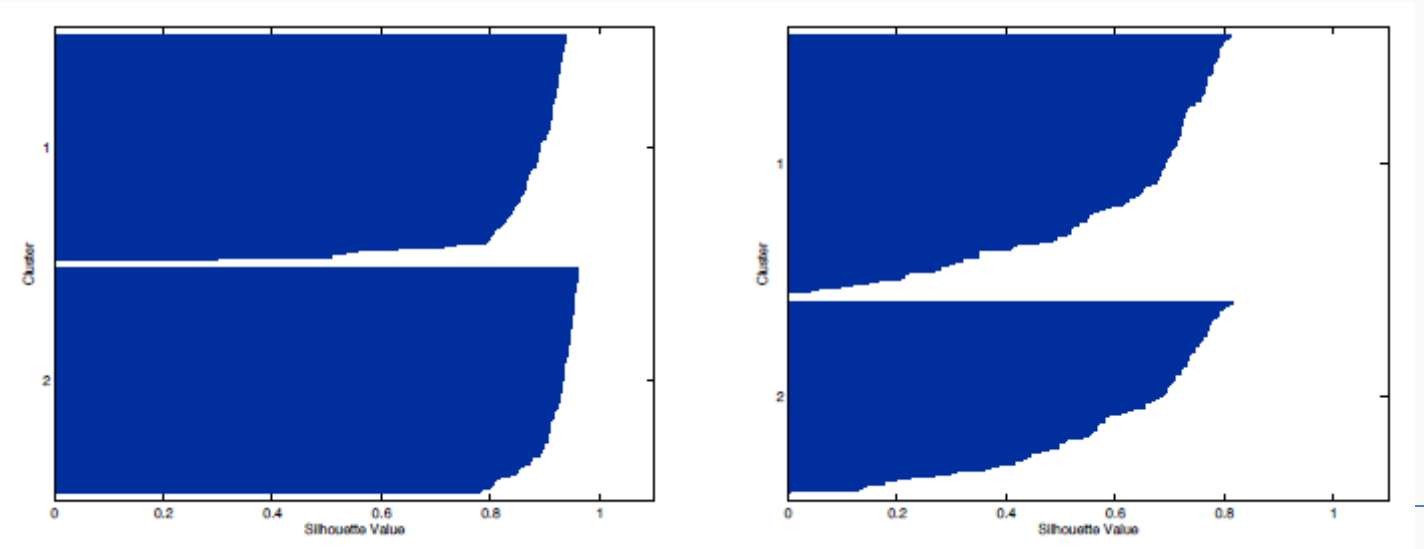
- En este caso tendríamos que dividir entre $a(\mathbf{x}_i)$. Esto nos lleva a la siguiente definición:

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max(a(\mathbf{x}_i), b(\mathbf{x}_i))}$$

- Una silueta ordena y grafica $s(\mathbf{x})$ para cada instancia, agrupada por cluster.

SILUETAS





- Puntos con S_i grandes (casi 1) están bien agrupadas
- Una S_i pequeña (circa 0) significa que el punto se encuentra entre dos clusters.
- Puntos con una S_i negativa están probablemente ubicados en el cluster equivocado.

MÉTODO DEL PROMEDIO DE SILUETAS

1. Corre el algoritmo de clustering para distintos valores de K (e.g. K=1 a 10).
2. Para cada K, calcula el valor promedio de S_i ($\bar{S}$) de todos los puntos.
3. Traza $\bar{S}$ en función de K
4. La ubicación del máximo es considerado un indicador del número apropiado de cúmulos.

MÉTODO DE LA ESTADÍSTICA DE BRECHA (GAP STATISTIC METHOD)

- Propuesto por Robert Tibshirani, Guenther Walther y Trevor Hastie (2002). *Estimating the number of clusters in a data set via the gap statistic*. Journal of the Royal Statistical Society Series B (Statistical Methodology) 63(2):411-423. DOI: 10.1111/1467-9868.00293

- El principio es el siguiente:

  - *Los datos* $\{x_{ij}\}$, $i = 1,2,\dots,n$, $j = 1,2,\dots,p$ *consisten en* p *features medidas en* n *observaciones.*

  - *Sea* $d_{ii'} = \sum_j \left\| x_{ij} - x_{i'j} \right\|^2$ *la distancia Euclideana al cuadrado entre dos observaciones i e i'.*

MÉTODO DE LA ESTADÍSTICA DE BRECHA (GAP STATISTIC METHOD)

- El principio es el siguiente (cont.):

  - *Supongamos que tenemos los datos agrupados en k cúmulos* $C_1, C_2, \dots, C_k$, *con* C_r *denotando los índices de observaciones en el cluster* r *y* $n_r = |C_r|$.

  - *Sea* $D_r = \sum_{i,i' \in C_r} d_{ii'}$ *la suma de las distancias entre pares de puntos en el cluster* r.

  - *Sea* $W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$ *la suma de cuadrados intra-cluster alrededor de las medias de los clusters agrupada.*

MÉTODO DE LA ESTADÍSTICA DE BRECHA (GAP STATISTIC METHOD)

- El principio es el siguiente (cont.):

  - *La idea es estandarizar el grafo de* $\log(W_k)$ *comparándolo contra su valor esperado bajo una apropiada distribución de datos nula.*

  - *El estimado del número óptimo de clusters es el valor de* k *para el cual* $\log(W_k)$ *cae lo más lejos por debajo de la curva de referencia.*

  - *Se define entonces:*

$$\mathrm{Gap}_n(k) = E_n^*\{\log(W_k)\} - \log(W_k)$$

MÉTODO DE LA ESTADÍSTICA DE BRECHA (GAP STATISTIC METHOD)

1. Corre el algoritmo de clustering para distintos valores de K (e.g. K=1 a 10), y calcula el valor correspondiente W_k.

2. Genera **B** conjuntos de datos de referencia mediante una distribución uniforme. Agrupa cada uno de estos conjuntos de datos, variando el número de clusters $k = 1, \dots, k_{max}$ y calcula el valor correspondiente W_{kb}.

3. Calcula la estimación de la estadística de brecha como la desviación de los valores de W_k con respecto a los valores esperados W_{kb} bajo la hipótesis nula:

$$\mathrm{Gap}(k) = \frac{1}{\mathrm{B}} \sum_{b=1}^{B} \log(W_{k\mathrm{b}}) - \log(W_k)$$

MÉTODO DE LA ESTADÍSTICA DE BRECHA (GAP STATISTIC METHOD)

4. Calcula las desviaciones estándar de la estadística de brecha.

5. Elige el número de clusters como el valor más pequeño de k, tal que la estadística de gap se encuentra dentro de una desviación estándar de la brecha en k+1:

$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - \sigma_{k+1}$$

AGRUPAMIENTO JERÁRQUICO

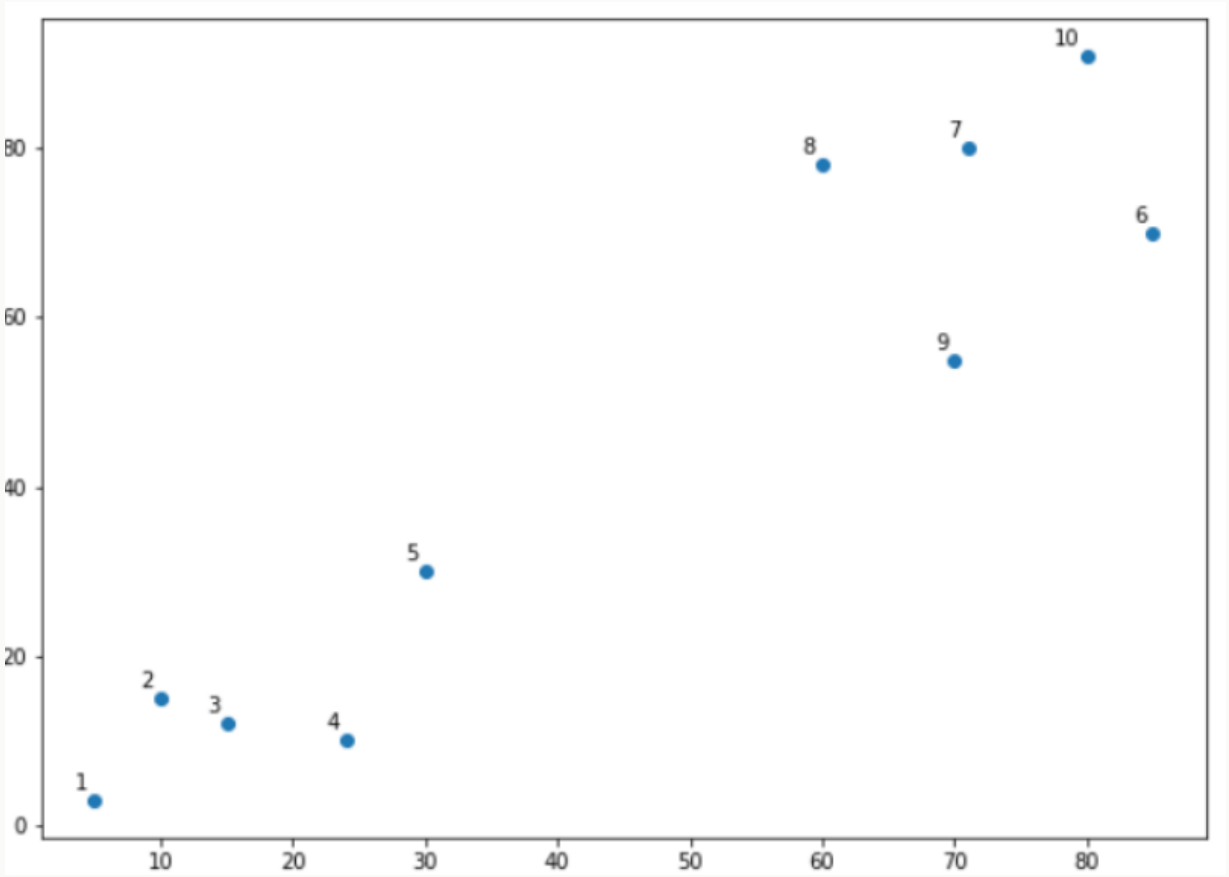
¿QUÉ ES EL AGRUPAMIENTO JERÁRQUICO?

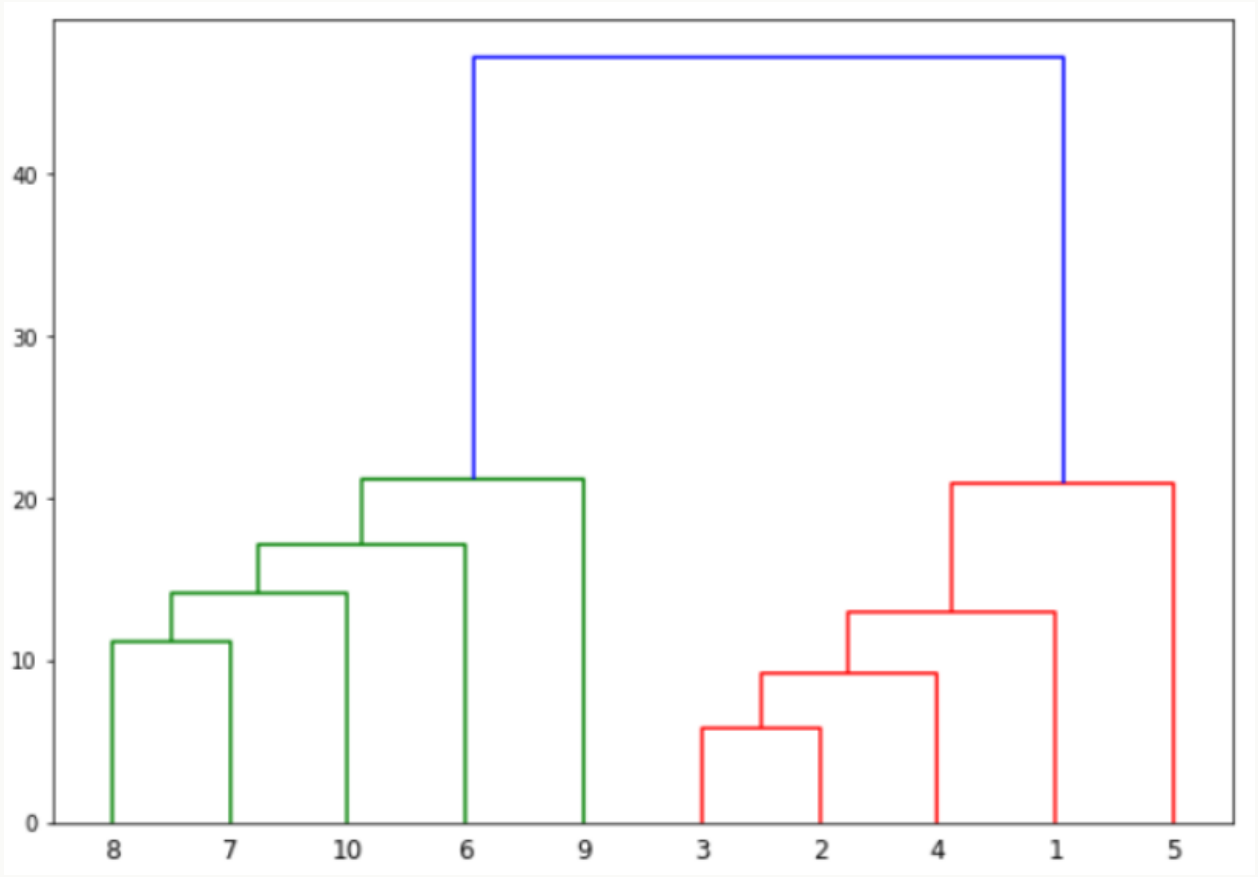
- Es un método de análisis de agrupamiento exploratorio.
- No requiere un número k a priori de clusters.
- El análisis se hace mediante particiones secuenciales de los datos.
- Construye particiones capa por capa a través del agrupamiento de objetos en árboles de clusters.

AGRUPAMIENTO JERÁRQUICO AGLOMERATIVO

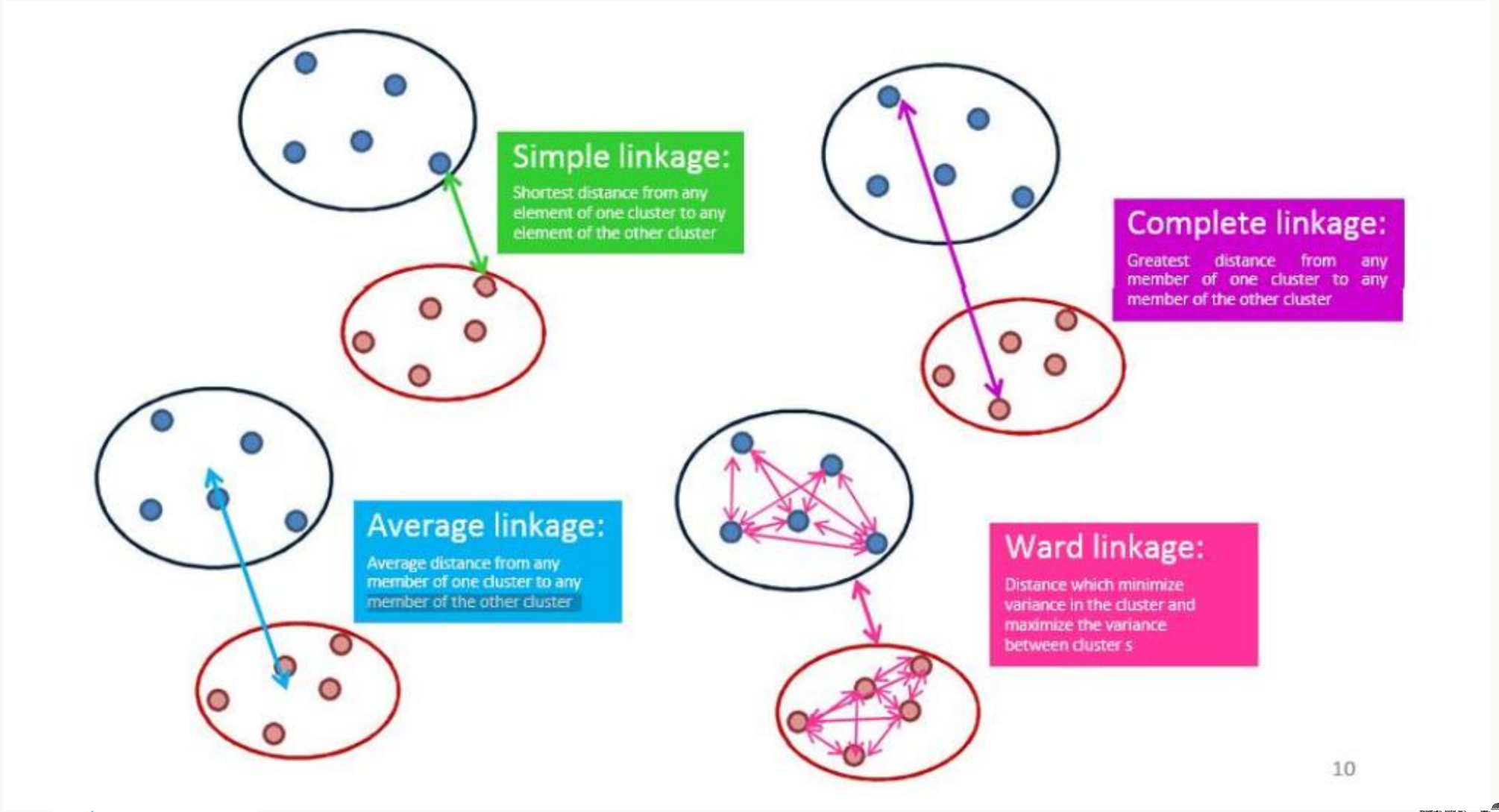
1. Comienza tartando cada punto como un cluster. K clusters si hay K puntos.
2. Forma un cluster uniendo los dos puntos más cercanos; esto resulta en K-1 clusters.
3. Forma más clusters uniendo los dos clusters más cercanos -> K-2 clusters.
4. Repite los tres pasos anteriores hasta que sólo haya un solo cluster.
5. Una vez que queda un solo cluster, se utilizan dendogramas para dividir en múltiples clusters dependiendo del problema.

EJEMPLO


10
8
7
6
9
5
2
3
4
1
0
20
40
60
80
10
20
30
40
50
60
70
80


0
10
20
30
40
8
7
10
6
9
3
2
4
1
5

MÉTODOS DE AGLOMERACIÓN



AGRUPAMIENTO JERÁRQUICO DIVISIVO

- Trabaja en sentido opuesto al aglomerativo.
- Comienza con un solo cluster que abarque todos los puntos.
- Ahora, en cada iteración separa el punto más lejano y repite el proceso hasta llegar a que cada punto es un cluster.

OTROS MÉTODOS DE AGRUPAMIENTO

MÉTODOS EN SCIKIT-LEARN

- https://scikit-learn.org/stable/modules/clustering.html