UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

OVA 1: Embedding Methods

CS3102 EDA

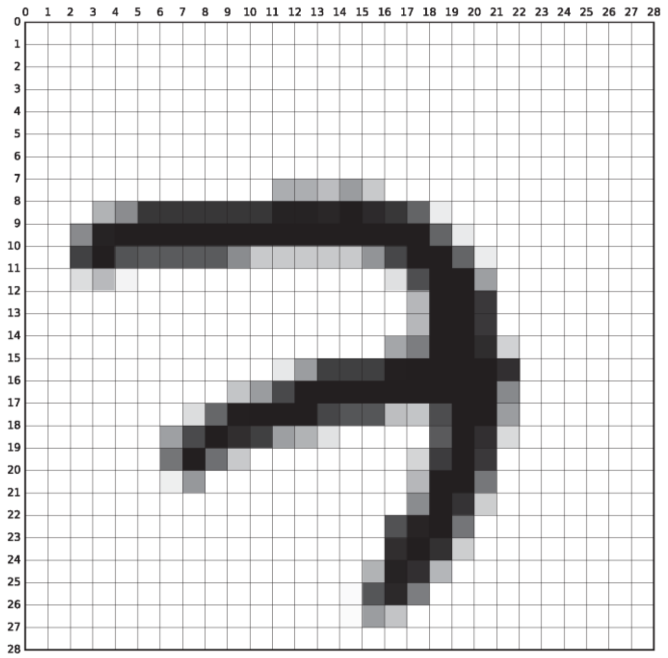
Índice

1. Embedding Methods

$\begin{bmatrix}1&2\\3&4\end{bmatrix} \to [1\ 2\ 3\ 4]$

Embedding *Methods*

6

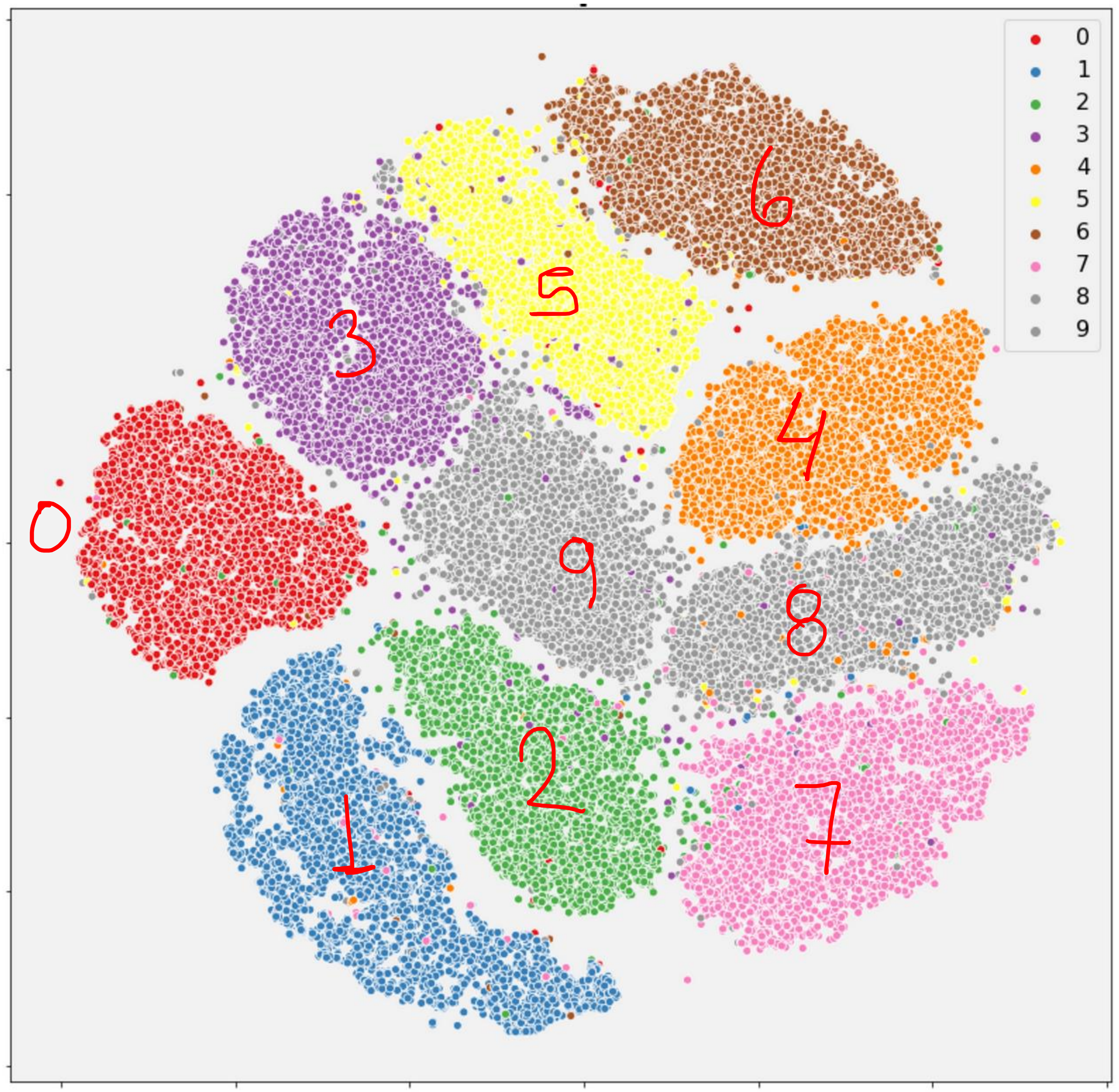


$\underbrace{28 \times 28}$

784

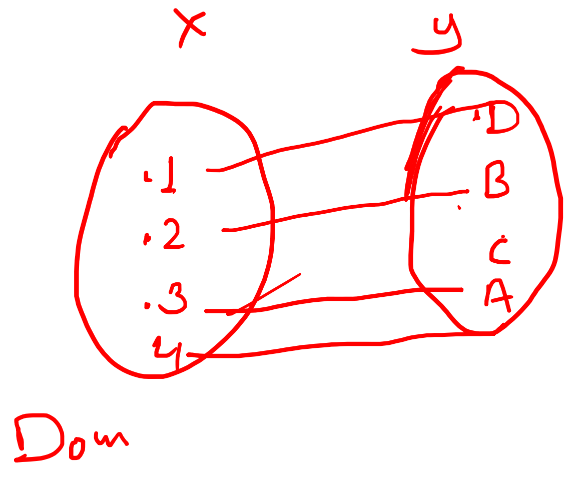
t-SNE

MNIST

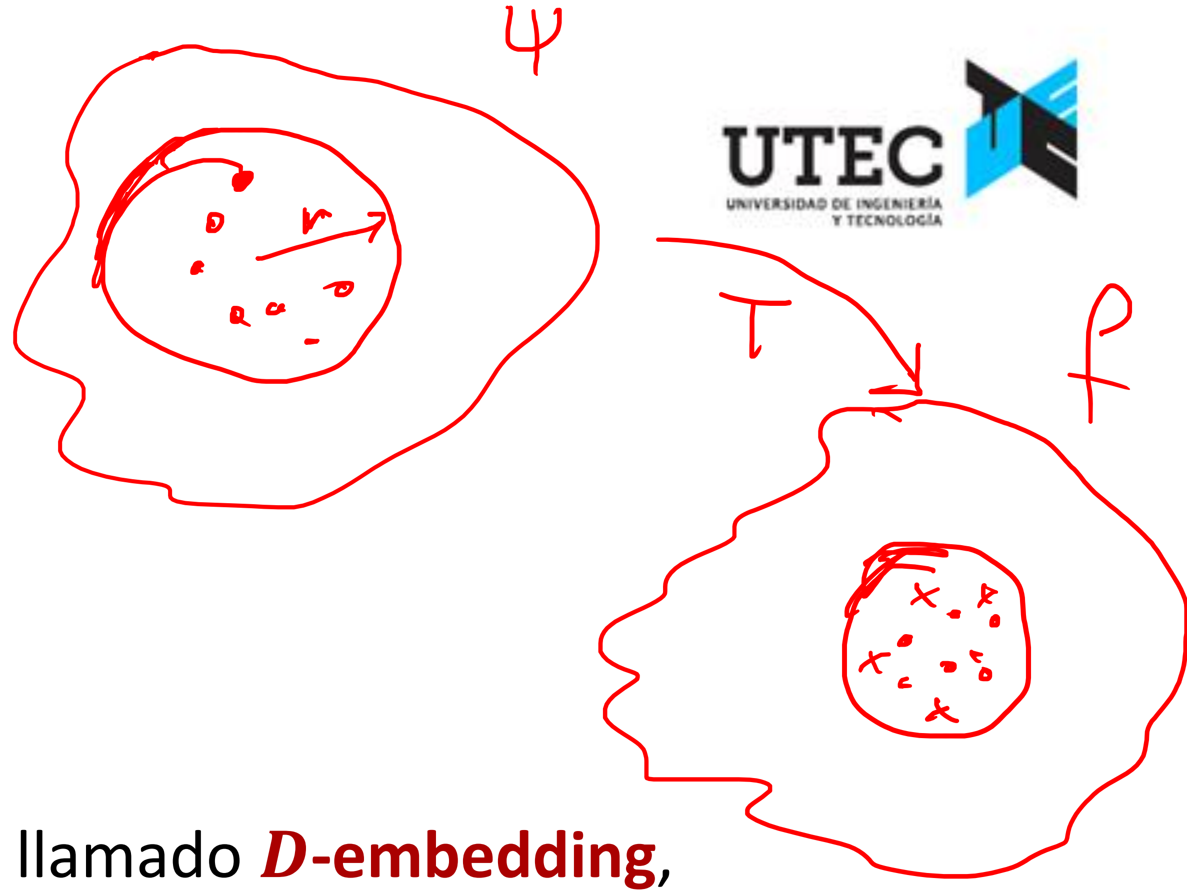


$\boxed{d(a,b)} \approx \boxed{d'(F(a),F(b))}$

isométrico



Embedding *Methods*



Textos Palabras

Euclidian Space

Sea $(\mathcal{O}, d)$ y $(\mathcal{F}, d')$ son *metric space*. Un mapeo inyectivo $f: (\mathcal{O}, d) \to (\mathcal{F}, d')$ es llamado ***D*-embedding**, donde $D \geq 1$ es un número real, si un número $r > 0$ de modo que para todo $o_1, o_2 \in \mathcal{O}$

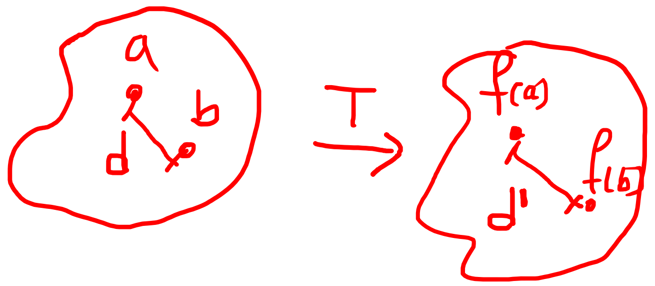
$0 < \gamma < 1$

$$\underbrace{\overset{\to 0}{r} \cdot d(o_1, o_2)}_{d_1} \leq \underbrace{d'\big(f(o_1), f(o_2)\big) \leq \boxed{\overset{\gamma}{D \cdot r}} \cdot d(o_1, o_2)}_{d_2}$$

El menor valor de D que cumpla la condición anterior es llamada **distorsión** de f.

$$\boxed{D = \frac{d_2}{d_1}}$$

Embedding *Methods*



$\left[\frac{d'}{d}\right]$

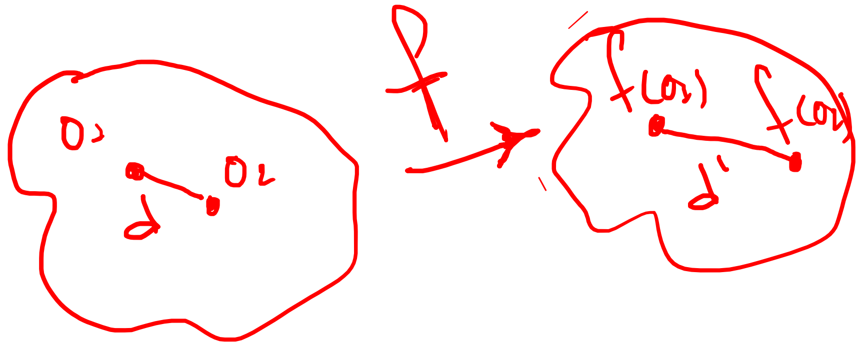
$C \leq d$

El mapeo $f: (\mathcal{O}, d) \to (\mathcal{F}, d')$ es llamado ***C*-Lipschitz** si $\boxed{d'\big(f(o_1), f(o_2)\big) \leq C \cdot d(o_1, o_2)}$ para todo $o_1, o_2 \in \mathcal{O}$.
Sea:

$$\|f\|_{\mathrm{Lip}} = \sup\left\{\frac{d'\big(f(o_1), f(o_2)\big)}{d(o_1, o_2)} : o_1, o_2 \in \mathcal{O}, o_1 \neq o_2\right\}$$

la **Lipschitz-norm** de f, la más pequeña posible tal que sea ***C*-Lipschitz**. Ahora bien, si f es un **mapa biyectivo**, no es difícil comprobar que su distorsión es igual a $\|f\|_{\mathrm{Lip}}\|f^{-1}\|_{\mathrm{Lip}}$. Por esta razón, los mapas con una distorsión finita se denominan a veces **bi-Lipschitz**.

Embedding *Methods*

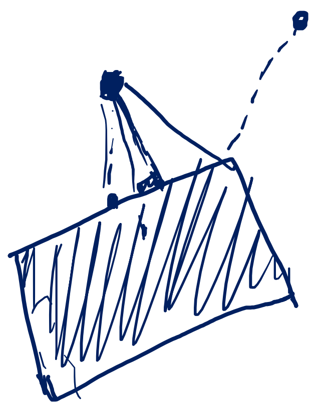


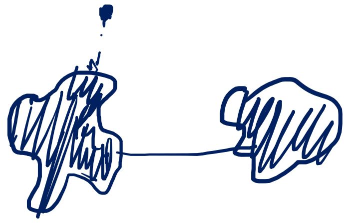
Stress

$$\frac{\sum_{o_1,o_2}\left[d'\big(f(o_1),f(o_2)\big)-d(o_1,o_2)\right]^2}{\sum_{o_1,o_2} d(o_1,o_2)^2}$$

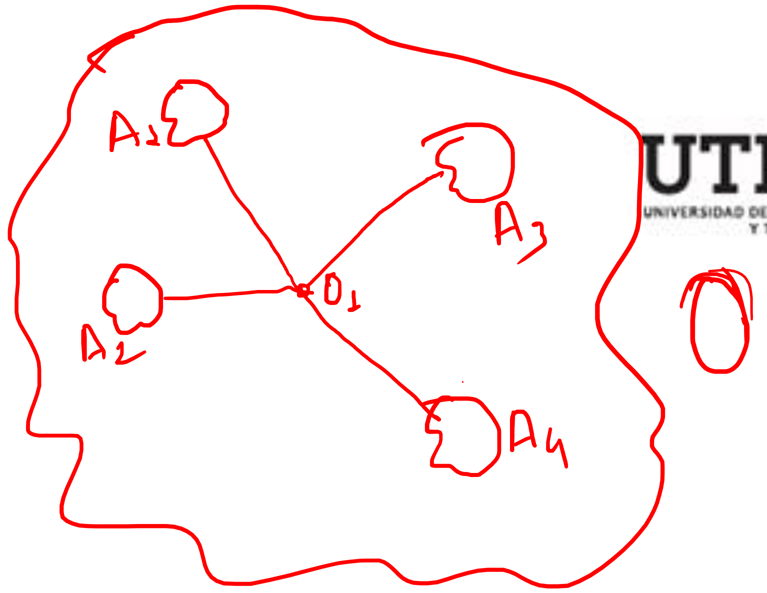
Multidimensional scaling (MDS): método de construcción de f que se basa en la minimización de la tensión

2. Lipschitz Embeddings





$d(p,O) = \min_{p_k \in O} d(p,p_k)$



Lipschitz *Embeddings*

Un **Lipschitz *embedding*** se define en términos de un conjunto R de subconjuntos de S, $R = \{A_1, A_2, \ldots, A_k\}$. Los subconjuntos A_i se denominan conjuntos de referencia de la incrustación. Sea $d(o, A)$ una extensión de la función de distancia d a un subconjunto $A \subset S$ tal que $d(o, A) = \min_{x \in A} d(o, x)$. Un *embedding* con respecto a R se define como un mapeo f tal que:

$$f(o) = \big(d(o, A_1), d(o, A_2), \ldots, d(o, A_k)\big)$$

En otras palabras, estamos definiendo un espacio de coordenadas donde cada eje corresponde a un subconjunto $A_i \subset S$ de los objetos, y los valores de coordenadas del objeto o son las distancias de o al elemento más cercano en cada uno de A_i.

Lipschitz *Embeddings*

Si d' es la métrica L_p, f se define de tal manera que $f(o) = \left(\frac{d(o,A_1)}{q}, \frac{d(o,A_2)}{q}, \dots, \frac{d(o,A_k)}{q}\right)$, donde $\boxed{q = k^{1/p}}$

$$\boxed{\frac{c}{\lceil \log_2 N \rceil} \cdot d(o_1, o_2) \leq d'\big(f(o_1), f(o_2)\big) \leq d(o_1, o_2)}$$

para cualquier par de objetos $o_1, o_2 \in S$, donde $c > 0$ es una constante.

UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

$K^{1/p}$

Lipschitz *Embeddings*

$p=2$

Ejemplo:

Object	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0	2	13	7	3	8	11	4	9	10
o_2	2	0	11	9	3	10	9	2	11	8
o_3	13	11	0	6	10	9	4	9	6	3
o_4	7	9	6	0	6	3	8	9	2	5
o_5	3	3	10	6	0	7	8	3	8	7
o_6	8	10	9	3	7	0	9	10	3	6
o_7	11	9	4	8	8	9	0	7	10	3
o_8	4	2	9	9	3	10	7	0	11	6
o_9	9	11	6	2	8	3	10	11	0	7
o_{10}	10	8	3	5	7	6	3	6	7	0

$f(o_3) = (9, 10, 3)$

$f(o_8) = (0, 3, 0)$

$A_1 = \{o_2, o_8\}$

$A_2 = \{o_1, o_5\}$

$A_3 = \{o_6, o_8, o_9, o_{10}\}$

$d(f(o_3), f(o_8)) = 11.79$

$d(o_3, o_8) = 9$

$\frac{d(f(o_3), f(o_8))}{k^{1/p}} = 5,895$

Lipschitz *Embeddings*

Ejemplo:

Object	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0	2	13	7	3	8	11	4	9	10
o_2	2	0	11	9	3	10	9	2	11	8
o_3	13	11	0	6	10	9	4	9	6	3
o_4	7	9	6	0	6	3	8	9	2	5
o_5	3	3	10	6	0	7	8	3	8	7
o_6	8	10	9	3	7	0	9	10	3	6
o_7	11	9	4	8	8	9	0	7	10	3
o_8	4	2	9	9	3	10	7	0	11	6
o_9	9	11	6	2	8	3	10	11	0	7
o_{10}	10	8	3	5	7	6	3	6	7	0

Reference Set	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
A_1	2	0	9	9	3	10	7	0	11	6
A_2	0	2	10	6	0	7	8	3	8	7
A_3	4	2	3	2	3	0	3	0	0	0
A_4	0	2	4	0	3	3	0	0	2	3

$$A_1 = \{o_2, o_8\}$$

$$A_2 = \{o_1, o_5\}$$

$$A_3 = \{o_6, o_8, o_9, o_{10}\}$$

$$A_4 = \{o_1, o_4, o_7, o_8\}$$

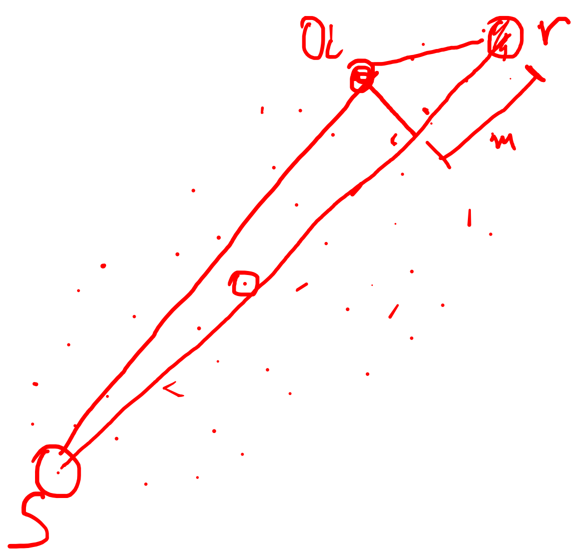
3. FastMap

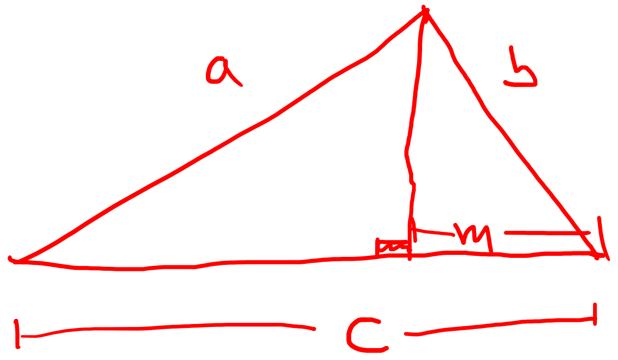
UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

Fast*Map*

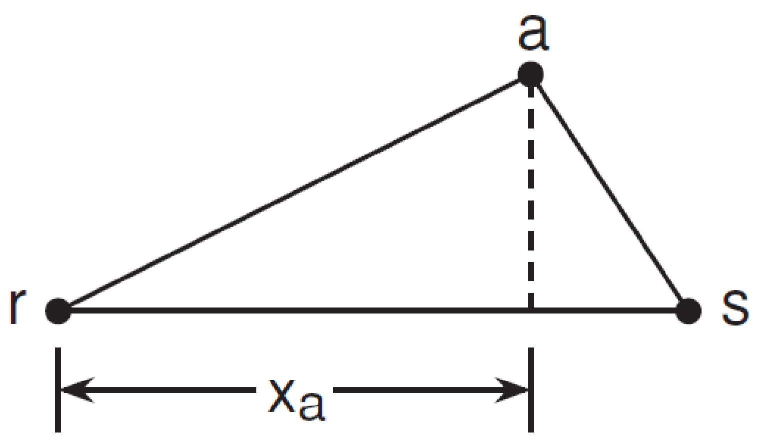
1) Objetos extremos (p_1, p_2)
2) Calcular proyección sobre $\mathcal{L}(p_1,p_2)$
3) Proyector sobre un plano perpendicular a $\mathcal{L}$

$$d'^2(O_1,O_2) = d^2(O_1,O_2) - (x_1 - x_2)^2$$

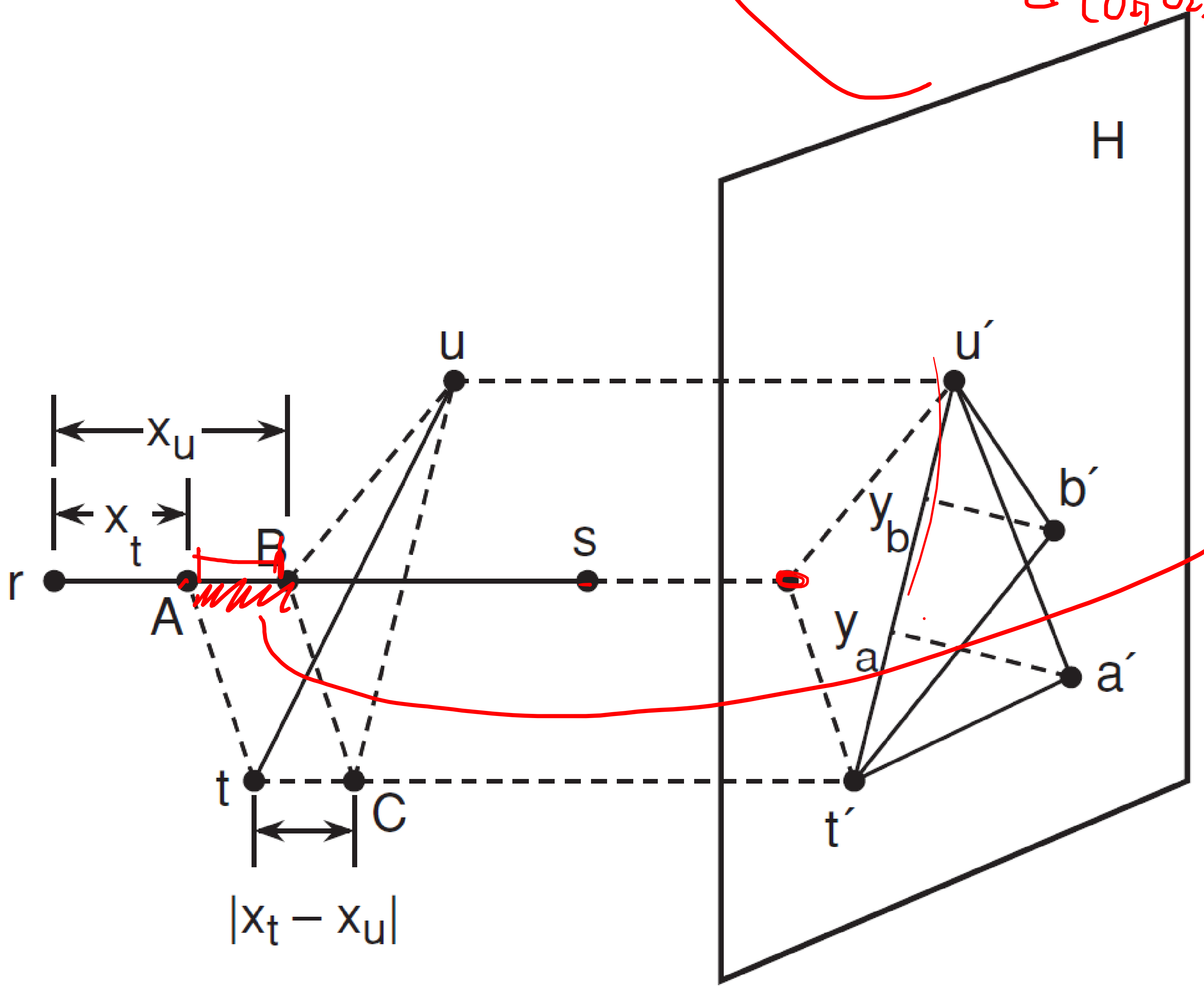




$$a^2 = b^2 + c^2 - 2cm$$



(a)



(b)

Christos Faloutsos and King-Ip Lin (1995) "FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets". Proceedings of the 1995 ACM SIGMOD international conference on Management of data (pp. 163-174).


UTEC
UNIVERSIDAD DE INGENIERÍA
Y TECNOLOGÍA
INGENIERÍA
MECATRÓNICA
BIOINGENIERÍA
CIENCIA DE
LA COMPUTACIÓN
AMBIENTAL