



Clasificación por KDE con Distancia de Fermat en variedades desconocidas: una aproximación empírica.

Lic. Gonzalo Barrera Borla (IC), Dr. Pablo Groisman (DM)
(Facultad de Ciencias Exactas y Naturales, UBA)

I. SÍNTESIS

En el presente trabajo, analizamos empíricamente el efecto de la elección de distancia en la performance de un clasificador por *Kernel Density Estimation* (KDC) en variedades desconocidas. En particular, reemplazamos la distancia euclídea con una distancia «aprendida de los datos» propuesta por Groisman et al, la *Distancia de Fermat*, e intentamos medir la mejora marginal en la exactitud («accuracy») de clasificación por sobre (a) KDC con distancia euclídea y (b) otros algoritmos estándares: SVC, regresión logística, kNN y Naive Bayes. Los resultados preliminares muestran que KDC es un método consistentemente performante, pero (a) nunca superior a SVC, y (b) que la performance de FKDC *no supera* a la de su par euclídeo, que técnicamente es un caso particular de FKDC. Finalmente, consideramos por qué sería éste el caso, y listamos oportunidades de mejora del algoritmo.

II. CONTEXTO

En aras de la brevedad, notamos aquí sólo conceptos fundamentales, siguiendo a Pelletier et al para la definición de KDC y el clasificador asociado y Groisman et al para la Distancia de Fermat.

i. KDE en variedades de Riemann

Sea (\mathcal{M}, g) resp. una variedad Riemanniana \mathcal{M} y su métrica g , compacta y sin frontera de dimensión d , y denotemos d_g la distancia cpte. Sea X un elemento aleatorio (e.a.) con soporte en \mathcal{M} y función de densidad f , y $\{X_1, \dots, X_N\}$ una muestra de e.e. aa. i.i.d. a X . Sean, además, K una «función núcleo» y $h > 0$ un «ancho de banda». Entonces, la estimación de f por KDE es la estimación de densidad por KDE es

$$\hat{f}(x) = N^{-1} \sum_{i=1}^N \frac{1}{h^d} \frac{1}{\theta_{X_i}(x)} K\left(\frac{d_g(x, X_i)}{h}\right) \quad (1)$$

donde $\theta_p(q)$ es la *función de densidad volumétrica* en \mathcal{M} alrededor de p . Obsérvese que cuando $\mathcal{M} = \mathbb{R}^d$ y g es la métrica euclídea, $\theta_p(q) = 1 \forall (p, q)$, y \hat{f} se reduce a la más conocida

$$\hat{f}(x) = N^{-1} \sum_{i=1}^N \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right) \quad (2)$$

Tomar por «núcleo gaussiano» $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$ es práctica casi universal, mientras que la elección de h es crucial para encontrar un buen estimador y está ampliamente tratada en la literatura. Menos estudiada está la elección de la distancia d_g en variedades desconocidas, y es aquí donde entra en juego la Distancia de Fermat.

ii. Clasificación por KDE

Sean ahora $k \in \mathbb{N}$ «clases», y contemos con una muestra $\{(X_1, Y_1), \dots, (X_N, Y_N)\}$, $Y_i \in \{1, \dots, k\} \forall i \in \{1, \dots, N\}$, de manera que los N elementos se separen en k submuestras de tamaño N_1, \dots, N_k , cada una soportada en su propia variedad (no necesariamente la misma), y función de densidad f_j , $j \in 1, \dots, k$. Sea p_j la proporción poblacional de la clase j en la muestra completa, que aproximamos como $\hat{p}_j = \frac{n_j}{N}$, y \hat{f}_j el estimador por KDE de f_j ya descrito. Loubes y Pelletier, basándose en el criterio de Bayes, plantean como regla de clasificación para un nuevo e.a. (x, y)

$$\hat{y} = \arg \max_{j \in 1, \dots, k} \hat{f}_j(x) \hat{p}_j = \sum_{i=1}^N \mathbb{1}\{Y_i = j\} K_h(x, X_i) \quad (3)$$

donde $\mathbb{1}\{\cdot\}$ es la función indicadora, y $K_h(x, X_i) = \frac{1}{h^d} \frac{1}{\theta_{X_i}(x)} K\left(\frac{d_g(x, X_i)}{h}\right)$.

Una implementación práctica de la regla en Ecuación 3 requiere conocer la geometría de la(s) variedad(es) en la(s) que se soportan las muestras. En la práctica, es harto común asumir la hipótesis de la variedad, pero desconocer su exacta geometría. En tal contexto, una alternativa es *aprender la distancia de los datos*. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua quaeerat.

iii. Aprendizaje de Distancias: Isomap, Distancia de Fermat

Un punto de partida natural, es asumir que los elementos muestrales $X_i \in \mathcal{M}$, y que si la variedad es suficientemente regular, el segmento $\overline{X_i X_j}$ también pertenece a \mathcal{M} . Con esta lógica, Tenenbaum et al [1] desarrollan Isomap, un algoritmo precursor en el aprendizaje de distancias:

1. Construya G , el grafo de k o ε vecinos más cercanos y pese cada arista por su métrica euclídea,
2. Tome por distancia aprendida d entre dos nodos la longitud del camino mínimo entre ellos,
3. Compute una representación de menor dimensión en un espacio euclídeo.

Más allá de su efectividad, tal algoritmo no deja de ser una heurística inteligente, y depende crucialmente de una correcta elección del parámetro de cercanía (k/ε). En una propuesta similar pero superadora, Groisman et al [2] proponen la «Distancia de Fermat», una distancia propiamente dicha en variedades, y muestran cómo ésta se puede aproximar «microscópicamente» a partir de una muestra. Sea Q el grafo completo de la muestra, y $\alpha \geq 1$, luego

$$D_{Q,\alpha}(x, y) = \inf \left\{ \sum_{i=1}^K \|q_{i-1} - q_i\|^\alpha : (q_0, \dots, q_K) \text{ es un camino de } x \text{ a } y \right\} \quad (4)$$

es la «distancia muestral» de Fermat. Nótese que usar el grafo completo obvia la necesidad de elegir (k/ε), mientras que $\alpha > 1$ «infla» el espacio y desalienta los «saltos largos» por espacio vacío. Cuando $\alpha = 1$, la distancia de Fermat se reduce a la distancia euclídea.

III. PROPUESTA

En la tesis desarrollamos un clasificador compatible con el *framework* de *scikit-learn* según los lineamientos de [3]. Asumiendo que \mathcal{M} es \mathbb{R}^d , obtenemos un clasificador que apodamos KDC. Luego, implementamos el estimador de Ecuación 4, y reemplazamos la distancia euclídea de KDC por la distancia muestral de fermat, para compeltar nuestra propuesta, FKDC.

IV. METODOLOGÍA

Deseamos evaluar la *exactitud* («accuracy») de los clasificadores propuestos en diferentes *datasets*, relativa a técnicas bien establecidas:

- regresión logística (LR)
- k-vecinos-más-cercanos (KN)
- clasificador de soporte vectorial (SVC)
- Naive Bayes Gaussiano (GNB)

Para comparar equitativamente estos *algoritmos* de clasificación,

- partiremos la muestra en entrenamiento y testeo,
- elegiremos los *hiperparámetros óptimos* por *validación cruzada en 5 pliegos* entre los datos de entrenamiento, y
- mediremos la exactitud de todos los algoritmos en el mismo conjunto de testeo.

V. ANÁLISIS

i. Fan

Para tener una idea «sistémica» de la performance de los algoritmos, evaluaremos su performance con diferentes *datasets*. Muchos factores en la definición de un dataset pueden afectar la exactitud de la clasificación; nos interesará explorar en particular 3 que a su vez figuran en el cálculo de la densidad en variedades:

- n , el tamaño de la muestra,
- d , la dimensión de las observaciones y
- k , la cantidad de categorías.

VI. RESULTADOS Y ANÁLISIS

i. Fantasías en 2D

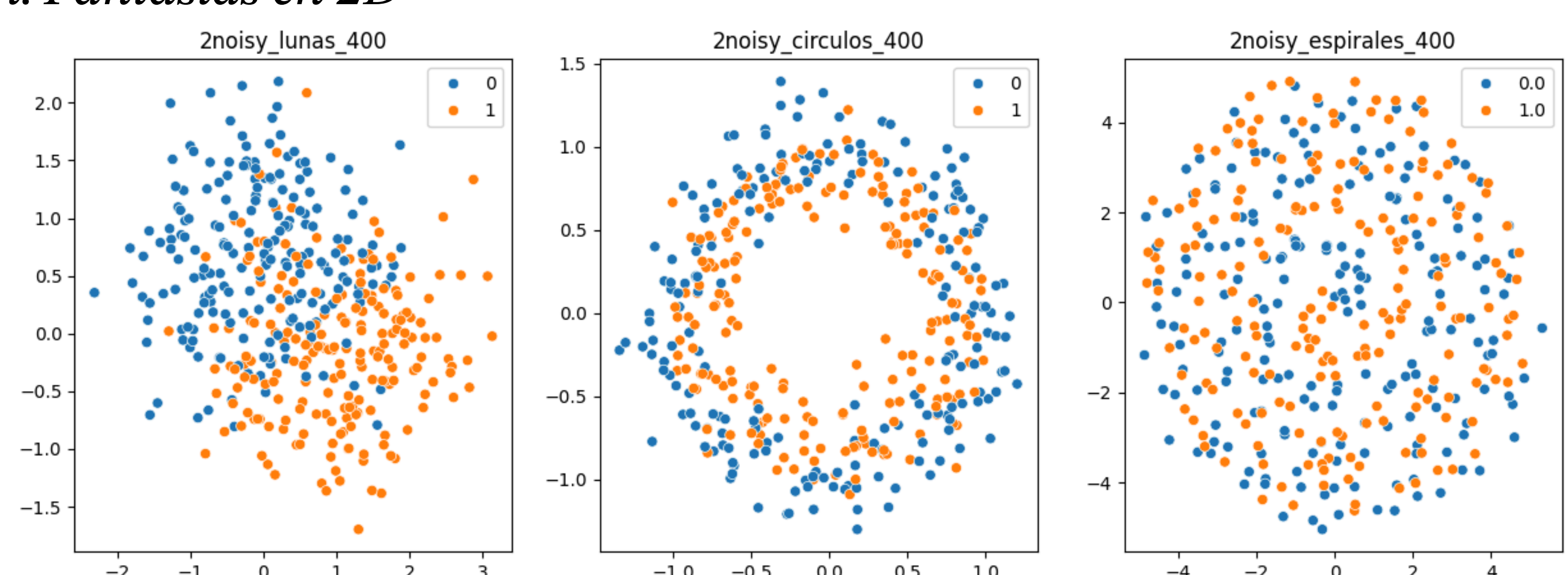


Figura 1: Datasets sintéticos en \mathbb{R}^2

ii. Iris

iii. digits

VII. OBSERVACIONES GENERALES

VIII. PREGUNTAS ABIERTAS

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aequaleamur animo, cum corpore dolemus, fieri.

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.
2. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.
3. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.

Lorem ipsum dolor sit.	Lorem ipsum.	Lorem ipsum.
Lorem ipsum dolor.	Lorem ipsum.	α
Lorem ipsum.	Lorem.	β
Lorem.	Lorem.	γ
Lorem ipsum.	Lorem ipsum dolor.	θ

$$\begin{pmatrix} 1 & 2 & \dots & 8 & 9 & 10 \\ 2 & 2 & \dots & 8 & 9 & 10 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 10 & 10 & \dots & 10 & 10 & 10 \end{pmatrix}$$

(5)

i. Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aequaleamur animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre.

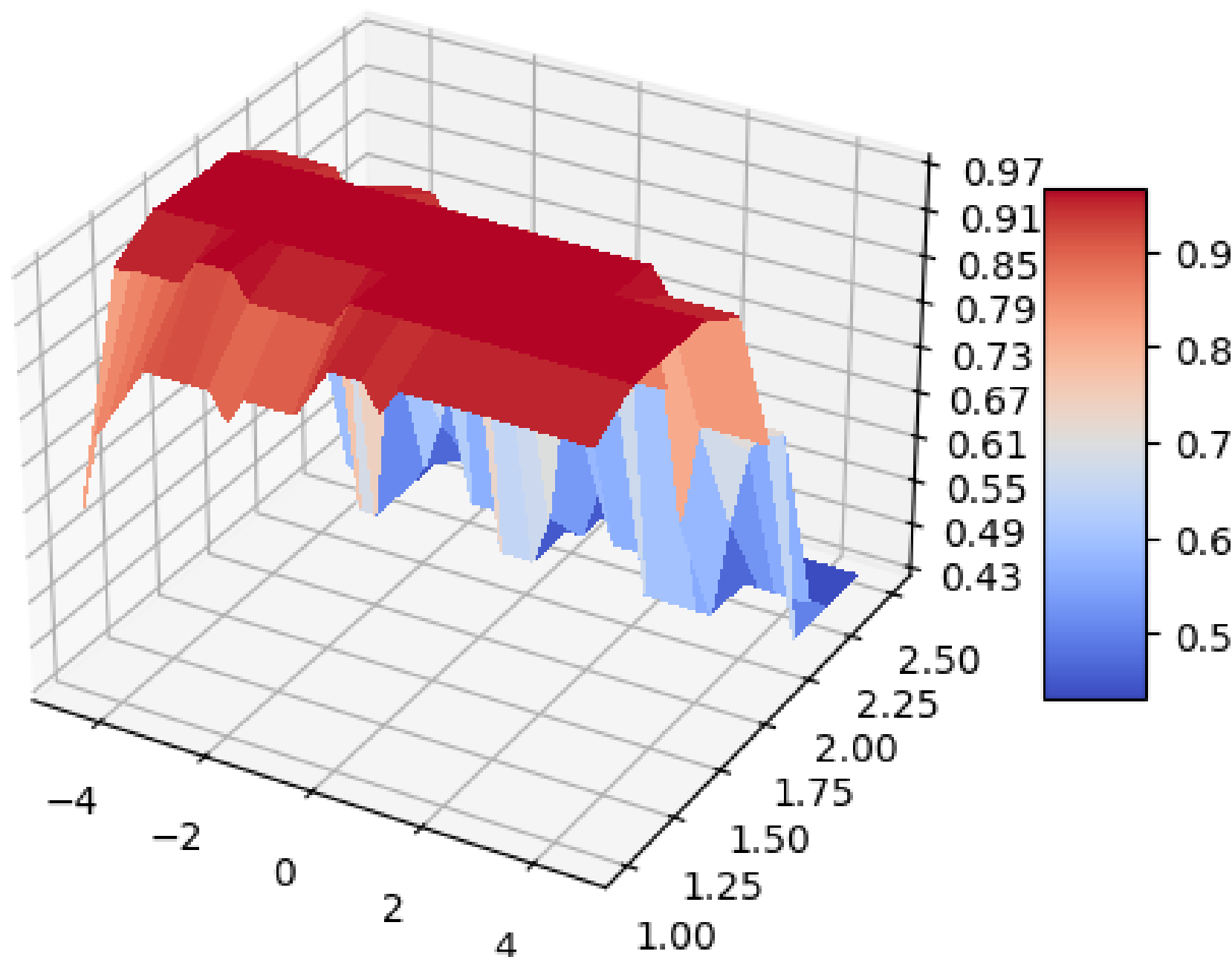


Figura 2: Lorem ipsum dolor sit amet, consectetur adipiscing elit.

IX. LOREM IPSUM DOLOR.

- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.,
- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.,
- Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do.,

BIBLIOGRAFÍA

[1] J. B. Tenenbaum, V. de Silva, y J. C. Langford, «A Global Geometric Framework for Nonlinear Dimensionality Reduction», *Science*, vol. 290, n.º 5500, pp. 2319-2323, dic. 2000, doi: 10.1126/science.290.5500.2319.

[2] P. Groisman, M. Jonckheere, y F. Sapienza, «Nonhomogeneous Euclidean First-Passage Percolation and Distance Learning», n.º arXiv:1810.09398. arXiv, diciembre de 2019.

[3] J.-M. Loubes y B. Pelletier, «A Kernel-Based Classifier on a Riemannian Manifold», *Statistics & Decisions*, vol. 26, n.º 1, pp. 35-51, mar. 2008, doi: 10.1524/std.2008.0911.