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 $\{X_1, ..., X_N\}$ una muestra de ee. 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

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

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

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

Tomar por «núcleo gaussiano» $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$ 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 $\{(X_1, Y_1), ..., (X_N, Y_N)\}, Y_i \in \{1, ..., k\} \forall i \in \{1, ...N\}$, de manera que los N elementos se separen en k submuestas de tamaño $N_1, ..., N_k$, cada una soportada en su propia variedad (no encesariamente la misma), y función de densidad $f_j, j \in 1, ..., k$. 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)

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

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 magnam aliquam quaerat.

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

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

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`)
- clasificador de soporte vectorial (`SVC`)
- k-vecinos-más-cercanos (`KN`)
- 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. Análisiss

Para tener una idea «sistémica» de la performance de los algoritmos, evaluaremos su performance con diferentes *datasets*. $\mathcal{M}$chos 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:, el tamaño de la muestra n, la dimensión p de las observaciones y la cantidad k de categorías.

i. Fantasías en $\mathbb{R}^2$

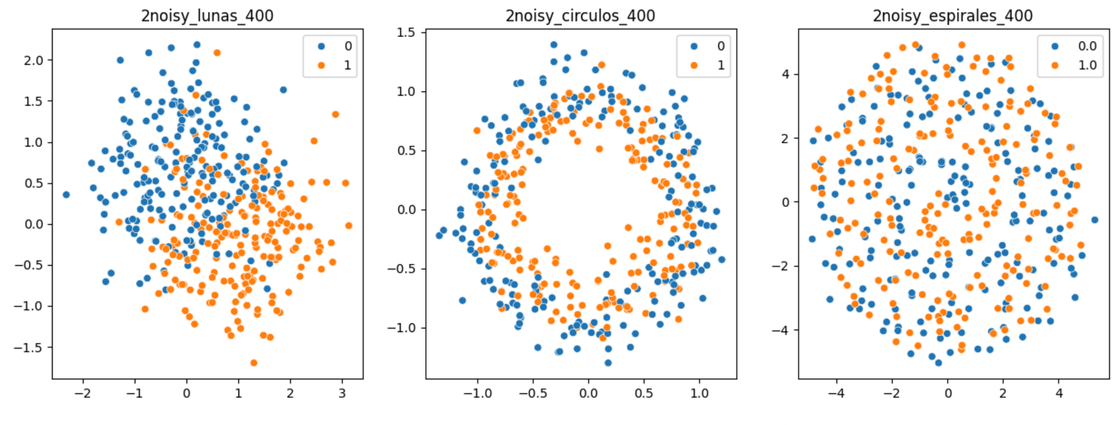

Figura 1: Datasets sintéticos en $\mathbb{R}^2$

Ruido	Dataset	FKDC	GNB	KDC	KN	LR	SVC
Alto	Circulos	67.2 (4.4)	63.5 (7.0)	67.0 (4.3)	67.3 (4.5)	44.8 (4.6)	71.3 (5.1)
	Espirales	76.2 (4.8)	48.5 (6.2)	76.6 (4.4)	76.0 (5.2)	48.6 (5.7)	78.7 (4.0)
	Lunas	79.7 (5.6)	80.4 (3.9)	81.3 (4.8)	80.9 (4.4)	80.7 (3.9)	81.2 (5.0)
Bajo	Circulos	78.4 (4.1)	67.7 (11.3)	78.5 (4.1)	79.1 (4.2)	45.0 (4.5)	81.2 (5.4)
	Espirales	90.0 (3.2)	49.6 (6.2)	90.4 (3.2)	90.3 (2.9)	49.5 (6.5)	92.9 (1.7)
	Lunas	88.0 (4.6)	83.6 (4.3)	88.1 (4.6)	87.8 (4.6)	83.9 (4.0)	88.0 (3.7)

Exactitud (espresada en porcentaje), con sus respectivos desviós estándares a lo largo de 16 repeticiones de cada experimento.

Los tres datasets, `lunas, circulos, espirales`, tienen $k = 2, p = 2, n = 400, n_1 = n_2 = 200$, y presentan variedades de dimensión intrínsica $d = 1$, a las cuales se les agrego «ruido» gaussiano con «bajo» y «alto» desvío estándar ($\sigma_{\text{alto}} \approx 1.5\sigma_{\text{bajo}}$). En los tres datasets, la performance de SVC es consistentemente la mejor, aunque KN, KDC y FKDC no son significativamente distintos si consideramos un intervalo de confianza razonable. Es alentador ver que la performance de KDC es siempre com-

petitiva, pero descorazonador ver que FKDC es sistemáticamente igual o ligeramente peor que KDC.

ii. `vino`, `pinguinos`, `iris` *y* `anteojos`

El siguiente conjunto de datos contiene $k = 3$ con diferente cantidad de predictores y n. Salvo por «anteojos», todos los datasets son pequeños pero reales.

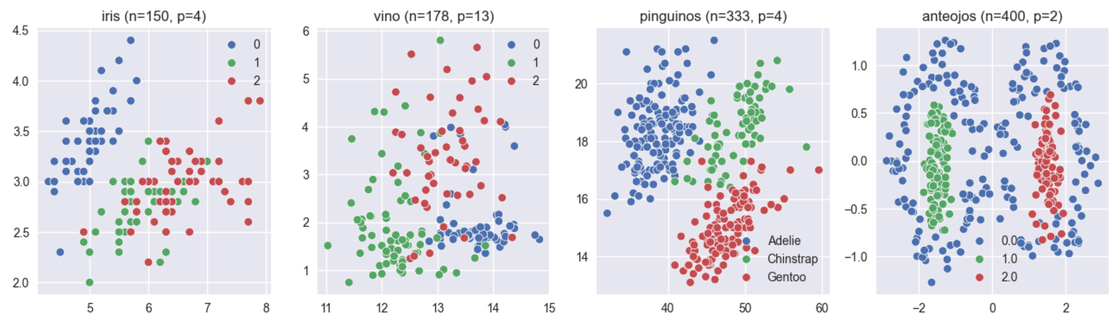

Figura 2: Datasets con $k = 3$

Dataset	FKDC	GNB	KDC	KN	LR	SVC
Anteojos	97.5 (1.4)	97.0 (1.8)	97.4 (1.4)	97.7 (1.4)	50.5 (5.0)	97.7 (1.8)
Iris	94.4 (4.3)	94.6 (5.0)	94.0 (4.4)	95.4 (4.0)	97.5 (2.3)	94.2 (5.8)
Pinguinos	84.0 (4.2)	97.8 (1.6)	84.1 (4.2)	85.2 (3.8)	66.6 (4.5)	98.2 (1.0)
Vino	71.9 (7.1)	96.9 (2.2)	73.8 (6.3)	71.0 (6.5)	66.0 (6.7)	95.3 (2.6)

En los datasets de `anteojos` e `iris`, se observa el mismo fenómeno que en los datasets «2D»: (F)KDC es competitivo con los mejores métodos (SVC y LR, resp.), pero no superador. En los datasets de `pinguinos` y `vino`, la *performance* de los métodos propuestos es significativamente peor. En todos los casos, no conseguimos mejoras significativas sobre KDC con FKDC.

iii. digitos

Los ee.aa. a estudiar son imaágenes de 8x8 (*id est*, en $\mathbb{R}^{64}$) que representan dígitos manuscritos. En este caso, aunque $p = 64$, es de esperar que la variedad donde yacen los trazos sea de mucha menor dimensión, y mejores resultados esperaríamos de la «estimación de la variedad» que promete FKDC.

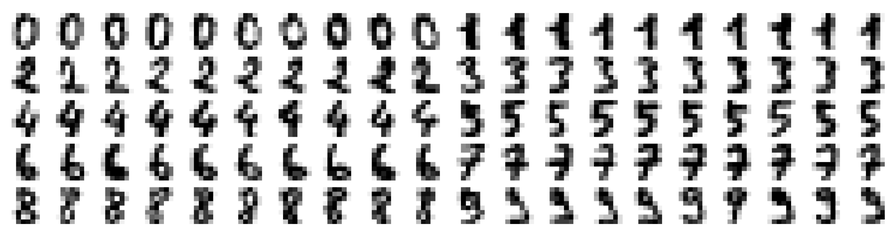

Figura 3: Dígitos manuscritos en B&N, 8x8 píxeles

Eval.	FKDC	GNB	KDC	KN	LR	SVC
20%	98.8 (0.7)	92.1 (1.1)	98.9 (0.6)	98.9 (0.4)	96.7 (0.6)	99.0 (0.6)
80%	97.0 (0.4)	90.2 (0.6)	96.9 (0.5)	96.6 (0.8)	94.5 (0.7)	97.5 (0.4)

VI. Observaciones Generales

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/stnd.2008.0911.