

**Proyecto de tesis**  
**para optar al título de Magister en Estadística Matemática**

**Nombre del postulante:** Lic. Gonzalo Barrera Borla

**Directora:** Dr. Pablo Groisman

**Tema de trabajo:** Distancia de Fermat en Clasificadores de Densidad Nuclear

**Lugar de trabajo:** Departamento de Matemática

## 1 Antecedentes existentes sobre el tema

El concepto de *distancia* entre las observaciones disponibles, es central a casi cualquier tarea estadística, tanto en descripción como inferencia. Consideremos, por caso, un ejercicio de clasificación. Sea  $\mathbf{x} = (x_i)_{i=1}^N$  una muestra de  $N$  observaciones de vv.aa. i.i.d.  $(X_1, \dots, X_N : X_i \sim \mathcal{L}(X) \ \forall i \in [N])$ , con  $x_i \in \mathbb{R}^{d_x} \ \forall i \in [N] \equiv \{1, \dots, N\}$ , donde cada observación pertenece a una de  $M$  clases  $C_1, \dots, C_M$  mutuamente excluyentes y conjuntamente exhaustivas.

Dada una nueva observación  $x$  cuya clase es desconocida ¿a qué clase deberíamos asignarla? Cualquier respuesta a esta pregunta implicará combinar toda la información muestral disponible, ponderando las  $N$  observaciones de manera relativa a su cercanía o similitud con  $x$ . Cuando el dominio de las  $x_i$  es un espacio euclídeo  $\mathbb{R}^{d_x}$ , es costumbre tomar la *distancia euclídea* para cuantificar la cercanía entre elementos. Así, por ejemplo,  $k$ -vecinos más cercanos ( $k$ -NN) asignará la nueva observación  $x$  a la clase modal entre las  $k$  observaciones de entrenamiento más cercanas (es decir, que minimizan  $\|x - \cdot\|$ ).

Una dificultad bien conocida con los métodos basados en distancias, es la *maldición de la dimensionalidad*: a medida que la dimensión  $d_x$  del espacio euclídeo en consideración crece, el espacio se vuelve tan grande que todos los elementos de la muestra están indistinguiblemente lejos entre sí; o lo que es equivalente, a igual  $N$ , la densidad de observaciones en el espacio decae exponencialmente con  $d_x$ .

En estos casos, es de suponer que el dominio de las  $X$  no cubre *todo*  $\mathbb{R}^{d_x}$ , sino que éstas se encuentran embebidas en una variedad  $\mathcal{M} \subset \mathbb{R}^{d_x}$  cuya dimensión intrínseca  $\dim(\mathcal{M})$ , es potencialmente mucho menor a  $d_x$ , y por ende la distancia *en la variedad* es más informativa que la distancia (euclídea) en el espacio ambiente  $\mathbb{R}^{d_x}$ . A este supuesto se lo suele llamar “hipótesis de la variedad” (*manifold hypothesis*), y suele ser particularmente acertado cuando las observaciones provienen “del mundo real” (e.g., imágenes, sonido y texto). Según Bengio et al. (2013), *aprender* la estructura de  $\mathcal{M}$  a partir de  $\mathbf{x}$  es una forma (entre muchas) de *aprendizaje de representaciones* (representation learning), donde la representación de  $x_i$  en base a sus coordenadas en  $\mathcal{M}$  (en  $\mathbb{R}^d$ ) es tanto o más útil que la representación original en  $\mathbb{R}^{d_x}$  para tareas de descripción e inferencia.

La ganancia en reducción de dimensionalidad con la hipótesis de la variedad debe ser contrastada con la dificultad extra de tener que trabajar en una variedad arbitraria  $\mathcal{M}$  en lugar de  $\mathbb{R}^{d_x}$ , a priori desconocida y que debemos estimar. Pelletier (2005) describe un estimador “nuclear” para la función de densidad de vv.aa. i.i.d. en variedades Riemannianas compactas sin borde, junto con resultados de consistencia y convergencia; Henry & Rodriguez (2009) los amplían para probar la consistencia uniforme fuerte y la distribución asintótica de estos estimadores.

Tanto Pelletier (2005) como Henry & Rodriguez (2009) asumen que la distancia geodésica es conocida. Trabajos recientes (Sapienza et al. (2018); Groisman et al. (2022); McKenzie & Damelin (2019); Little et al. (2022)) proponen aprender una distancia geodésica  $\mathcal{D}_f^p$  entre los nodos del grafo (aleatorio) completo de la muestra  $\mathbb{X}_n^1$ , con cada arista pesada por una potencia  $p$  de la distancia

<sup>1</sup>O por simplicidad de cómputo, su aproximación por el grafo de  $k$ -vecinos más cercanos

euclídea entre sus extremos. En Sapienza et al. (2018), el uso de esta distancia - que los autores llaman “de Fermat”, por su analogía con el fenómeno óptico -, parece rendir considerables mejoras de *performance* empírica en tareas de clasificación. Cuando  $p = 1$ , el estimador de la distancia geodésica  $\mathcal{D}_f^1$  resultante es idéntico al que usa *Isomap* (Tenenbaum et al. (2000)) para construir los *embeddings* de dimensión reducida.

## 2 Naturaleza del aporte original sobre el tema y objetivos

Uniando los elementos enunciados anteriormente, nos proponemos estudiar sistemáticamente qué valor aporta el uso de una distancia basada en datos (la distancia de Fermat  $\mathcal{D}_f^p$ ) frente a la elección canónica (la distancia euclídea  $\|x - \cdot\|$ ), en el aprendizaje de *estimadores de densidad nuclear* (KDEs, por sus siglas en inglés). Nos proponemos luego comparar sus bondades relativas usándolos en tareas de *clasificación* bajo una amplia gama de condiciones:

- en datasets “reales” y “sintéticos”,
- en relación a la dimensión  $d_x$  del espacio ambiente y
- en relación a las  $k$  categorías posibles para  $Y \in \{C_1, \dots, C_k\}$

Aprender un clasificador a partir de KDEs con distancia euclídea (Hastie et al. (cap. 6.6, 2009)) es un método bastante eficiente en términos de cómputo. En cambio, un calculo exacto del estimador muestral de  $\mathcal{D}_f^p$  requiere  $n^3$  pasos. La pregunta al respecto de su eficacia, entonces, debe considerar además comparativamente los costos computacionales de ambas distancias, que en datasets “grandes” podrían ser demasiado altos para obtener ganancias de *performance* relativamente menores. Para poner en contexto la capacidad predictiva de estos clasificadores y su costo computacional, incluiremos como métodos de referencias

- clasificadores de *Naive Bayes* (Hastie et al. (cap. 6.6.3, 2009)), que usan  $d$  KDEs unidimensionales en lugar de un KDE  $d$ -dimensional por clase),
- *gradient boosting trees* (GBTs, Hastie et al. (cap. 10, 2009)), un método reconocido en la actualidad por su simplicidad de uso y escasez de requerimientos y
- *random forests* Hastie et al. (cap. 15, 2009), que capturan buena parte de las bondades de los GBTs con una estructura sencilla.

Además, nos proponemos dar argumentos teóricos que garanticen la consistencia de la estimación de la densidad en el estilo de Pelletier, cuando se reemplaza la distancia geodésica por una estimación empírica (plug-in).

## References

- Y. Bengio, A. Courville, & P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, aug 2013. doi: 10.1109/tpami.2013.50.
- Pablo Groisman, Matthieu Jonckheere, & Facundo Sapienza. Nonhomogeneous euclidean first-passage percolation and distance learning. *Bernoulli*, 28(1), feb 2022. doi: 10.3150/21-bej1341.

- Trevor Hastie, Robert Tibshirani, & Jerome Friedman. *Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer London, Limited, 2009. ISBN 9780387216065.
- Guillermo Henry & Daniela Rodriguez. Kernel density estimation on riemannian manifolds: Asymptotic results. *Journal of Mathematical Imaging and Vision*, 34(3):235–239, feb 2009. doi: 10.1007/s10851-009-0145-2.
- Anna Little, Daniel McKenzie, & James M. Murphy. Balancing geometry and density: Path distances on high-dimensional data. *SIAM Journal on Mathematics of Data Science*, 4(1):72–99, jan 2022. doi: 10.1137/20m1386657.
- Daniel Mckenzie & Steven Damelin. Power weighted shortest paths for clustering euclidean data. May 2019.
- Bruno Pelletier. Kernel density estimation on riemannian manifolds. *Statistics & Probability Letters*, 73(3):297–304, jul 2005. doi: 10.1016/j.spl.2005.04.004.
- Facundo Sapienza, Pablo Groisman, & Matthieu Jonckheere. Weighted geodesic distance following fermat’s principle. 2018. URL <https://openreview.net/forum?id=BJfaMIJwG>.
- Joshua B. Tenenbaum, Vin de Silva, & John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, dec 2000. doi: 10.1126/science.290.5500.2319.