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: Aprendizaje de distancia basada en datos para clasificación por densidad de núcleos. // 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 $\boldsymbol{x} = (x_i)_{i=1}^N$una muestra de N observaciones de vv.aa. i.i.d. $(X_1, \dots, X_n : X \sim \mathcal{L}(X) \ \forall \ i \in [N])$, 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, codificadas en un vector $\boldsymbol{y} = (y_1, \dots, y_N)$ donde $y_i = j \iff x_i \in C_j$.

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 (kNN) 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 cae 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 $\dim(\mathcal{M}) = d_{\mathcal{M}}$, donde potencialmente $d_{\mathcal{M}} \ll d_x$, y por ende la distancia *en la variedad* es más inormativa 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, *aprender* la estructura de $\mathcal{M}$ a partir de $\boldsymbol{x}$ es una forma (enre 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, se compensa 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 compactas de Riemann sin fronteras, junto con resultados de consistencia y convergencia; Henry y Rodriguez [2009] los amplían para probar la consistencia uniforme fuerte y la distribución asintótica de $f_{N,k}$.

Sin embargo, tanto Pelletier Henry y Rodríguez discuten “un estimador de densidad [...] basado en núcleos que son funciones de *la distancia geodésica Riemanniana en la variedad*, cuya expresión es consistente con su equivalente en el caso euclídeo”.[p. 298 intro, Pelletier 2005]. Ahora, la distancia euclídea en $\mathbb{R}^{d_x}$, $|\cdot|_{d_x}$ (que podemos calcular fácilmente) no tiene por qué coincidir con la distancia geodésica en $\mathcal{M}$, $dg_{\mathcal{M}}$ (que nos interesa). Entre otros factores, tanto la diferencia en dimensiones $d_x - d_{\mathcal{M}}$ como la curvatura de $\mathcal{M}$ afectan drásticamente la relación entre $|\cdot|_{d_x}$ y $dg_{\mathcal{M}}$.

Trabajos recientes [McKenzie 2019, Little 2021, Sapienza 2018, Groisman 2019] proponen aprender la distancia $dg_{\mathcal{M}}$ a partir del del grafo (aleatorio) completo de la muestra $\mathbb{X}_n$, y las geodésicas (caminos mínimos) entre sus elementos, que resultan de pesar las aristas de acuerdo a cierta potencia p de la distancia euclídea entre sus extremos. Empíricamente, el uso de esta distancia basada en datos que los autores llaman "distancia de Fermat" [Groisman 2019].

2 Naturaleza del aporte original sobre el tema y objetivos

Uniendo los elementos enunciados anteriormente, nos proponemos una comparación sistemáticamente entre el uso de la distancia de Fermat $\mathcal{D}_f$ [McKenzie 2019, Little 2021, Sapienza 2018, Groisman 2019], aprendiéndola de los datos disponibles, contra la elección canónica (la distancia euclídea), para el cómputo de densidades por núcleo *en variedades de dimensión desconocida*, $d_{\mathcal{M}} \leq d_x$ según propone [Pelletier 2005] y expande [Henry y Rodriguez 2009]. Como prueba para comparar sus ventajas relativas, aplicaremos los estimadores de densidad (KDEs, "kernel density estimator(s)") resultantes, a tareas varias de *clasificación* bajo una amplia gama de condiciones:

- datasets "reales" y "sintéticos",
- tanto en relacióna la dimensión D de $X \in \mathbb{R}^D$, como
- las k categorías posibles para $Y \in \{C_1, \ldots, C_k\}$)
- distintas razones entre $d : D$

La KDE con distancia euclídea es un método bastante eficiente en términos de cómputo. En cambio, un calculo exacto del estimador muestral de $\mathcal{D}_f$ requiere n^3 pasos. La pregunta al respecto de su eficacia, entonces, debe considerar además comparativamente los costos computacionales de ambos métodos, que en datasets lo suficientemnete grandes podrían no justificar ganancias de *performance* relativamente menores. En este aspecto, y para poner en contexto la capacidad predictiva de KDE, incluiremos como método de referencia (tanto estadística como computacionalmente) a los *gradient boosting trees*, un método reconocido en a actualidad por su simplicidad de uso y escasez de requerimientos [ESL Cap. 9, Additive Trees and Boosting Methods].

En el orden teórico, nos proponemos ???????

References

[Gr19] NONHOMOGENEOUS EUCLIDEAN FIRST-PASSAGE PERCOLATION AND DISTANCE LEARNING P. GROISMAN, M. JONCKHEERE, AND F. SAPIENZA

[1] W EIGHTED G EODESIC D ISTANCE F OLLOWING F ERMAT ' S P RINCIPLE Pablo Groisman IMAS-CONICET, NYU-ECNU IMS at NYU Shanghai, and Universidad de Buenos Aires, Argentina. pgroisma@dm.uba.ar Facundo Sapienza Aristas SRL, Buenos Aires, Argentina. f.sapienza@aristas.com.ar Matthieu Jonckheere

[2] 2019 Power Weighted Shortest Paths for Clustering Euclidean Data Daniel Mckenzie 1 ∗1 and Steven Damelin

[3] Balancing Geometry and Density: Path Distances on High-Dimensional Data Anna Little $*$ Daniel McKenzie † James M. Murphy ‡ June 9, 2021

[4] ESL II

[5] 2014 Representation Learning: A Review and New Perspectives Yoshua Bengio † , Aaron Courville, and Pascal Vincent †

[6] Kernel Density Estimation on Riemannian Manifolds: Asymptotic Results Guillermo Henry · Daniela Rodriguez Published online: 21 February 2009

[7] Kernel density estimation on Riemannian manifolds Bruno Pelletier 2005

[8] ACA PRA ABAJO NO SE USARON

[9] Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting José E. Chacón and Tarn Duong

[10] A Comprehensive Approach to Mode Clustering Yen-Chi Chen, and Christopher R. Genovese, and Larry Wasserman 2015

[11] Nonparametric Density Estimation for High-Dimensional Data - Algorithms and Applications Zhipeng Wang $*$ and David W. Scott †