Distancia de Fermat en Clasificadores de Densidad Nuclear

Lic. Gonzalo Barrera Borla

Buenos Aires, 02/03/23

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Instituto del Cálculo

Tesis presentada para optar al título de Magíster en Estadística Matemática de la Universidad de Buenos Aires

Director: Dr. Pablo Groisman

Abstract

TODO

Contents

1 Introducción

1.1 El problema de clasificacion

Consideremos el problema de clasificación:

Definition 1. (Problema de clasificación). Sea $\boldsymbol{x} = (x_i)_{i=1}^N$ una muestra de N observaciones, repartidas en M clases $C_1, \ldots, C_M$ mutuamente excluyentes y conjntamente exhaustivas (es decir, $\forall\ i \in [N] \equiv \{1, \ldots, N\}, x_i \in C_j \iff x_i \notin C_k, k \in [M], k \neq j$). Asumamos además que la muestra está compuesta de observaciones independientes entre sí, y en particular, cada clase tiene su propia ley: si $\|C_j\| = N_j$ y $x_i^{(j)}$representa la i-ésima observación de la clase j, resulta que $X_i^{(j)} \sim \mathcal{L}_j(X)\ \forall\ j \in [M], i \in [N_j]$.

Dada una nueva observación x_0 cuya clase es desconocida,

1. (clasificación dura) ¿a qué clase deberíamos asignarla?
2. (clasificación suave) ¿qué probabilidad tiene de pertenecer a cada clase $C_j, j \in [M]$?

Todo método o algoritmo que pretenda responder el problema de clasificación, prescribe un modo u otro de combinar toda la información muestral disponible, ponderando las N observaciones de manera relativa a su cercanía o similitud con x_0. Por caso, $k-$vecinos más cercanos ($k-$NN) asignará la nueva observación x_0 a la clase modal entre las k observaciones de entrenamiento más cercanas (es decir, que minimizan la distancia euclídea $\|x_0 - \cdot\|$). $k-$NN no hace ninguna mención explícita de las leyes de clase $\mathcal{L}_j$, lo cual lo mantiene sencillo a costa de ignorar la estructura del problema.

1.2 Estimación de densidad

Una familia bastante genérica de métodos para resolver el problema de calsificación, consisten aproximadamente de los siguientes pasos:

1. Hacer algunos supuestos sobre la forma de las leyes $\mathcal{L}_j$
2. Hallar estimadores $\hat{\mathcal{L}}_j$ de cada ley $\mathcal{L}_j$ usando las muestras de cada clase, $\boldsymbol{x}^{(j)} = \left(x_i^{(j)}\right)_{i=1}^{N_j}$ y algún procedimiento estándar (e.g.: máxima verosimilitud)
3. Definir una regla de decisión $\mathcal{R}\left(\cdot|\hat{\mathcal{L}}_j, j \in [M]\right) : \mathbb{R}^{d_x} \to [M]$ que dados los estimadores de (2), asigne la observación x_0 a la clase $\mathcal{R}(x_0)$.

Esta familia de clasificadores, se distinguen por una explícita *estimación de densidades* que más tarde se utilizarán para la tarea de clasificación en sí. Por

ejemplo, al considerar el problema de clasificación binaria, el análisis de discriminante lineal (LDA) de Fisher[1] queda encuadrado en esta familia de la siguiene manera:

En (1), asumimos que las las leyes $\mathcal{L}_j$

(a) son todas distribuciones normales con media μ_j y

(b) homocedásticas: $\Sigma_j = \Sigma \; \forall \; j \in [M]$).

En (2), estimamos $\hat{\mu_j}, \hat{\Sigma}$ por máxima verosimilitud,
$\hat{\mu_j} = N_j^{-1} \sum_{i=1}^{N_j} x_i^{(j)}$
$\hat{\Sigma} = N^{-1} \sum_{j=1}^{M} \sum_{i=1}^{N_j} (x_i^{(j)} - \hat{\mu_j})(x_i^{(j)} - \hat{\mu_j})$.
Y la regla de (3) es la indicadora $1\,(\cdot)$ del discriminante lineal

$$\begin{aligned} \mathcal{R}\,(x) &= 1\,(w \cdot x > c) \\ w &= \Sigma^{-1}(\mu_1 - \mu_0) \\ c &= w \cdot \frac{1}{2}(\mu_1 + \mu_0) \end{aligned}$$

con los parámetros μ_j, Σ reemplazados por las estimaciones de (2).

Inevitablemente, existe un *trade-off* entre lo restrictivo de los supuestos de (1), y la generalidad del clasificador resultante. En el caso de LDA, los supuestos (leyes normales y homocedasticidad) son inverosímiles en casi cualquier escenario real, pero el clasificador resultante es muy sencillo de computar. En general, este será el caso para todos los métodos *paramétricos* de estimación de densidad, en que de todas las posibles funciones de densidad, quedan acotadas a aquellas que se pueden expresar de forma cerrada con una expresión predefinida (en este caso, la densidad normal), y Q parámetros (aquí, μ y Σ).

Alternativamente, existen métodos en que los supuestos de (1) se obvian del todo, o al menos son lo suficientemente generales como para representar todas salvo las más patológicas leyes (e.g.: asumir que la media y dispersión son finitas). A estos se los conoce, naturalmente, como métodos *no paramétricos* de estimación de densidad.

Estimación de densidad por núcleos

La estimación de densidad por núcleos (o KDE, por sus siglas en inglés), es uno de los métodos mejor estudiados dentro del amplio universo no-paramétrico[2]. Introducidos hacia 1960 (Rosenblatt 1958, Parzen 1962) para variables aleatorias unidimensionales, han sido ampliamente desarrollados y adaptados a espacios mucho más generales. El objetivo es encontrar un estimador *suave* de la densidad poblacional f de una v.a. X a partir de una muestra discreta, usando una función no-negativa K llamada *núcleo* ("kernel") y un parámetro de suavización h, el *ancho de banda* ("bandwith").

[1] https://en.wikipedia.org/wiki/Linear_discriminant_analysis
[2] Algo sobre NNs, otros metodos nopa

Definition 2. (función núcleo) Una función ϕ es un *núcleo* ("kernel"), si

- toma únicamente valores reales no-negativos: $\phi(x) \geq 0 \forall x$,
- está normalizada: $\int_{-\infty}^{+\infty} \phi(u)\, du = 1$ y
- es simétrica: $K(u) = \phi(-u)\, \forall u$

Remark 3. Si K es un núcleo, entonces $K_\lambda(u) = \lambda K(\lambda u)$ también lo es, lo cual permite construir un núcleo adecuadamente escalado a los datos.

Definition 4. (KDE univariado) Sea $(x_1, \ldots, x_N)$ una muestra de elementos i.i.d. tomada de cierta distribución univariada con densidad desconocida f, cuya forma deseamos conocer. Su estimador de densidad por núcleos (su "KDE") es

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \phi_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} \phi\Big(\frac{x - x_i}{h}\Big)$$

Dejando por un momento de lado qué par (K, h) usar, podemos derivar un clasificador "duro" de manera bastante directa para la versión univariada del problema 1:

Definition 5. (clasificador KDE univariado). Sea $C : \mathbb{R}^{d_x} \to [M]$ la "función de clase", tal que $\forall x \in \mathbb{R}^{d_x},\ C(x) = j \iff x \in C_j$. Sean además $\hat{f}_h^{(1)}, \ldots, \hat{f}_h^{(M)}$ los estimadores de densidad obtenidos según 4. El "clasificador por estimación de densidad nuclear" correspondiente será:

$$\hat{C}(x) = \arg\max_{\mathrm{j}\in[\mathrm{M}]} \hat{f}_h^{(j)}(x)$$

asignando cada observación a la clase en la que maximiza la densidad estimada.

Cuando las clases de las cuales se compone la población se encuentran muy "separadas" entre sí (es decir, $\exists k \in [M] : f_h^{(k)}(x_0) \gg 0\ ,\ f_h^{(j)} \simeq 0\ \forall\ j \in [M]/k$), la clasificación "dura" de 5 será suficiente. Ahora bien, ¿cómo hacemos para cuantificar la incertidumbre asociada a la clasificación, cuando existe más de una clase con densidad estimada no despreciable? Como las $\hat{f}_h^{(j)}$ estimadas identifican distribuciones, es razonable decir que $p(C(x) = j) \propto f_h^{(j)}(x)$. Usando la regla de Bayes y un *a priori* sobre las probabilidades de clase basado en las proporciones muestrales $\hat{p}(C_j) = N_j/N$, podemos conseguir una regla *suave* de clasificación:

Definition 6. (clasificador KDE univariado suave) Sea el problema 1 y los estimadores de densidad de 4. Por la regla de bayes,

$$p(C(x) = j) = \frac{f^{(j)}(x) \cdot p(C_j)}{p(x)}$$

Reemplazando el a priori $p(C_j)$ por su estimación muestral, las densidades $f^{(j)}$ por sus estimadores y usando la ley de la probabilidad total para expandir $p(x)$, obtenemos:

$$\hat{p}\left(C\left(x\right)=j\right)=\frac{\hat{f}_h^{(j)}\left(x\right)\cdot N_j}{\sum_{i\in[M]}\hat{f}_h^{(i)}\left(x\right)\cdot N_i}$$

1.3 La noción de distancia en KDE

El peso de cada x_i en $\hat{f}\left(x_0\right)$ es $\phi\left(x_0-x_i\right)$, y como ϕ es simétrica respecto al 0, sólo importa la *distancia* entre el nuevo punto y cada muestra, x_0, x_i; no así la *dirección*. En una dimensión al menos, el núcleo ϕ pondera - escalando por h - la distancia (euclídea) entre el punto a clasificar y cada datum:

$$\phi_h\left(x_0-x_i\right)=\phi_h\left(\left|x_0-x_i\right|\right)=\frac{1}{h}\phi\left(\frac{\left|x_0-x_i\right|}{h}\right)$$

En mayore dimensiones, la situación es más compleja, pero análoga

Definition 7. (KDE multivariado, Hwang 1994) Sea $\{\mathbf{x}\}=\{x_1,\ldots,x_N\}$ una muestra de elementos i.i.d. tomada de cierta distribución $d-$dimensional con densidad desconocida f, cuya forma deseamos conocer. Su estimador de densidad por núcleos (su "KDE") será

$\widehat{f}_h(x)=\frac{1}{Nh^d}\sum_{i=1}^{N}\phi\left(\frac{1}{h}\left(x-x_i\right)\right)$
donde el núcleo ϕ debe satisfacer
$\phi\left(x\right)\geq,0$, y $\int_{\mathbb{R}^{d_x}}(x)\,dx=1$

Un núcleo muy popular es el gaussiano $\phi\left(x\right)=\left(2\pi\right)^{-d/2}\exp\left(-\frac{\|x\|^2}{2}\right)$,
un núcleo simétrico con su valor decayendo suavemente a medida que se aleja del centro.

Usualmente los datos no se encontrarán distribuidos uniformemente en todas las direcciones, y será deseable pre-escalarlos para evitar diferencias extremas en su dispersión y locación. Un enfoque atractivo, es primero "esferar" o "blanquear" los datos mediante una transformación afín que devuelva data con media cero y matriz de covarianza unitaria; y luego aplicar 7. Más específicamente, dada una muesta $\{x\}$,podemos definir su versión "esférica" como

$$z=\mathbf{S}^{-1/2}\left(x-Ex\right)$$

donde la esperanza E es evaluada a través de la media muestral, y $\mathbf{S}\in\mathbb{R}^{d\times d}$es la matriz de covarianza de los datos:

$$\mathbf{S}=E\left[\left(x-Ex\right)\left(x-Ex\right)^T\right]=\mathbf{U}\mathbf{D}\mathbf{U}^T$$
$$\mathbf{S}^{-1/2}=\mathbf{U}\mathbf{D}^{-1/2}\mathbf{U}^T$$

Donde $\mathbf{U}$ es una matriz ortonormal y $\mathbf{D}$ es una matriz diagonal. De preocuparse por la influencia de *outliers* en los datos, existen métodos robustos para su derivación (Huber 1981).

Se puede mostrar fácilment que luego del "blanqueo", $Ez = 0$ y $E\left[zz^T\right] = \mathbf{I_d}$. El estimador resultante para los datos esféricos realiza un estimación de densidad más sofisticada

Acá sigo a Hwang, que hace una presentación algo distinta a Wikipedia/Wand&Jones, primero blanquea y luego usa núcleo con $\mathbf{H} = \mathbf{I_d}$; W&J no recomienda blanquear porque luego usa $\mathbf{H}$ arbitraria. Debería revisar bien la equivalencia, qué conviene más.:

$$\hat{f}(z) = \frac{1}{Nh^d} \sum_{i=1}^{N} \phi\left(\frac{1}{h}(z - z_i)\right)$$

$$\hat{f}(x) = \frac{(\det \mathbf{S})^{-1/2}}{Nh^d} \sum_{i=1}^{N} \phi\left(\frac{1}{h}\mathbf{S}^{-1/2}(x - x_i)\right)$$

Remark 8. Dada una distribución de probabilidad Q en $\mathbb{R}^d$, con media $\mu \in \mathbb{R}^d$ y matriz de covarianza positiva definida $\mathbf{S} \in \mathbb{R}^{d \times d}$, la *distancia de Mahalanobis*[3] de un punto x a Q es

$$d_M(x, Q) = \sqrt{(x - \mu)^T \mathbf{S}^{-1} (x - \mu)}$$

Dados dos puntos x, y en $\mathbb{R}^n$, la distancia de Mahalanobis *entre si* con respecto a Q es

$$d_M(x, Q) = d_M(x, \mu; Q)$$

Como $\mathbf{S}$ es definida positiva, también lo es S^{-1}, con lo que las raíces cuadradas están bien definidas. Por el teorema espectral, S^{-1} se puede descomponer en $S^{-1} = \left(S^{-1/2}\right)^T S^{-1/2}$ para alguna matriz real $d \times d$, lo cual sugiere una definición equivalente

$d_M(x, y; Q) = \left\| S^{-1/2}(x - y) \right\|$

donde $\|\cdot\|$ es la norma euclídea. Es decir, la distancia de Mahalanobis es la distancia euclídea luego de una transformación de blanqueo.

Reemplazando $W = S^{-1/2}, \mu = x_1, \ldots, x_N$, podemos redefinir el estimador de 1.3 para $f(x)$ como un estimador de núcleos basado en la distancia de Mahalanobis de x a cada observación de $\{\mathbf{x}\}$.

La relativa sencillez para el cómputo del método hasta aquí descrito lo hace un perenne favorito entre los estimadores de densidad no paramétricos. Quedará a criterio del investigador considerar si sus bondades vuelven tolerables las limitaciones impuestas o no. A saber,

1. Salvo en casos excepcionalmente bien portados, la dirección y dispersión *local* de la muestra alrededor de un cierto punto x_i típicamente no coincidirá con la dirección $\mathbf{U}$ y dispersión $\mathbf{D}$ *global* que se obtienen de computar $\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ en la muestra completa.

[3] https://en.wikipedia.org/wiki/Mahalanobis_distance

2. Aún cuando la estimación global de $\mathbf{S}$ sea localmente adecuada, no resulta inmediatamente obvio que la suavización $\mathbf{H} = \mathbf{S}$ inducida por la muestra sea óptima en términos de representación de la densidad, super- y sub-suavizando[4] regiones de alta densidad y *outliers*, respectivamente.

3. Al ubicar una "montañita" de densidad en *cada* dato de la muestra, el cómputo del estimador hasta aquí expuesto se vuelve prohibitamente costoso para N relativamente grande.

Wand & Jones (1993) realiza un estudio exhaustivo de las consecuencias de distintas parametrizaciones de $\mathbf{H}$ para el caso multivariado más sencillo, $d = 2$, considerando familias de creciente complejidad para $\mathbf{H}$, siempre positivas definidas:

- en términos generales,
  - productos escalares de la identidad: $\mathcal{H}_1 \coloneqq \left\{h_1^2\mathbf{I}; h_1 > 0\right\}$
  - matrices diagonales con distintas escalas en cada eje: $\mathcal{H}_2 \coloneqq \left(\text{diag}\left(h_1^2, h_2^2\right); h_1, h_2 > 0\right)$
  - matrices completas:

$$\mathcal{H}_3 \coloneqq \left\{ \begin{bmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{bmatrix}; h_1, h_2 > 0, |h_{12}| < h_1 h_2 \right\}$$

    ,

- basadas en una "esferización" de los datos vía matriz de covarianza $\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{bmatrix}$ de la densidad objetivo f:
  - ignorando la correlación $\mathcal{C}_2 \coloneqq \left\{h^2\mathbf{D}; h^2 > 0\right\}$, con $\mathbf{D} = \text{diag}\left(c_{11}, c_{22}\right)$,
  - completa $\mathcal{C}_3 \coloneqq \left\{h^2\mathbf{C}; h^2 > 0\right\}$ e
  - *híbridas*, con suavizado independiente en cada dirección

$$\mathcal{Y} \coloneqq \left\{ \begin{bmatrix} h_1^2 & \rho_{12} h_1 h_2 \\ \rho_{12} h_1 h_2 & h_2^2 \end{bmatrix}; h_1, h_2 > 0 \right\}$$

    y coeficiente de correlación $\rho_{12} = c_{12}/\sqrt{c_{11}c_{22}}$

Nótese que $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3$, $\mathcal{C}_2 \subseteq \mathcal{H}_2$, $\mathcal{C}_3 \subseteq \mathcal{H}_3$, $\mathcal{Y} \subseteq \mathcal{H}_3$. Wand & Jones se "desembarazan" del problema de *selección* de anchos de banda, eligiendo enfocarse en la *eficiencia relativa asintótica*[5] de cada clase con respecto a la más general $\mathcal{H}_3$. A tal fin, toman como medida del error global incurrido por cierto estimador $\hat{f}_{\mathbf{H}}$ el error cuadrático medio integrado (MISE, por sus siglas en inglés)

$$MISE\left(\mathbf{H}\right) = MISE\left(\hat{f}_{\mathbf{H}}, f\right) = E\int_{\mathbb{R}^d}\left(\hat{f}_{\mathbf{H}}\left(y\right) - f\left(y\right)\right)^2 dy$$

[4] oversmoothing y undersmoothing

[5] burdamente, la relación entre los tamaños muestrales necesarios para conseguir el mismo error asintótico restringiendo $\mathbf{H}$ a cierto par de las clases aquí descritas

y su aproximación asintótica.

Los autores notan una dificultad cualitativamente nueva en el caso multivariado en comparación al univariado: definir la *orientación* de $\mathbf{H}$. Aún en el relativamente sencillo contexto bivariado, muestran cómo la estrategia "ingenua" de depender para ello de la covarianza muestral conlleva enormes pérdidas de eficiencia, aún para la familia $\mathcal{Y}$, sobre todo para densidades multimodales y otras que se alejan de la normalidad. En su recomendación final, los autores sugieren que en general "hay mucho para ganar incluyendo parámetros de orientación" (es decir, elementos no-diagonales) en la parametrización de $\mathbf{H}$.

Autores posteriores han tomado el desafío y considerado métodos para la elección de un suavizador $\mathbf{H} \in \mathcal{H}_3$ para el caso general $d-$dimensional. Los mismos autores en un trabajo posterior (WandJones94) proponen un estimador "plug-in" del $\mathbf{H}$ óptimo que se puede aplicar a $\mathcal{H}_3$, pero luego se limitan a la familia diagonal $\mathcal{H}_2$ para su aplicación concreta. Duong2005 sintetiza sus aportes propios y otros precedentes alrededor de la estimación de $\mathbf{H}$ completa según tres métodos de validación cruzada[6]: CV sesgada (BCV), CV insesgada (UCV), y CV "suavizada" (SCV). Todos los métodos propuestos buscan minimizar un error cuadrático (UCV usa MISE; BCV el asintótico AMISE y SCV una combinación lineal de ambos), en el contexto de validación cruzada "dejar-uno-afuera". El método con el que mejores resultados obtienen, SCV, es también el más complejo en su implementación, pues requiere considerar un "suavizador piloto" $\mathbf{G} \in \mathbb{R}^{d\times d}$ cuya elección no es transparente.

Hall2005, por su parte, motivado por la aplicación concreta de estimación de densidad al problema de clasificación, toma un camino distinto para la optimización: en lugar de elegir $\mathbf{H}$ minimizando (A)MISE, se propone elegir $\mathbf{H}$ de manera que minimice una función relacionada directamente con la tarea propuesta: el riesgo de Bayes. Sea $\mathcal{R}$ una regla de decisión como se planteó en 1, diremos que el *riesgo de Bayes* en una región Γ es

$$\begin{aligned}&\mathrm{err}_{\mathcal{R}}(f_1,\ldots,f_K|\gamma)\\ &\qquad=\sum_{j=1}^{K} p_j \int_{\Gamma} Pr\left(x \textbf{ no} \text{ sea clasificado por } \mathcal{R} \text{ como} \in C_j\right) f_j\left(x\right) dx\end{aligned}$$

En este sentido, el clasificador de 5 es óptimo, y por ende es razonable argumentar que elegir $\mathbf{H}$ como Hall propone es superador a optimizar $\mathbf{H}$ para el "resultado intermedio" de estimar las densidades de cada clase. En efecto, para el caso más sencillo $K = 2, d = 1$ y las densidades se "cruzan" en un solo punto con un mismo signo, los anchos de banda encontrados por minimización del riesgo de Bayes son un orden de magnitud distintos de los ya cubiertos. Sin embargo, para $d > 1, K \geq 2$, resulta ser el caso que el ancho de banda óptimo según el error de Bayes es el mismo que via (A)MISE.

Hwang (1994) comienza estudiando explícitamente cómo elegir h para datos esferizados (la familia $\mathcal{C}_3$ en Wand&Jones93), luego nota las dificultades (2)

[6]"cross-validation", o CV, por sus siglas en inglés.

y (3) previamente mencionadas, y compara varios algoritmos superadores en algún sentido al KDE con ancho de banda fijo (FKDE):

- KDE adaptativo (AKDE), similar a FKDE esferizado pero con un factor de ancho local λ_n para cada núcleo

$$\hat{f}_{AKDE}(z) = \frac{1}{Nh^d} \sum_{i=1}^{N} \lambda_i^{-d} \phi \left(\frac{1}{h\lambda_i} (z - z_i) \right)$$

  - El cómputo de los factores λ_i ha de resolverse iterativamente, comenzando por el caso FKDE, $\lambda_i = 1 \forall i \in [N]$, con lo cual el costo computacional será aún más alto que en el caso base.
  - Aunque cada núcleo estará mejor escalado a su contexto local, el enfoque sigue utilizando una misma orientación global para todos los núcleos.

- KDE de base funcional radial (RBF): para minimizar la cantidad de núcleos a ajustar a los datos, divide el proceso de estimación de densidad en dos partes: (i) agrupar los datos en clusters según cierto algoritmo no-supervisado, y luego (ii) ajustar un núcleo gaussiano, su altura y su ancho a cada cluster de (i). Por esto, también se lo conoce com "modelado de mezclas gaussianas".

  - Aunque el estimador final se puede expresar con muy pocos términos, el procedimiento completo es considerablemente más complejo que el de FKDE, dependiendo críticamente de la esferización y remoción de *outliers* para la detección de clusters.
  - Dependiendo del tamaño muestral, la dimensionalidad de los datos y la cantidad de bases utilizadas, ciertos clusters pueden resultar en núcleos demasiado "empinados" o demasiado "planos". Así, una de las principales ventajas de este método - la posibilidad de ajustar una matriz de covarianza distinta a cada cluster de datos - implicará una minuciosa inspección de los datos para saber qué escala y orientación es razonable para cada base.

- KDE por "persecución de la proyección" (PPDE): El espíritu de este método, está basado en buscar iterativamente proyecciones "interesantes" de los datos en bajas dimensiones (típicamente 1-D), modificar la muestra original $\{\mathbf{x}\}^{(0)}$ para remover la estructura encontrada en la proyección, y repetir el proceso en los datos resultantes. Siguiendo a Huber 85, la distribución normal se considera la "menos interesante", y será "más interesante" aquella proyección de los datos que más se le aleje.

  - Para evitar confundir la dirección y escala de la muestra con proyecciones verdaderamente interesante, el método de PPDE requiere también esferizar los datos e ignorar *outliers* juiciosamente (p. 29 Huber85).

– Un problema específico a PPDE, es que no puede lidiar satisfactoriamente con estructuras "escondidas" detrás de otras. E.g., las proyecciones de una densidad 2-D con forma de dona a 1-D no dan cuenta fehaciente de la estructura original.

En resumen:

- Wand, Jones y Duong, entre tantos otros, pretenden buscar selectores basados en MISE para **H** completa, pero terminan encontrando dificultades que los restringen, en la práctica, a matrices diagonales, o los enriedan en la selección de parámetros auxiliares con complejidad propia.
- Hall, motivado por el problema que nos compete, intenta un selector basado en el riesgo de Bayes, pero el método resulta tan complejo en su propio derecho, que aún al tratar muestras multivariadas, lo hace con un suavizador escalar h, $\mathbf{H} \in \mathcal{H}_1$, y nota que los resultados no difieren significativamente de los obtenidos minimizando el error cuadrático integrado.
- Hwang explora todo tipo de anchos de banda escalares (fijos, adaptativos, clusterizados), y por último "abandona" esta línea y considera un método que no requiere definir explícitamente un suavizador **H**: PPDE.

1.4 La maldición de la dimensionalidad

Hasta aquí, pareciera ser que el enfoque de estimación de densidad por núcleos para el caso multivariado está irremediablemente condenado al fracaso, o al menos a una agotadora complejidad. Sin embargo, antes de claudicar, vale la pena entender algunas de las razones de tamaña complejidad.

Una dificultad obvia es que aún considerando un único suavizador global **H**, en d dimensiones hacen falta estimar $\binom{d}{1} + \binom{d}{2} = \left(d^2 + d\right)/2$ varianzas y covarianzas, respectivamente. El crecimiento cuadrático en la cantidad de parámetros implicará que el tamaño muestral N necesario para obtener estimaciones razonables crezca insosteniblemente. El fenómeno, conocido como "maldición de la dimensionalidad", se puede entender intuitivamente considerando el siguiente escenario:

Remark 9. Sea $B\left(x, r, d\right)$ la bola $d-$dimensional de radio r centrada en $x \in \mathbb{R}^d$, y consideremos una v.a. uniformemente distribuida dentro de ella (por volumen), $X \sim Unif\left(B\left(0, r, d\right)\right)$. Sea $\epsilon > 0$; cuál es la probabilidad de que X se encuentre al "interior" de la bola (sustrayendo un "cascarón" externo de espesor ϵ) $Pr(X \in B\left(0, r - \epsilon, d\right))$?

Como la distribución de X es uniforme en volumen, y $B\left(x, r - \epsilon, d\right) \subset B\left(x, r, d\right)$, basta con comparar los volúmenes de de ambas $d-$esferas para encontrar la solución. El volumen$d-$dimensional de una bola es

$$Vol\left(B\left(x, r, d\right)\right) = Vol_B\left(r, d\right) = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2} + 1\right)} r^d$$

donde $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$ es la función gamma. Luego,

$$Pr(X \in B(0, r-\epsilon, d)) = \frac{Vol(B(0, r-\epsilon, d))}{Vol(B(0, r, d))} = \left(\frac{r-\epsilon}{r}\right)^d$$

Como $\left(\frac{r-\epsilon}{r}\right) < 1$, $\lim_{d\to\infty} Pr(X \in B(0, r-\epsilon, d)) \to 0$. Es decir, a medida que crece la dimensión del soporte de X, el "interior" de la bola esta (casi) vacío, y la distribución de X se concentra en el "cascarón" exterior. Aún para valores moderados de d, ϵ el efecto es pronunciado. Por ejemplo, en 20 dimensiones, un cascarón de 2% de espesor ($\epsilon = 0.02r$) concentrará $1 - \left(\frac{r-\epsilon}{r}\right)^d = 1 - 0.98^{20} = 0.6676\cdots \approx$ ¡2/3 de la masa de probabilidad de X!

Este enorme "vacío" en el espacio de alta dimensión, se traduce en una irrelevancia de las métricas "ingenuas" de distancia. Como $x \in B(0, r, d) \iff \|x\| \leq r\sqrt{d}$, y similarmente $x \notin B(0, r-\epsilon, d) \iff \|x\| > (r-\epsilon)\sqrt{d}$, podemos escribir

$$Pr(X \notin B(0, r-\epsilon, d)) = Pr(X \notin B(0, r-\epsilon, d), X \in B(0, r, d))$$
$$1 - \left(\frac{r-\epsilon}{r}\right)^d = Pr\left((r-\epsilon)\sqrt{d} < \|X\| \leq r\sqrt{d}\right)$$

De manera que $\lim_{d\to\infty} Pr\left((r-\epsilon)\sqrt{d} < \|X\| \leq r\sqrt{d}\right) \to 1$. Es decir, a medida que $d \to \infty$ y para ϵ arbitrariamente pequeño, la distancia euclídea de (casi) toda la distribución al centro de la esfera tiende a ser aproximadamente $r\sqrt{d}$, lo cual hace que esta distancia euclídea sea inútil para diferenciar entre elementos de la muestra.

1.5 Reducción de dimensionalidad y la hipótesis de la variedad

A pesar de lo sorprendente del resultado, vale notar que descansa sobre el hecho de que la distribución de X sobre su soporte $\text{supp}(X) = B(0, r, d) \subset \mathbb{R}^d$ es uniforme, e independiente en todas las dimensiones. En casi cualquier contexto material, este supuesto no es sostenible. Por poner un ejemplo, podemos representar todas las posibles imágenes en escala de grises de 1 megapixel como puntos X pertenecientes al espacio $\mathbb{R}^{1024\times1024}$, pero la basta mayoría de ellas consistirían en "puro ruido blanco" y no significarían nada para un observador. Las imágenes que sí tiene sentido reconocer y clasificar (un gato, una bicicleta, etc.) son un conjunto muchísimo más restringido - aún teniendo en cuenta todo tipo de posiciones y contrastes posibles -, y sus diferentes elementos (como la posición de los ojos y las orejas del gato) guardan relaciones específicas entre sí. Es decir, están *correlacionados*.

Si nos suponemos en esta situación, el camino más directo para aliviarla, es *reducir la dimensionalidad* del problema. Al fin y al cabo, es el crecimiento en d lo que nos embrolló en un principio. Dadas $\{\mathbf{x}\} = \{x_i | x_i \in \mathbb{R}^{d_x},\ i \in [N]\}$, buscaremos una *representación* $f : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$, que preserve fielmente los atributos

más relevantes de $x \in \mathbb{R}^{d_x}$, en la menor cantidad de dimensiones d_y. Encontrar compromisos ideales entre la "fidelidad" y la dimensionalidad de estas representaciones, dió lugar al campo de *aprendizaje de representaciones,* del cual Bengio2012[7] hace un excelente censo. El autor relaciona la tarea del área con la noción geométrica de una *variedad.*

Remark 10. A nuestros fines, una variedad $\mathcal{M}$ es un espacio de dimensión $d_{\mathcal{M}}$ que *localmente*, se asemeja a $\mathbb{R}^{d_{\mathcal{M}}}$. En efecto, una variedad puede ser vista como un objeto compuesto de parches $d_{\mathcal{M}}$-dimensionales pegados. Una variedad se llama *cerrada* si no tiene borde y es compacta.

La *hipótesis de la variedad* ("manifold hypothesis") postula que los datos x obtenidos del mundo real con alta dimensionalidad d_x habrían de concentrarse en una variedad $\mathcal{M}$ de -potencialmente - mucha menor dimensionalidad $d_{\mathcal{M}} \ll d_x$, embebido en el espacio original $\mathbb{R}^{d_x}$.

> Esta asunción parece particularmente adecuada en tareas de aprendizaje para las cuales las configuraciones muestreadas aleatoriamente no son como las que ocurren naturalmente: ya mencionamos imágenes, pero esperamos lo mismo de sonidos, texto, secuencias genómicas y hasta aún las respuestas a algunos cuestionarios inverosímilmente exhaustivos de los departamentos estatales de estadística.
>
> Ni bien tenemos una *representación,* uno piensa en una variedad considerando las variaciones en el dominio original que están bien capturadas o reflejadas (por correspondientes cambios) en la representación aprendida. A *grosso modo*, algunas direcciones estarán bien preservadas (las direcciones *localmente tangentes* a cada punto en la variedad), mientras que otras se perderán - las ortogonales a $\mathcal{M}$. Desde esta perspectiva, la principal tarea del aprendizaje no-supervisado, puede ser vista como el modelado de la estructura de la variedad que soporta los datos observados. La representación que se aprenda, puede asociarse a un sistema intrínseco de coordenadas en la variedad embebida. El algoritmo arquetípico de modelado de variedades es, oh sorpresa, también el algoritmo arquetípico de aprendizaje de representaciones de baja dimensionalidad: Análisis de Componenetes Principales (PCA).
>
> PCA modela una *variedad lineal.* Fue inicialmente diseñado con el objetivo de encontrar la variedad lineal más cercana a una nube de puntos. Las componentes principales, i.e., la representación $f_\theta(x)$ que devuelve PCA para un input x, ubica unívocamente su proyección en esa variedad: se corresponde con coordenadas intrínsecas de la variedad. Las variedad que soportan dominios complejos del mundo real, sin embargo, se esperan que sean fuertemente no-lineales.

[7] https://www.reddit.com/r/MachineLearning/comments/mzjshl/comment/gwq8szw/?utm_source=share&utm_medium=wel
Bengio himself sobre el origen del término.

Más que una propiamente dicha hipótesis falsificable al respecto de la distribución de los datos, mencionamos la *hipótesis de la variedad* en tanto resulta modelo mental útil para entender cómo estimar la densidad generadora de los datos en altas dimensiones. Ya mencionamos que a medida que d_x crece, la distancia euclídea en $\mathbb{R}^{d_x}$ se vuelve menos informativa. Trabajar dentro de $\mathcal{M}$, con dimensión $d_{\mathcal{M}}$ puede aliviar la situación sobre todo cuando $d_{\mathcal{M}} \ll d_x$, pero hay una ventaja más escondida en el hecho de que una variedad es sólo localmente semejante al espacio euclídeo - es decir, *lineal* -, pero puede "arrugarse" en el espacio ambiente.

Imaginemos un conjunto de datos $\{\mathbf{u}\} = \left\{u_i, i \in [N], u_i \in \mathcal{U} \subseteq \mathbb{R}^2\right\}$, con forma de letra "U", justamente. $\mathcal{U}$ es una variedad 1-dimensional - una curva - embebida en el espacio cartesiano - $\mathbb{R}^2$, una variedad 2-dimensional., Llamemos u_α, u_ω a los dos puntos que se encuentran más cerca de los extremos superiores del dibujo de la "U". En la variedad latente, estos dos puntos están tan separados entre sí como es posible; sin embargo, si medimos la distancia entre ambos en el espacio ambiente - $\mathbb{R}^2$ - obtendremos que están mucho más cerca entre sí que, por ejemplo, el punto medio donde la "U" corta su eje de simetría axial. La razón de tal insensatez, es simplemente, que hemos tomado una medida de distancia que no se ajusta bien al espacio latente.

1.6 KDE en variedades

¡Excelente! Fieles a la hipótesis de la variedad, podemos sugerir un camino alternativo a los complejos derroteros por los que nos llevó de paseo KDE multivariado en alta dimensión: en lugar de calcular un KDE en el espacio ambiente $\mathbb{R}^{d_x}$, hipotetizamos que $X \in \mathcal{M} \subseteq \mathbb{R}^{d_x}, \dim \mathcal{M} = d_{\mathcal{M}} \ll d_x$, y por lo tanto podemos restringir la definición de su densidad $f : \mathcal{M} \to (0, \infty)$ para obtener una mejor representación. Pero: ¿cómo se construye una función de densidad *en una variedad*? Algunas variedades particularmente interesantes, como el círculo S^1 y la esfera S^2, fueron estudiadas temprano en el siglo XX (Rao, Fisher, citar bien), pero la estimación de densidad en variedades arbitrarias no parece haber sido tratado antes que en Pelletier2005, quien - convenientemente - hizo exactamente eso, "Kernel density estimation on Riemannian Manifolds". En lo que sigue, (intentamos) ser fieles a lo que entendimos de la exposición de Bruno.

Definition 11. (estimación de densidad por núcleos en variedades, Pelletier2005, seccion 2; Muñoz2011 en su tesis de lic. con directores Henry y Rodriguez)

Sea $(\mathcal{M}, g)$ una variedad Riemanniana compacta sin frontera de dimensión d. Asumiremos que $(\mathcal{M}, g)$ es completo, es decir, $(\mathcal{M}, d_g)$ es un espacio métrico completo, donde d_g denota la distancia de Riemann.

Sea X un elemento aleatorio en $\mathcal{M}$[8] con densidad f continua en casi todo punto. Sea $\{\mathbf{X}\}$ un conjunto de N elementos aleatorios i.i.d. a X. Sea $K : \mathbb{R}_+ \to \mathbb{R}$ un mapa no-negativo tal que

[8] i.e., un mapa medible en un espacio de probabilidad $(\Omega, \mathcal{A}, P)$ que toma valores en $(\mathcal{M}, \mathcal{B})$,donde $\mathcal{B}$ representa el σ-campo de Borel de $\mathcal{M}$. Asumiremos que la medida imagen de P por X es absolutamente continua con respecto a la medida Riemanniana de volumen - que notaremos v_g -, admitiendo una densidad f continua en c.t.p. sobre $\mathcal{M}$.

1. $\int_{\mathbb{R}^d} K(\|x\|)\mathrm{d}\lambda(x) = 1$ (K es una función de densidad)
2. $\int_{\mathbb{R}^d} xK(\|x\|)\mathrm{d}\lambda(x) = 0$ ($EX = 0$, K es simétrica),
3. $\int_{\mathbb{R}^d} \|x\|^2 K(\|x\|)\mathrm{d}\lambda(x) < \infty$ ($VarX < \infty$),
4. $\mathrm{sop}K = [0, 1]$,
5. $\sup K(x) = K(0)$,

donde λ es la medida de Lebesgue en R^d. Luego, el mapa $\mathbb{R}^d \ni x \to K(\|x\|) \in \mathbb{R}$ es un núcleo isotrópico en $\mathbb{R}^d$ con soporte en la bola unitaria.

Sean p, q dos puntos de $\mathcal{M}$. Sea $\theta_p(q)$ la *función de densidad volumétrica* en $\mathcal{M}$[9]. Definimos el estimador de densidad de f como el mapa $f_{N,K} : \mathcal{M} \to \mathbb{R}$ que a cada $p \in \mathcal{M}$ le asocia el valor $f_{N,K}(p)$ definido como

$$f_{N,K}(p) = N^{-1} \sum_{i=1}^{N} \frac{1}{h^d} \frac{1}{\theta_{X_i}(p)} K\left(\frac{d_g(p, X_i)}{h}\right)$$

Remark 12. (concordancia con espacios euclídeos) Sea $\mathcal{M} = \mathbb{R}^d$ con su típica métrica euclídea. Luego, $\theta_p(q) = 1 \ \forall \ p, q \in \mathcal{M}$ y $f_{N,K}$ se puede escribir como $f_{N,K} = N^{-1} \sum_{i=1}^{N} r^{-d} K(\|p - X_i\| / r)$. La expresión de $f_{N,K}$ es consistente con la expresión de KDEs en el caso euclídeo.

Para asegurarse de que $f_{N,K}$ sea integrable sobre $\mathcal{M}$, habremos de imponer una restricción más sobre el ancho de banda:

$$h_n < h_0 < \mathrm{inj}_g \mathcal{M}$$

, donde $\mathrm{inj}_g \mathcal{M}$ es el *radio de inyectividad* de $\mathcal{M}$ (Chavel1993, p. 108; Muñoz2011, p. 23)[10]. Sin entrar en demasiados detalles, siempre y cuando $\mathcal{M}$ sea compacta

[9] Besse 1978 (p. 154) lo define aproximadamente como

$$\theta_p : q \to \theta_p(q) = \frac{\mu_{\exp_p^* g}}{\mu_{g_p}} \left(\exp_p^{-1}(q)\right)$$

i.e., el cociente entre la medida canónica de la métrica Riemanniana $\exp_p^* g$ en espacio tangente $T_p(M)$, y la medida de Lebesgue en la estructura euclídea $g_p$en $T_p(M)$. La función de densidad volumétrica está ciertamente definida para q en un vecindario de p. En términos de coordenadas normales geod´sicas en p, $\theta_p(q)$ es igual al determinante de la métrica g expresado dichas estas coordenadas en $\exp_P^{-1}(q)$.

[10]

Definition 13. (Muñoz2011, p. 23, definición 3.3.16) Sea $(\mathcal{M}, g)$ una variedad Riemanniana de dimensión d. Llamamos *radio de inyectividad* a

$$\mathrm{inj}_g \mathcal{M} = \inf_{p \in \mathcal{M}} \sup \{s \in \mathbb{R} > 0 : B(p, s) \text{ es una bola normal}\}$$

Burdamente, diremos que B es una *bola normal* centrada en p si existe una bola V en $T_p(\mathcal{M})$ (un vecindario de p) en el que las coordenadas de cada punto $q \in V$ se pueden mapear biyectivamente a coordenadas en $\mathbb{R}^d$: por ejemplo, si $\mathcal{M}_1 = \mathbb{R}^d$ con la métrica canónica ($\|\cdot\|$) entonces $\mathrm{inj}_g \mathcal{M}_1 = \infty$, pues todo el espacio comparte un único mapa de coordenadas global. Si le quitamos un punto, $\mathcal{M}_2 = \mathcal{M}_1 - \{p\}$ entonces $\mathrm{inj}_g \mathcal{M}_2 = 0$ (para un punto "muy cercano a p", $q \in \mathcal{M}, q \approx p$, no habrá bola normal posible) . Si $\mathcal{M} = S^1 \times \mathbb{R}$ (un cilindro vacío en $\mathbb{R}^3$) con la métrica inducida de $\mathbb{R}^3$, el radio de inyectividad es π. REPASAR.

este radio de inyectividad será > 0, y al menos para los resultados asintóticos (cuando el tamaño muestral es lo suficientemente grande como para que $h \to 0$), $0 < h < h_0$.

Pelletier2005 avanza algunas propiedades elementales de este estimador: adapta el concepto de "media" para elementos aleatorios en $\mathbb{R}^d$ a e.a. en variedades Riemannianas compactas sin frontera $\mathcal{M}$ (Proposición II, "media intrínseca"), y prueba que $f_{N,K}$ es un estimador consistente [11] de f (Teorema 5), en el siguiente sentido

Theorem 14. *(Teorema 5, Pelletier 2005)* Sea f una densidad de probabilidad dos veces diferenciable en $\mathcal{M}$ con segunda derivada covariante acotada. Sea $f_{N,K}$ su estimador definido en 11 con ancho de banda h que satisface la condición 1.6. Luego, existe una constante C_f tal que

$$E_f \left\| f_{N,K} - f \right\|^2_{L^2(\mathcal{M})} \leq C_f \left(\frac{1}{Nh^d} + h^4 \right)$$

En consecuencia, para $h \sim N^{\frac{-1}{d+4}}$,

$$E_f \left\| f_{N,K} - f \right\|^2_{L^2(\mathcal{M})} \leq O\left(N^{\frac{-4}{d+4}} \right)$$

Henry & Rodríguez2009 continúan el estudio de este estimador, probando

1. bajo ciertas condiciones de regularidad sobre conjuntos compactos $\mathcal{M}_0 \subseteq \mathcal{M}$ - la consistencia fuerte

$$\sup_{p \in \mathcal{M}_0} \left| f_{n,K}(p) - f(p) \right| \xrightarrow{c.t.p.} 0$$

2. bajo condiciones extras sobre f y la serie h_n, $f - f_{N,K}$ converge en distribución a ciert ley normal, con tasa $\sqrt{nh^d}$

$$\sqrt{nh^d} \left(f(p) - f_{n,K}(p) \right) \xrightarrow{\mathcal{D}} \mathcal{N}(\mu, \Sigma)$$

Loubes y Pelletier (2010) extienden el trabajo de Pelletier 2005, proponiendo un clasificador binario basado en núcleos, para e.a. soportados sobre variedades compactas y cerradas de Riemann. Recordemos que en 6 propusimos un clasificador suave que asignase a cada clase, una probabilidad de pertenencia

$$p\left(C(x) = j\right) = \frac{f^{(j)}(x) \cdot p(C_j)}{p(x)}$$

[11] Pelletier considera la convergencia en $L^2(\mathcal{M})$, ¿esto sería consistencia débil? ¿Qué diferencia hay con la que estudian Henry&Rodríguez2009?

de manera que podemos describir una reglas de clasificación dura, como

$$\begin{aligned}\mathcal{R}\left(x|f_1,\ldots,f_K\right) &= \arg\max_{j\in[K]} p\left(C\left(x\right)=j\right)\\ &= \arg\max_{j\in[K]} \frac{f^{(j)}\left(x\right)\cdot p(C_j)}{\sum_{j\in[K]} f^{(j)}\left(x\right)\cdot p(C_j)}\\ &\arg\max_{j\in[K]} f^{(j)}\left(x\right)\cdot p(C_j)\end{aligned}$$

y su estimador muestral

$$\begin{aligned}\hat{\mathcal{R}}\left(x|\hat{f}_1,\ldots,\hat{f}_K\right) &= \arg\max_{j\in[K]} \hat{f}^{(j)}\left(x\right)\cdot\hat{p}\left(C_j\right)\\ &= \arg\max_{j\in[K]} N_j^{-1}\sum_{i=1}^{N}\frac{1}{h^d}\frac{1}{\theta_{X_i}\left(p\right)}K\left(\frac{d_g\left(p,X_i\right)}{h}\right)\cdot\frac{N_j}{N}\\ &= \arg\max_{j\in[K]} \sum_{i=1}^{N}\mathbf{1}\left\{C\left(X_i\right)=j\right\}K_h\left(p,X_i\right)\end{aligned}$$

donde

$$K_h\left(p,X_i\right)=\frac{1}{h^d}\frac{1}{\theta_{X_i}\left(p\right)}K\left(\frac{d_g\left(p,X_i\right)}{h}\right)$$

Este es, precisamente, el clasificador que Loubes y Pelletier proponen. Considerando como función objetivo a minimizar la misma probabilidad de error de clasificación que vimos con Hall2005,

$$L\left(\mathcal{R}\right)=Pr\left(\mathcal{R}\left(X\right)=C\left(X\right)\right)$$

los autores muestran que el clasificador propuesto $\hat{\mathcal{R}}$ alcanza asintóticamente la misma pérdida que el clasificador óptimo de bayes, $\mathcal{R}^*$

$$\lim_{n\to\infty} Pr\left(L\left(\hat{\mathcal{R}}_n\right)=L\left(\mathcal{R}^*\right)\right)=1$$

con

$$\mathcal{R}^*\left(x\right)=\arg\max_{j\in[K]} Pr\left(C\left(X\right)=j|X=x\right)$$

Siguiendo a Drevoye1996, diremos que el clasificador es *fuertemente consistente*, en tanto alcanza el error de Bayes cuando $n\to\infty$. Los autores dejan las consideraciones prácticas de su funcionamiento fuera del trabajo.

Los resultados combinados de Pelletier, Henry, Rodríguez y Loubes nos dejan bastante cerca de lo que venimos buscando - construir un clasificador basado en densidades -, con una diferencia fatal: estos trabajos consideran variedades *conocidas*, mientras que nosotros trabajamos bajo la *hipótesis de la variedad*, pero en principio no conocemos la variedad en sí. Crucialmente, desconocer la variedad implica desconocer dos cosas: la distancia geodésica d_g y la función de densidad de volumen θ_p, que tendremos que *aprender de los datos* de alguna manera, junto con la densidad f.

1.6.1 Isomap

Consideraremos brevemente uno de los algoritmos más famosos de aprendizaje de variedades no-lineales: Isomap[12]. Según Tenenbaum2000, el objetivo explícito de su algoritmo es "aprender la geometría global subyacente de un dataset, usando información métrica local fácilmente medible", para un conjunto amplio de variedades no-lineales. La tarea central, consiste en aproximar adecuadamente las distancias geodésicas en la variedad $d_g(x, y)$ entre puntos alejados, conociendo únicamente las distancias euclídeas en la muestra $\|x - y\|$.

Dada la naturaleza localmente euclídea de las variedades, para puntos "vecinos" entre sí, la distancia en $\mathbb{R}^{d_x}$ (en el espacio de las X) será una aproximación razonable a la distancia geodésica en la variedad. Luego, para puntos alejados entre sí, podemos aproximar d_g como la suma de una secuencia de "pequeños saltos" entre puntos vecinos en el grafo de la muestra.

El algoritmo completo, consta de tres pasos principales (Tabla 1, Tenenbaum2000):

1. **Constrúyase un grafo de vecinos muestrales** $G = (X, E)$ sobre el dataset completo, donde la arista $x \leftrightarrow y$ está incluida si $\|x - y\|_{d_x} < \epsilon$ ("ϵ-Isomap"), o si y es uno de los K vecinos más cercanos de x . Tómese $\|x - y\|$ como el valor de la arista $x \leftrightarrow y$.

2. **Compútense los caminos mínimos**, usando - según convenga - el algoritmo de Floyd-Warshall o Dijsktra en el grafo G. Los costos de los caminos mínimos $d_G(x, y)$ constituyen una aproximación de las distnacias geodésicas $d_{\mathcal{M}}(x, y)$.

3. **Constrúyase un *embedding* d−dimensional.** Utilizando escalamiento multidimensional (MDS, MultiDimensional Scaling, un algoritmo de reducción de dimensionalidad), crear una representación (embedding) en el espacio euclídeo $\mathbb{R}^d$ que minimice una métrica de discrepancia entre las distancias computadas en (2), con las distancias en la representación a construir, llamada "stress" o "strain".

Los resultados de este algoritmo - que han sido bastante espectaculares para lo relativamente sencillo de su estructura, descansan en una prueba de la convergencia asintótica, a medida que N crece, de que las distancias en el grafo d_G proveen aproximaciones incrementalmente mejores a las distancias geodésicas intrínsecas $d_{\mathcal{M}}$, volviéndose arbitrariamente precisas en el límite de $N \to \infty$. La tasa a la que esta convergencia sucede, depende de ciertos parámetros de la variedad (su dimensión $d_{\mathcal{M}}$, la función de volumen θ_p), de cómo esta yace en el espacio ambiente (radio de curvatura r_0 y separación de ramass_0) y de la densidad $f : \mathcal{M} \to \mathbb{R} > 0$ de la que estamos sampleando.

Allende los costos computacionales, hay dos parámetros a fijar en este algoritmo: el parámetro de "vecindad" ϵ ó K en (1), y la dimensión d en (3). Inspeccionando el gráfico de "estrés" de MDS como función de la dimensión d

[12] **Iso**metric feature **map**ping, en inglés

escogida, se pueden buscar un punto de inflexión ("codo") en que seguir aumentando d no aliviana significativamente la tensión del algoritmo, y son por tanto candidatos naturales a la dimensión intrínseca de la variedad $d_{\mathcal{M}}$.

Por su parte, el valor óptimo de ϵ ó K no cuenta con una regla inmediata para su determinación. Consideremos $\epsilon-$Isomap: valores demasiado pequeños de ϵ podrían dejar muchos vértices de G - muchas observaciones muestrales - desconectadas de la componente gigante - la componente conexa de mayor tamaño - de G;valores demasiado grandes de ϵ podrían llevar a incluir en G aristas $e \in E$ que cruzan el espacio ambiente $\mathbb{R}^{d_x}$ completamente por fuera de $\mathcal{M}$, "cortocircuitando" la representación. Consideraciones análogas complican la elección de cantidad de vecinos en $K-$Isomap.

1.7 Distancia de Fermat

La idea central en Isomap, es *primero* aprender/computar una distancia (en el grafo de vecinos más cercanos) y *luego* construir un *embedding* - una representación - en cierta dimensión dada, en lugar de usar una distancia dada, para aprender una representación de menor dimensión. Sin embargo, por cómo está definido el algoritmo, lo que Isomap aproxima es la distancia euclídea en la dimensión intrínseca de la variedad subyacente. A nuestros fines, será necesario también considerar la *densidad f en la variedad $\mathcal{M}$.*

Por ejemplo, si f fuese una mezcla con iguales pesos de dos leyes gaussianas unidimensionales, $\mathcal{N}(0,1)$ y $\mathcal{N}(10,2)$, quisiéramos que el punto $x = 5$ - euclídeamente equidistante de ambas - estuviese más cerca de $\mathcal{N}(10,2)$ que de $\mathcal{N}(0,1)$ - exactamente como sucedería en $\mathbb{R}^d$ si consideramos la distancia de Mahalanobis 8.

En casos reales, ni f ni $\mathcal{M}$ se conocerán de antemano, así que pareciera conveniente tratar de aprender una distancia que considere ambas a la vez. Es exactamente en ese sentido que Groisman et al. (2019) proponen la "distancia de Fermat", una distancia basada en densidades aplicable a variedades que, de alguna manera, generaliza el trabajo de Tenenbaum2000.

Definition 15. (distancia de Fermat muestral, adaptado de Definición 2.1 en Groisman2019)

Sea Q un conjunto no-vacío, localmente finito, contenido en $\mathbb{R}^d$. Para $\alpha \geq 1$ y $s, t \in Q$, definimos la *distancia de Fermat muestral* como

$$D_{Q,\alpha}(s,t) = \inf \left\{ \sum_{j=1}^{K-1} \left\| x_{i_j+1} - x_{i_j} \right\|^{\alpha} : (q_1, \ldots, q_K) \text{ es un camino de } s \text{ a } t, q_i \in Q \forall i \in [K], K \geq 1 \right\}$$

Los autores definen esta distancia muestral para conjuntos arbitrarios Q, pero de aquí en más consideraremos $Q = \{\mathbf{x}\}$, la muestra d_x-dimensional de interés. Nótese que $D_{Q,\alpha}$ satisface la desigualdad triangular, y define una métrica sobre Q. Cuando no sea estrictamente necesario, omitiremos en la notación la dependencia en Q, α.

Remark 16. La distancia de Fermat muestral es el costo del camino mínimo en el grafo *completo* de Q, con las aristas pesadas por una potencia α de la distancia euclídea. Cuando $Q = \{\mathbf{x}\}$, $K = N$, $\alpha = 1$, la distancia del paso (1) en Isomap es idéntica a la distancia muestral de Fermat.

A continuación, definen la versión macroscópica de la distancia de Fermat muestral,

Definition 17. (distancia de Fermat, definicion 2.2) Sea $\mathcal{M}$ una variedad de Riemann, $f : \mathcal{M} \to \mathbb{R}_+$ una función continua y positiva en $\mathcal{M}$, $\beta \geq 0$ y $s, t \in \mathcal{M}$. Definimos la *distancia de Fermat* $\mathcal{D}_{f,\beta}(s,t)$ como

$$\mathcal{T}_{f,\beta}(\gamma) = \int_\gamma f^{-\beta}, \quad \mathcal{D}_{f,\beta}(s,t) = \inf_{\gamma \in \Gamma} \mathcal{T}_{f,\beta}(\gamma)$$

donde el ínfimo esta tomado sobre el conjunto de todos los caminos continuos y rectificables contenidos en $\bar{\mathcal{M}}$ (la clausura de $\mathcal{M}$) que comienzan en s y terminan en t, y la integral es respecto de la longitud de arco dada por la distancia euclídea.

Se omitirán las dependencias de f, β cuando no haya confusión posible.

Uniendo las dos definiciones previas, el teorema central del trabajo es el siguiente:

Theorem 18. (Teorema 2.7, Groisman2019) Sea $\mathcal{M}$ una variedad d-dimensional, isométrica y C^1 embebida en $\mathbb{R}^D$[13]. Sea $Q_n = \{q_1, \ldots, q_n\}$ puntos independientes con densidad común f. Luego, para $\alpha > 1$ y $x, y \in \mathcal{M}$ se tiene

$$\lim_{n\to\infty} n^\beta D_{Q_n,\alpha}(x,y) = \mu \mathcal{D}_{f,\beta}(x,y) \text{ casi seguramente.}$$

Aquí, $\beta = (\alpha - 1)/d$ y μ es una constante que depende únicamente de α y la dimensión de la variedad d.

En otras palabras, correctamente escalada, la distancia muestral de Fermat converge a la distancia "poblacional" de Fermat, y $D_{Q_n,\alpha}$ es un estimador consistente de $\mathcal{D}_{f,\beta}$. Los autores prueban el caso en que f corresponde a un proceso puntual de Poisson homogéneo en $\mathcal{M}$, y conjeturan que es cierto para f arbitraria.

2 Propuesta

En la Introducción hemos repasado en detalle un método eficiente y lo sumamente estudiado para responder al problema de clasificación en dominios de alta dimensionalidad: la estimación de densidad por núcleos (KDE), específicamente en variedades de Riemann. Notamos que de los tres parámetros a elegir -

[13] Es decir, existe un conjunto abierto y conexo $S \subset \mathbb{R}^d$ y $\phi : \bar{S} \to \mathbb{R}^D$ una transformación isométrica tal que $\backslash \text{phi}\phi(\bar{S}) = \mathcal{M}$. En aplicaciones reales se espera que $d \ll D$, pero no es necesario.

el núcleo, el ancho de banda y la distancia - tanto el ancho de banda como la distancia son problemáticos en HD, aunque para el ancho de banda el tratamiento encontrado en la literatura es mucho más extenso. Nos proponemos investigar si es posible mejorar la performance de los métodos descritos hasta ahora, con una noción de distancia aprendida de los datos, la distancia muestral de Fermat propuesta por Groisman2019. Más específicamente, aplicaremos

- un clasificador basado en estimaciones de densidad por núcleos (gaussianos) en variedades según Loubes2010
- con matriz de suavización **H** individualmente orientada en cada elemento muestral según Vincent2003
- y distancia varietal aprendida según Groisman2019

Evaluaremos al clasificador resultante en un conjunto de datasets sintéticos y naturales que representen un espectro amplio de casos de alta dimensionalidad, a través de un estudio de ablación, para entender cuál es la ventaja marginal de utilizar una distancia aprendida por sobre el clasificador equivalente con distancia euclídea.

Los métodos de estimación por núcleos, aunque simples en su concepción, tienen altos requerimientos computacionales, y el aprendizaje de distancias basadas en grafos, más aún. Por ello, en el estudio ablativo comparado, incluiremos como referencia de precisión:

- un clasificador KNN con distancia euclídea - la versión más sencilla posible de un clasificador KDE, y
- un clasificador por GBT - gradient boosting trees -, uno de los métodos más "plug & play" disponibles hoy en día.

Incluiremos algunos comentarios sobre el costo computacional de cada método, comparando la expectativa teórica con los resultados de nuestras - sencillas y caseras - implementaciones.

Finalmente, nos proponemos dar algunas garantías teóricas sobre el comportamiento asintótico de la distancia muestral de Fermat como estimador de la distancia (macroscópica / poblacional) homónima.

3 Otros papers

Hay varios papers con ideas muy piolas sobre como aprender una variedad, y como usar la info (las cartas generadas) para clasificar. Se aleja de nuestro interes principal, pero tal vez ameriten mención?

3.0.1 Manifold Tangent Classifier (+TangentProp)

Incluye un buen detalle de 3 versiones interreleacionadas de la hipotesis de la variedad.

Usa una NN para encontrar en cada punto, direcciones tangentes en las que la funcion de activacion no cambia significativamente. Luego, usa tangentprop (una forma de gradient backpropagatiojn con restricciones sobre las derivadas primeras) para incluir esa info en la optimizacion y mejorar los resultados de clasificacion.

3.0.2 Shell Theory

Por la maldicion de la dimensionalidad, debería ser directamente imposible ML en alta dimension, pero en la practica se ve que funciona. Propone una teoría general pero accesible de "generadores jerárquicos", "shell theory", que imita las clasificaciones jerárquicas semánticas que buscamos entender (gato siamés $\subset$ gato $\subset$ animal).

3.0.3 Brand2003 - Charting a Manifold

Menciona una especie de hipotesis de la variedad antes que otros cuantos, aunque no la llama así.

Ofrece una idea empírica de cómo estimar la dimensión de la variedad mirando cómo crece la función de conteo de puntos $c(r, y) = \sum_{i=1}^{N} \mathbf{1}\{x_i \in B(y, r, d_x)\}$ en relación al radio de la bola considerada.

Aplica el método propuesto al trefoil que consideramos recientemente.

3.0.4 Vincent, Bengio 2003 - Manifold Parzen Windows

Esencialmente MVKDE con $\mathbf{H}_i$definida para *cada* obserbvación en un vecindario "suave" (usando kernels sobre la distancia a c/otra obs) o duro (KNN) alrededor de la obs. Usa algunos "trucos" para evitar $\mathbf{H}_i$ mal condicionadas.

Considera "negative conditional log likhelihood" como medida de bondad del clasificador, como alernativa continua al error de clasificación y otras.

3.0.5 The Curse of Highly Variable Functions for Local Kernel Machines

Muestra cómo todos los métodos basados en núcleos (KNN,m KDE, hasta isomap) comparten la necesidad de un tamaño muestral enorme cuando la función objetivo a aprender tiene muchas variaciones, por depender de entornos locales a cada observacion para mapear la variedad. Aún funciones de baja "complejidad de Kolmogorov" (paridad, seno) son muy difíciles de aprender con kernels, y sin info global.

3.0.6 Learning Eigenfunctions Links Spectral Embedding and Kernel PCA

Une un monton monton de metodos de estimacion de densidad / embeddings demtro de un marco unificado de funciones basadas en nucleos. En particular,

Isomap (y landmark-Isomap) se pueden ampliar a puntos out-of-sample computando la aproximacion a la distancia geodésica en el grafo de kNN, a traves de los puntos de entrenamiento, basicamente como estamos por proponer nosotros para extender distancia de fermat a out-of-sample. Duro pero interesante.

3.0.7 Chu2018 - Exploration of a Graph-based Density-Sensitive Metric

We consider a simple graph-based metric on points in Euclidean space known as the edge-squared metric. This metric is defined by squaring the Euclidean distance between points, and taking the shortest paths on the resulting graph. This metric has been studied before in wireless networks and machine learning, and has the density- sensitive property: distances between two points in the same cluster are short, even if their Euclidean distance is long. This property is desirable in machine learning.

3.0.8 Biijral2012 - Semi-supervised Learning with Density Based Distances

Denoting the probability density function in R d by f (x), we can define a path length measure through R d that assigns short lengths to paths through highly den- sity regions and longer lengths to paths through low density regions. We can express such a path length measure as

$$J_f\left(x_1 \overset{\gamma}{\rightsquigarrow} x_2\right) = \int_0^1 g(f(\gamma(t)))\left\|\gamma'(t)\right\|_p dt,$$

where $\gamma : [0,1] \to \mathbb{R}$ d is a continuous path from $\gamma(0) = x_1$ $\gamma(1) = x_2$ and $g : \mathbb{R}^+ \to \mathbb{R}$ is monotonically decreasing (e.g. $g(u) = 1/u$). Using Equation 1 as a density-based measure of path length, we can now define the density based distance (DBD) between any two points $x_1, x_2 \in \mathbb{R}$ d as the density-based length of a shortest path between the two points

$$D_f(x_1, x_2) = \inf_{\gamma} J_f(x_1 \overset{\gamma}{\rightsquigarrow} x_2)$$

Alternatively, a simple heuristic was suggested by Vincent and Bengio (2003) in the context of clustering, and is based on constructing a weighted graph over the data set, with weights equal to the squared dis- tances between the endpoints and calculating shortest paths on this graph.

N.delA.: El paper de Vincent y Bengio que mencionan no está disponible en internet, sólo aparece citado en otros trabajos: "*Vincent, P., & Bengio, Y. (2003). Density sensitive metrics and kernels. Proceedings of the Snowbird Workshop.*", pero todo indica que la formulación es como la de Groisman2019, con $\beta = 2$

Más adelante, considera funciones $g = f^{-r}$ y pareciera llegar a una formulación idéntica a la de Groisman2019.

4 Notas sueltas

- soft clf chen
- (¿Es lo mismo $\|\cdot\|$ que la geodésica en R^d_x? Creo que sí)
- mencion a t-SNE? como esta basada en distancia euclidea, no parece que vaya a ayudar mucho
- RKHS - reproducing kernel hilbert spaces -: alguito para entender a que cuernos ser refieren?
- biblio: No subirla, pero esconder script ligeramente disimulado que la baje por uno?

5 Análisis experimental

6 Cuentita

7 Conclusiones

References

[1] Yoshua Bengio. The consciousness prior.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives.

[3] Yoshua Bengio, Olivier Delalleau, and Nicolas Roux. The curse of highly variable functions for local kernel machines. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

[4] Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux, Jean-François Paiement, Pascal Vincent, and Marie Ouimet. Learning eigenfunctions links spectral embedding and kernel PCA. 16(10):2197–2219.

[5] Yoshua Bengio, Hugo Larochelle, and Pascal Vincent. Non-local manifold parzen windows. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

[6] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In Catriel Beeri and Peter Buneman, editors, *Database Theory - ICDT'99*, Lecture Notes in Computer Science, pages 217–235. Springer.

[7] Avleen S. Bijral, Nathan Ratliff, and Nathan Srebro. Semi-supervised learning with density based distances.

[8] Matthew Brand. Charting a manifold. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

[9] Jeff Calder and Mahmood Ettehad. Hamilton-jacobi equations on graphs with applications to semi-supervised learning and data depth. 23(318):1–62.

[10] A. Carpio, L. L. Bonilla, J. C. Mathews, and A. R. Tannenbaum. Fingerprints of cancer by persistent homology.

[11] Lawrence Cayton. Algorithms for manifold learning.

[12] José E. Chacón and Tarn Duong. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. 7.

[13] Isaac Chavel. *Riemannian geometry: a modern introduction*. Number 98 in Cambridge studies in advanced mathematics. Cambridge University Press, 2nd ed edition. OCLC: ocm62089870.

[14] Yen-Chi Chen, Christopher R. Genovese, and Larry Wasserman. A comprehensive approach to mode clustering.

[15] Timothy Chu, Gary Miller, and Donald Sheehy. Exact computation of a manifold metric, via lipschitz embeddings and shortest paths on a graph.

[16] M. Davenport, J. Romberg, and J Rozell. Bayes rule for random variables. ECE 3077 Notes by M. Davenport, J. Romberg and C. Rozell. Last updated 21:27, June 25, 2014.

[17] Tarn Duong and Martin L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. 32(3):485–506.

[18] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Elsevier. Google-Books-ID: BIJZTGjTxBgC.

[19] Pablo Groisman, Matthieu Jonckheere, and Facundo Sapienza. Nonhomogeneous euclidean first-passage percolation and distance learning.

[20] Peter Hall and Kee-Hoon Kang. Bandwidth choice for nonparametric classification. 33(1).

[21] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer London, Limited.

[22] Guillermo Henry and Daniela Rodriguez. Kernel density estimation on riemannian manifolds: Asymptotic results. 34(3):235–239.

[23] Peter J. Huber. Projection pursuit. 13(2):435–475.

[24] Jenq-Neng Hwang, Shyh-Rong Lay, and A. Lippman. Nonparametric multivariate density estimation: a comparative study. 42(10):2795–2810.

[25] Wen-Yan Lin, Siying Liu, Changhao Ren, Ngai-Man Cheung, Hongdong Li, and Yasuyuki Matsushita. Shell theory: A statistical model of reality. pages 1–1.

[26] Anna Little, Daniel McKenzie, and James Murphy. Balancing geometry and density: Path distances on high-dimensional data.

[27] Jean-Michel Loubes and Bruno Pelletier. A kernel-based classifier on a riemannian manifold. 26(1):35–51.

[28] Daniel Mckenzie and Steven Damelin. Power weighted shortest paths for clustering euclidean data.

[29] Andres Leandro Muñoz. Estimación no paramétrica de la densidad en variedades riemannianas.

[30] Emanuel Parzen. On estimation of a probability density function and mode. 33(3):1065–1076. Publisher: JSTOR.

[31] Bruno Pelletier. Kernel density estimation on riemannian manifolds. 73(3):297–304.

[32] Salah Rifai, Yann N Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

[33] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. 27(3):832–837. Publisher: Institute of Mathematical Statistics.

[34] Facundo Sapienza, Matthieu Jonckheere, and Pablo Groisman. Weighted geodesic distance following fermat's principle.

[35] B. W. Silverman. Using kernel density estimates to investigate multimodality. 43(1):97–99.

[36] Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop - a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann.

[37] Yichuan Tang. Tutorial on tangent propagation.

[38] Charles Taylor. Classification and kernel density estimation. 41(3):411–417.

[39] Joshua Tenenbaum. Mapping a manifold of perceptual observations. In *Advances in Neural Information Processing Systems*, volume 10. MIT Press.

[40] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. 290(5500):2319–2323.

[41] Loring W. Tu. *An Introduction to Manifolds*. Universitext. Springer New York.

[42] Pascal Vincent and Yoshua Bengio. Manifold parzen windows. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

[43] M. P. Wand and M. C. Jones. Comparison of smoothing parameterizations in bivariate kernel density estimation. 88(422):520–528.

[44] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Springer US.

[45] Matt P. Wand and M. Chris Jones. Multivariate plug-in bandwidth selection. 9(2):97–116. Publisher: Heidelberg: Physica-Verlag,[1992-.

[46] Zhipeng Wang and David W. Scott. Nonparametric density estimation for high-dimensional data - algorithms and applications. 11(4).

[47] Jaehong Yu and Seoung Bum Kim. Density-based geodesic distance for identifying the noisy and nonlinear clusters. 360:231–243.