Distancia de Fermat en Clasificadores de Densidad por Núcleos

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

Resumen

TODO

Índice

Parte I
Forma Final

Notación

$\mathbb{R}$: los números reales

1. Preliminares

1.1. El problema de clasificación

1.1.1. Definición del problema unidimensional

Definición 1 (muestra aleatoria)**.** Sea

Consideremos el problema de clasificación:

Definición 2 (problema de clasificación)**.** Sea $\{\mathbf{X}\} = \{X_1, \ldots, X_N\}$, $X_i \in \mathbb{R}^{d_x}$ $\forall i \in [N]$ una muestra de N observaciones aleatorias d_x-dimensionales, repartidas en M clases $C_1, \ldots, C_M$ mutuamente excluyentes y conjuntamente exhaustivas[1]. Asumamos además que la muestra está compuesta de observaciones independientes entre sí, y las observaciones de cada clase están idénticamente distribuidas según 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 cuy

1.1.2. Clasificadores «duros» y «suaves»

Definición 3. a clase es desconocida,

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

Cualquier 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 relativamente 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 - la más frecuente - entre las k observaciones de entrenamiento más cercanas en distancia euclídea $\|x_0 - \cdot\|$. $k-$NN no menciona explícitamente las leyes de clase $\mathcal{L}_j$, lo cual lo mantiene sencillo a costa de ignorar la estructura del problema.

[1] es decir, $\forall\, i \in [N] \equiv \{1, \ldots, N\}, X_i \in C_j \iff X_i \notin C_k, k \in [M], k \neq j$

Uno sencillo: 1-NN - tal vez?

Uno conocido: LR - tal vez?

3.1.3. Datasets

Datasets sintéticos baja dimensión

Datasets reales en «mediana» dimensión

Dígitos

PCA-red MNIST

Consideremos el problema de clasificación:

Definición 4 (problema de clasificación)**.** Sea $\{\mathbf{X}\} = \{X_1, \ldots, X_N\}$, $X_i \in \mathbb{R}^{d_x}$ $\forall i \in [N]$ una muestra de N observaciones aleatorias d_x-dimensionales, repartidas en M clases $C_1, \ldots, C_M$ mutuamente excluyentes y conjuntamente exhaustivas[2]. Asumamos además que la muestra está compuesta de observaciones independientes entre sí, y las observaciones de cada clase están idéntica-

[2] es decir, $\forall\, i \in [N] \equiv \{1, \ldots, N\}, X_i \in C_j \iff X_i \notin C_k, k \in [M], k \neq j$

mente distribuidas según 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 $C_j, j \in [M]$?

Cualquier 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 relativamente 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 - la más frecuente - entre las k observaciones de entrenamiento más cercanas en distancia euclídea $\|x_0 - \cdot\|$. $k-$NN no menciona explícitamente las leyes de clase $\mathcal{L}_j$, lo cual lo mantiene sencillo a costa de ignorar la estructura del problema.

8.2. Clasificadores 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 $\{\mathbf{X}\}^{(j)} = \left\{X_1^{(j)}, \ldots, X_{N_j}^{(j)}\right\}$ y algún procedimiento estándar [3]

clasificador Definir una regla de decisión - un *clasificador* - $\mathcal{R}\left(x|\hat{\mathcal{L}}_j, j \in [M]\right) : x \in S \to [M] \ni j$ que dados los estimadores de (2), asigne la observación x_0 a la clase $\mathcal{R}(x_0)$.

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

Ejemplo 5. El análisis de discriminante lineal (LDA) de Fisher[**?**] para clasificación binaria ($j \in \{0, 1\}$) se encuadra en esta familia de la siguiene manera:

1. Las las leyes $\mathcal{L}_j$
   a) son todas distribuciones normales con media μ_j y
   b) homocedásticas: $\Sigma_j = \Sigma\ \forall\ j \in [M]$).

[3] e.g.: máxima verosimilitud, método de momentos, et cetera

2. Estimamos $\hat{\mathcal{L}}_j = N\left(\hat{\mu_j}, \hat{\Sigma}\right)$ como normales de medias independientes y varianza única por máxima verosimilitud,

$$\hat{\mu_j} = N_j^{-1} \sum_{i=1}^{N_j} x_i^{(j)},\ \forall j \in \{0, 1\}$$
$$\hat{\Sigma} = N^{-1} \sum_{j=1}^{M} \sum_{i=1}^{N_j} (x_i^{(j)} - \hat{\mu_j})(x_i^{(j)} - \hat{\mu_j}).$$

3. El clasificador es simplemente la función indicadora $\mathbf{1}\{\cdot\}$del discriminante lineal

$$\mathcal{R}(x) = \mathbf{1}\{w \cdot x > c\},\ \text{donde } w = \hat{\Sigma}^{-1}(\hat{\mu_1} - \hat{\mu_0}) \text{ y } c = w \cdot \frac{1}{2}(\hat{\mu_1} + \hat{\mu_0})$$

Generalmente, mientras más restrictivos sean los supuestos de (1), más sencilla de computar será la regla (3), y viceversa. y la generalidad del clasificador resultante: el trabajo del buen científico será encontrar el compromiso óptimo.[4]

En el ejemplo de **??**, 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, que acotan el espacio de densidades posibles a aquellas que se pueden expresar de forma cerrada con una expresión predefinida (en este caso, la densidad normal), y Q parámetros (aquí, μ_0, μ_1, 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 de probabilidad[5]. A estos se los conoce como métodos *no paramétricos* de estimación de densidad.

8.3. Estimación de densidad por núcleos

La estimación de densidad por núcleos (KDE[6], por sus siglas en inglés), es uno de los métodos mejor estudiados dentro del amplio universo no-paramétrico. Introducidos hacia 1960[**?**, **?**] 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

[4] N. del A.: En las famosas palabras de George Box,

> Todos los modelos son incorrectos, pero algunos son útiles.

El modelado estadístico es, aún hoy, más arte que técnica, y en palabras de Picasso,

> Todos sabemos que el arte no es la verdad. El arte es una mentira que nos acerca a la verdad, al menos aquella que no es dado comprender. El artista debe saber el modo de convencer a los demás de la verdad de sus mentiras.

[5] e.g.: asumir que la media y dispersión son finitas

[6] Kernel Density Estimation

llamada *núcleo* ("kernel") y un parámetro de suavización h, el *ancho de banda* ("bandwith"). La notación y nomenclatura para KDE es heterogénea; en la exposición que sigue, tomaremos de referencia el abarcador tratado de [**?**]

Definición 6 (función núcleo)**.** (función núcleo) Una función $K : \mathbb{R} \to \mathbb{R}$ es un *núcleo* ("kernel"), si

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

Observación 7. La no-negatividad y simetría no son forzosamente necesarias, pero van a otorgarles propiedades muy convenientes al estimador resultante. Cualquier función de densidad univariada cumple con la no-negatividad y normalización, y muchas, como la ley uniforme y la gaussiana, son además simétricas, convierti´ndolas en núcleos bastante populares.

Observación 8. Para todo núcleo K y $\lambda \in \mathbb{R}$, $J(u) = \lambda K(\lambda u)$ también es un núcleo, lo cual permite construir núcleos adecuadamente escalados a los datos. Usaremos la notación $K_h(u) = h^{-1}K(u/h)$ para referirnos a estos núcleos escalados.

Definición 9 (KDE univariado)**.** Sea $\{\mathbf{X}\}$ una muestra de N elementos aleatorios i.i.d. tomada de cierta distribución univariada con densidad desconocida f. Su estimador de densidad por núcleos (su "KDE") es

$$\widehat{f}(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big)$$

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

Definición 10 (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 $\hat{f}^{(1)}, \ldots, \hat{f}^{(M)}$ los M estimadores de densidad de cada clase obtenidos obtenidos según **??**. El "clasificador por estimación de densidad nuclear" correspondiente será:

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

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 [7], la clasificación "dura" de **??** 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? ¿Y si

[7] i.e., $\exists k \in [M] : f_h^{(k)}(x_0) \gg 0\ ,\ f_h^{(j)} \simeq 0 \ \forall\ j \in [M]\,/k$

creemos que las clases no son equiprobables a priori? Como las $\hat{f}_h^{(j)}$ estimadas identifican distribuciones, podemos utilizar la regla de Bayes para construir un «clasificador suave». Sea $p(A)$ la probabilidad de A, y consideremos que la proporción muestral de cada clase es una distribución *a priori* razonable para las clases bajo consideración (es decir, $\hat{p}(C_j) = N_j/N$ es un estimador insesgado de $p(C_j)$). Luego,

Definición 11 (clasificador KDE univariado suave)**.** Sea el **??** y los estimadores **??**. Por la regla de bayes,

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

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:

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

El vector $M-$dimensional $(\hat{p}_1, \ldots, \hat{p}_M)$ es una «clasificación suave» de x en las M posibles clases disponibles.

8.4. KDE multivariado

En el contexto univariado, no hay direcciones en el espacio, sólo sentido, positivo o negativo. Más aún, el peso de cada X_i en $\hat{f}(x)$ es $K(x - X_i)$, y como K es simétrica respecto al 0, sólo importa el *valor absoluto* - la *distancia* - entre el nuevo punto y cada observación. En una dimensión al menos, el núcleo K pondera - escalando por h - la distancia (euclídea) entre el punto a clasificar y cada datum:

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

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

Definición 12 (KDE multivariado)**.** (Sección 4.2 en [**?**]) En su forma más general, el estimador de densidad nuclear $d-$dimensional es

$$\widehat{f}(x; \mathbf{H}) = N^{-1} \sum_{i=1}^{N} K_{\mathbf{H}}(x - X_i)$$

donde $\mathbf{H}$ es una matriz $d \times d$ simétrica positiva definida, análoga al *ancho de banda* unidimensional h,

$$K_{\mathbf{H}}(x) = |\det \mathbf{H}|^{-1/2} K\left(\mathbf{H}^{-1/2} x\right)$$

y $K : \mathbb{R}^d \to \mathbb{R}_0^+$ es una función núcleo que satisface

$$\int_{\mathbb{R}^d} K(u)\, du = 1$$

Como en el contexto escalar, el núcleo suele ser una funciones de densidad aleatoria $d-$variada. Un núcleo muy popular es el la densidad normal estándar

$$K(x) = (2\pi)^{-d/2} \exp\left(-\frac{\|x\|^2}{2}\right)$$

,

un núcleo *esférico* o *radialmente simétrico*. En este caso, $K_{\mathbf{H}}(x - X_i)$ es equivalente a la densidad $\mathcal{N}(X_i, \mathbf{H})$ en el vector x.

Cuando $\mathbf{H} = h^2\mathbf{I}$, el estimador resultante es consistente con el producto de d estimadores **??**s:

$$\begin{aligned}\widehat{f}(x; h^2\mathbf{I}) &= N^{-1} \sum_{i=1}^{N} \left|\det h^2\mathbf{I}\right|^{-1/2} K\left(\left(h^2\mathbf{I}\right)^{-1/2}(x - X_i)\right) \\ &= N^{-1}h^{-d} \sum_{i=1}^{N} K\left((x - X_i)/h\right)\end{aligned}$$

Observación 13 (distancia de Mahalanobis). Dada una distribución de probabilidad Q en $\mathbb{R}^d$, con media $\mu \in \mathbb{R}^d$ y matriz de covarianza positiva definida $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, la *distancia de Mahalanobis*[8] de un punto x a Q es

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

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

$d_M(x, y; Q) = \|W(x - y)\|$

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

Reemplazando $W = \mathbf{H},\ \mu = X_1, \ldots, X_N$, podemos redefinir el estimador **??** como un estimador de núcleos basado en la distancia de Mahalanobis de x a las distribuciones $\mathcal{N}(X_i, \mathbf{H})$.

La relativa sencillez para el cómputo del método hasta aquí descrito lo hace un perenne favorito entre los estimadores de densidad no paramétricos, en particular cuando tomamos $\mathbf{H} = \mathbf{C}$, donde $\mathbf{C}$ es el estimador muestral de la covarianza de $\{\mathbf{X}\}$, lo cual simplifica radicalmente la cantidad de parámetros a ajustar. Sin embargo, el método posee algunas conocidas desventajas [**?**]:

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

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 y dispersión *global* $\mathbf{C}$ computada en la muestra completa.

2. Aún cuando la estimación global de $\mathbf{C}$ sea localmente adecuada, no resulta inmediatamente obvio que la suavización $\mathbf{H} = \mathbf{C}$ inducida por la muestra sea óptima en términos de representación de la densidad para regiones de alta densidad y outliers a la vez. Los estimadores por densidad nuclear así construidos suelen suavizar de más[9] en regiones de alta densidad, y de menos[10] alrededor de los *outliers* en la muestra.

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.

Distintos autores han intentado solucionar estas dificultades con éxito mixto.

8.4.1. Funciones de pérdida alternativas

Minimizar el (A)MISE como criterio de bondad en la evaluación de $\hat{f}$ responde antes que nada a conveniencias para la manipulación algebraica[11]. La diferencia *absoluta* del error integrado medio[12] es una alternativa atractiva: a diferencia del error *cuadrático*, el error absoluto es invariante con respecto a transformaciones monótonas de los datos[**?**]. A pesar de esta deseable propiedad, el tratamiento es arduo en $d = 1$ y excruciante en dimensiones mayores.

Motivado por la aplicación concreta de estimación de densidad al problema de clasificación, [**?**] toma un camino más directo: minimiza el *riesgo de Bayes* de $\hat{f}(x; h), h \in \mathbb{R}$, que tiene una interpretación inmediata en el **??**.

Definición 14 (riesgo de Bayes)**.** . Sea $\mathcal{R}(\cdot|f_1, \ldots, f_K)$ un **??** y una región $\Gamma \subseteq \text{dom } X$. El *riesgo de Bayes* asociado a $\mathcal{R}$ en Γ es

$$\text{err}(\mathcal{R}|\Gamma) = \sum_{j=1}^{K} p_j \int_{\Gamma} Pr(x \textbf{ no} \text{ sea clasificado por } \mathcal{R} \text{ como} \in C_j) f_j(x)\, dx$$

Hall muestra que el el clasificador de **??** es una regla óptima en el sentido del riesgo de Bayes - y por ende, para clasificación. Luego, sería razonable argumentar que elegir h como Hall propone es superador a optimizar h para el objetivo intermedio de estimar las verdaderas densidades de clase f_i. En un análisis concienzudo del caso $d = 1, K = 2, \Gamma \subset \mathbb{R}$, Hall halla que para el caso más sencillo $K = 2, d = 1$, según los signos de las derivadas f_1', f_2' en los puntos

[9] *oversmooth*
[10] *undersmooth*
[11] todos sabemos qué bien se portan los cuadrados
[12] $MIAE\left(\hat{f}_{\mathbf{H}}, f\right) = E\int_{\mathbb{R}^d}\left|\hat{f}_{\mathbf{H}}(y) - f(y)\right| dy$, y AMIAE análogo a AMISE**??**

de cruce de f_1, f_2 respectivas, el orden de magnitud del h óptimo varía drásticamente. El rango de «malos condicionamientos» que llevan a estas situaciones, sin embargo, se vuelve mucho más angosto en el problema multivariado ($d > 1$) o de múltiples clases ($K > 2$). En estos contextos, el ancho de banda óptimo es generalmente el mismo que es apropiado para estimación de densidad, según (A)M[I|S]E.

8.4.2. La elección de H en el caso bivariado

Para ilustrar la creciente complejidad en la elección de los coeficientes de $\mathbf{H}$, consideremos el caso multivariado más sencillo, $d = 2$, siguiendo a [**?**] que realizan un estudio exhaustivo de este problema, en relación a un conjunto de densidades f que se desea estimar vía KDE, con distintas propiedades[13] que dificulten la tarea.

Consideremos las 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 muestral $\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{bmatrix}$:
  - ignorando la correlación $\mathcal{C}_2 \coloneqq \left\{h^2\mathbf{D}; h^2 > 0\right\}$, con $\mathbf{D} = \text{diag}\,(c_{11}, c_{22})$,
  - 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 $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3$, $\mathcal{C}_2 \subseteq \mathcal{H}_2$, $\mathcal{C}_3 \subseteq \mathcal{H}_3$, $\mathcal{Y} \subseteq \mathcal{H}_3$. Para cada distribución estudiada y familia de matrices $\mathbf{H}$, se elige la matriz de ancho de banda

[13] e.g., con componentes con y sin correlación, sesgadas, kurtoticas, bi-, tri- y tetra-modales.

optimizando el *error cuadrático integrado medio asintótico*[14], y luego analizan la *eficiencia relativa asintótica*[15] de cada familia $\mathcal{A} \in \{\mathcal{H}_1, \mathcal{H}_2, \mathcal{C}_2, \mathcal{C}_3, \mathcal{Y}\}$, en comparación con la familia «irrestricta» $\mathcal{H}_3$, que en virtud de darle tantos grados de libertad como es posible a $\mathbf{H}$, tendrá siempre el menor AMISE - a costa de ser la más difícil de parametrizar.

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 considerables pérdidas de eficiencia, aún para la familia $\mathcal{Y}$, la más cercana a $\mathcal{H}_3$, cuando f es multimodal o se aleja de la normalidad.[16]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 $\mathbf{H}$.

8.4.3. El caso $d-$dimensional

Si el dominio bivariado ya presentaba suficientes dificultades para que ningún método de de elección de $\mathbf{H}$ «dominase» a los demás para cualquier densidad f, el caso $d-$dimensional general no es excepción. Como secuela de [**?**], los autores proponen un estimador "plug-in" del $\mathbf{H}$ óptimo - en el sentido de AMISE[17] - que se puede calcular para $\mathbf{H}$ «completa», a través de ciertos «funcionales» $\psi_{\mathbf{m}}$ que dependen de f y sus derivadas parciales de orden d[18]. Cuando se busca una matriz $\mathbf{H}$ completa, aún para dimensiones moderadas ($d \leq 5$) la cantidad de funcionales a estimar es enorme, por lo que luego se limitan a matrices diagonales para su aplicación concreta[**?**].

[**?**] sintetiza aportes propios y ajenos alrededor de la estimación de $\mathbf{H}$ completa según tres métodos de validación cruzada «deja-uno-afuera»: sesgada (BCV),

---

[14][AMISE] El error cuadrático medio integrado (MISE, por sus siglas en inglés) se define como

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

y su versión asintótica,

$$AMISE\left(\mathbf{H}\right) = \lim_{N\to\infty} MISE\left(\mathbf{H}\right)$$

Luego, fijada una densidad f cuyo KDE se desea estudiar, restringiendo $\mathbf{H}$ a una familia $\mathcal{A}$, se toma

$$\mathbf{H}^*_{\mathcal{A}} = \arg\inf_{\mathbf{H}\in\mathcal{A}} AMISE\left(\mathbf{H}\right)$$

[15][ARE] por *Asymptotic Relative Efficiency,* definido como

$$ARE\left(\mathcal{A} : \mathcal{B}\right) = AMISE\left(\mathbf{H}^*_{\mathcal{A}}\right) / AMISE\left(\mathbf{H}^*_{\mathcal{B}}\right)$$

[16]W&J (1993) tiene un lindo ejemplo "(F) Bimodal II" de cómo la covarianza estimada para una mezcla de dos gausianas con diferencias en la locación sobre el eje x, y mayor dispersión en el eje y, termina dando una estimación de la covarianza inútil para suavizado. Podría reproducirlo con scipy+matplotlib para ilustrar.

[17]ver **??**

[18]Sea $\mathbf{m}$ una $d-$tupla $\mathbf{m} = (m_1, \ldots, m_d)$ y $f^{(\mathbf{m})}$ la derivada parcial de f en $\mathbf{m}$, entonces $\psi_{\mathbf{m}} = \int f^{(\mathbf{m})}\left(x\right) f\left(x\right) dx$

insesgada (UCV), y (SCV)[19]. Con cada método, busca minimizar cierto error cuadrático: MISE para UCV; el asintótico AMISE en BCV y una combinación lineal de ambos define SCV. 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.

Una parametrización completa de $\mathbf{H}$ en d dimensiones requiere la hercúlea tarea de elegir $\binom{d}{1} + \binom{d}{2} = \left(d^2 + d\right)/2$ coeficientes[20]. El ya-mencionado trabajo de [**?**] toma un camino alternativo: pre-transformar los datos para que tengan media cero y matriz de covarianza unitaria[21], y luego intenta buscar h para los datos transformados. La práctica es equivalente a buscar $\mathbf{H}$ en la familia $\mathcal{C}_3 \coloneqq \left\{h^2\mathbf{C}; h^2 > 0\right\}$ de [**?**]. Hwang encuentra las dificultades listadas en (**??**), y compara varios algoritmos superadores en algún sentido al KDE con ancho de banda fijo (FKDE).

KDE Adaptativo (AKDE) Similar a FKDE, pero con un factor de ancho local λ_n para cada núcleo

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

Aunque cada núcleo estará mejor escalado a su contexto local, el enfoque sigue utilizando una misma orientación global $\mathbf{C}$ para todos los núcleos. 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.

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 a cada cluster un núcleo gaussiano, su altura y su ancho según la posición y cantidad de sus observaciones. Por esto, también se lo conoce com "modelado de mezclas gaussianas"[22]. Aunque el estimador final se puede expresar con tan pocos términos como clusters haya, 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 de N, d y la cantidad de clusters identificados, pueden aparecer 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.

[19] en inglés: Biased, Unbiased & Smoothed Cross Validation

[20] ¡y yo me trabo eligiendo entre té ver y té negro!

[21] práctica también conocida como «esferización» o «blanqueo». En particular, si $\bar{X}, \mathbf{C}$ aon respectivamente la media y covarianza muestral de $\{\mathbf{X}\}$, el conjunto $\{\mathbf{Z}\} \coloneqq \left\{Z_i = \mathbf{C}^{-1/2}\left(X_i - \bar{X}\right) \ \forall \ X \in \{\mathbf{X}\}\right\}$ es su equivalente blanqueado. Es fácil ver que $E\left[Z\right] = 0$, $E\left[ZZ^T\right] = \mathbf{I}$.

[22] GMM, o *Gaussian Mixture Modelling* en inglés

KDE por "persecución de la proyección" (PPDE) El espíritu de PPDE consiste en buscar iterativamente proyecciones "interesantes" de los datos en bajas dimensiones (típicamente 1-D), modificar la muestra original $\{\mathbf{Z}\}^{(0)}$ para remover la estructura encontrada en la proyección, e iterar el proceso en los datos resultantes. Siguiendo a [**?**], 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[23]. Un problema específico a PPDE, es que no puede lidiar satisfactoriamente con estructuras "escondidas" en alta dimensión, «detrás» de proyecciones en baja dimensión[24]

8.5. 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 $\mathbf{H}$, 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:

Observación 15 (maldición de la dimensionalidad). Sea $B\left(c, r, d\right)$ la bola $d-$dimensional de radio r centrada en $c \in \mathbb{R}^d$, y X v.a. uniformemente distribuida (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, $Pr\left(X \in B\left(0, r-\epsilon, d\right)\right)$?

Como la distribución de X es uniforme en volumen, y $B\left(0, r-\epsilon, d\right) \subset B\left(0, r, d\right)$, basta con comparar los volúmenes de de ambas $d-$esferas para encontrar la solución. El volumen$d-$dimensional de una bola es

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

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

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

[23] [**?**, p. 29]

[24] e.g., las proyecciones unidimensionales de una densidad en $\mathbb{R}^2$ con forma de dona no dan cuenta fehaciente de la estructura original. Estimar estas *variedades* escondida en el espacio ambiente euclídeo será un punto central de nuestro trabajo más adelante.

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 $x \in B(0, r, d) \iff \|x\| \leq r$, y similarmente $x \notin B(0, r-\epsilon, d) \iff \|x\| > (r-\epsilon)$, podemos escribir

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

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 cualquier observación al centro de la esfera tiende a r. En altas dimensiones, la distancia euclídea resulta inútil para diferenciar entre elementos muestrales.

8.6. La hipótesis de la variedad

El resultado previo descansa sobre el hecho de que la distribución de X sobre su soporte $\text{sop}(X) = B(0, r, d) \subset \mathbb{R}^d$ 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 megapíxel como puntos X pertenecientes al espacio $\mathbb{R}^{1024\times1024}$, pero si tomamos una imagen 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*[25]

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

[25] Un desarrollo contemporáneo sumamente interesante es el de [**?**], que se puede traducir como *Teoría de los «caparazones»: Un modelo estadístico de la realidad.* Los autores observan que en teoría, debido a la **??**, el aprendizaje estadístico debería ser lisa y llanamente imposible en altas dimensiones, pero en la práctica se ve que funciona. Propone un marco estadístico riguroso destinado a concebir el aprendizaje automático en alta dimensión, la «teoría de los caparazones» - aunque en inglés suena más bonito, *shell theory* . Fundado en la observación de que las relaciones entre objetos que deseamos entender forman una jerarquía (gato siamés $\subset$ gato $\subset$ animal), propone que las observaciones en alta dimensión son resultado de un proceso de "generadores jerárquicos". Desarrollando una noción de distancia adecuada, muestran que en dichos procesos generativos, las instancias de cada proceso en la jerarquía - casi siempre - están encapsuladas por un «caparazón» distintivo que excluye a (casi) cualquier otra instancia, y permite identificar clases rigurosamente.

lo que nos embrolló en un principio. Dadas $\{\mathbf{X}\} = \left\{X_i | X_i \in \mathbb{R}^{d_x},\ i \in [N]\right\}$, 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[**?**] hace un excelente censo. El autor relaciona la tarea del área con la noción geométrica de una *variedad,* a través de la *hipótesis de la variedad.*[26].

Observación 16. A nuestros fines, una variedad $\mathcal{M}$ de dimensión $d_{\mathcal{M}}$, es un espacio 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 cualquier tipo de observaciones multivariadas obtenidas «naturalmente»: sonidos, texto, secuencias genómicas y hasta las respuestas al eterno cuestionario del censo. Siguiendo a [**?**],

> 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 variedades que soportan dominios complejos del mundo real, sin embargo, se espera que sean fuertemente no-lineales.

[26] El término no es del todo riguroso pero figura frecuentemente en la literatura sobre aprendizaje automático. El mismo Bengio se explaya sobre el origen del término en Reddit, y [**?**] distingue entre varias formulaciones íntimamente relacionadas de la misma hipótesis.

Más que una propiamente dicha hipótesis falsificable al respecto de la distribución de los datos, mencionamos la *hipótesis de la variedad* como un 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 - un segmento curvo - embebida en el espacio cartesiano - $\mathbb{R}^2$, una variedad 2-dimensional. En la variedad latente, los dos puntos extremos del dibujo de la "U" 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. La razón de tal insensatez, es simplemente, que hemos tomado una medida de distancia que no se ajusta bien al espacio latente.

8.7. 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 en la circunferencia S^1 y la esfera S^2, fueron estudiadas temprano en el siglo XX [27], pero la estimación de densidad en variedades arbitrarias no parece haber sido tratado sistemáticamente antes de [**?**], quien convenientemente hizo exactamente eso: proponer un estimador de densidad por núcleos en variedades Riemannianas. En lo que sigue, intentamos ser fieles a lo que entendimos de la exposición de Bruno, y suplimos algunos agujeros teóricos con la más amena tesis de licenciatura de [**?**].

Definición 17 (variedad Riemanniana)**.** ([**?**, §3.3.2]). Una variedad Riemanniana es una variedad diferenciable $\mathcal{M}$ dotada de un métrica Riemanniana g, que denotaremos con $(\mathcal{M}, g)$[28].

[27] Motivado por el estudio de los pesos atómicos de elementos químicos, [**?**] (en alemán) introduce la «distribución von Mises» en la circunferencia S^1, adaptando la densidad normal estándar univariada. [**?**] propone la «distribución de Fisher» en la esfera S^2para desarrollar un test sobre la dirección de flujos de lava en Islandia. A mediados del siglo XX el campo de la «estadística direccional» - un antecesor directo de la estadística en variedades arbitrarias - estaba bien desarrolada, y [**?**] estudia en detalle la «distribución de von Mises-Fisher», generalización $d-$dimensional de los aportes antedichos.

[28] Para quienes no entienden casi absolutamente nada de geometría diferencial como yo, la tesis de [**?**] es un excelente puente a los trabajos que se citan en las siguientes secciones, en especial §3 de íbid., «Preliminares Geométricos».

Definición 18 (KDE en variedades)**.** ([**?**, §2]) 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 $\mathcal{M}$ [29] 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

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[30] 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 $\mathcal{M}$[31]. Definimos el estimador de densidad de $\hat{f}(p|N, K)$ como el mapa $\hat{f} : \mathcal{M} \to \mathbb{R}$ que a cada $p \in \mathcal{M}$ le asocia el valor $\hat{f}(p)$ definido como

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

Observación 19. (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 $f_{N,K} = N^{-1} \sum_{i=1}^{N} h^{-d} K(\|p - X_i\|/h)$. La expresión de $f_{N,K}$ es consistente con la expresión de **??** como producto de d **??** de idéntico ancho de banda h . Sea $\mathcal{M} = \mathbb{R}^d$ y $d_M(p, q|\Sigma)$ la **??** con covarianza Σ. Como $\left\|\Sigma^{-1/2}(p - q)\right\| = d_M(p, q|\Sigma)$ y Σ es una transformación lineal, $\forall p, q \in$

[29] Más precisamente, $X : \Omega \to \mathcal{M}$ es 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 $\mathcal{M}$.

[30] iso-tropos: igual (iso) en toda dirección (tropos). El núcleo gaussiano estándar es isotrópico.

[31] 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 $\exp_P^{-1}(q)$.

$\mathbb{R}^d$, $\theta_p(q) = \det \Sigma^{1/2} = (\det \Sigma)^{1/2}$ y

$$\begin{aligned}
\widehat{f}(x|\Sigma) &= N^{-1} \sum_{i=1}^{N} \frac{1}{h^d} \frac{1}{\theta_{X_i}(p)} K\left(\frac{d_g(p, X_i)}{h}\right) \\
&= N^{-1} \sum_{i=1}^{N} \frac{1}{h^d} \frac{1}{(\det \Sigma)^{-1/2}} K\left(\frac{d_M(p, X|\Sigma)}{h}\right) \\
&= N^{-1} \sum_{i=1}^{N} \frac{1}{h^d} \frac{1}{(\det \Sigma)^{-1/2}} K\left(\frac{\left\|\Sigma^{-1/2}(x - X_i)\right\|}{h}\right) \\
&= N^{-1} \sum_{i=1}^{N} \frac{1}{h^d} K_\Sigma\left(\frac{\|x - X_i\|}{h}\right)
\end{aligned}$$

y el estimador en variedades es consistente con el caso general del **??**.

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

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

, donde $\text{inj}_g \mathcal{M}$ es el *radio de inyectividad* de $\mathcal{M}$[32]. Sin entrar en demasiados detalles, siempre y cuando $\mathcal{M}$ sea compacta este radio de inyectividad será > 0, y al menos para los resultados asintóticos (cuando el tamaño muestral es lo suficientemente grande como para que $h \to 0$), siempre existe un $0 < h < h_0$ posible.

Pelletier avanza algunas propiedades elementales de este estimador: adapta el concepto de «media» para elementos aleatorios en $\mathbb{R}^d$ a el de «media intrínseca» en variedades Riemannianas compactas sin frontera [**?**, Prop. II]), y prueba que es consistente para f en el siguiente sentido

Teorema 21. *[**?**, Teorema 5]* Sea f una densidad de probabilidad dos veces diferenciable en $\mathcal{M}$ con segunda derivada covariante acotada. Sea $\hat{f}$ su estimador

[32]

Definición 20. [**?**, p. 108][**?**, p. 23] Sea $(\mathcal{M}, g)$ una variedad Riemanniana de dimensión d. Llamamos *radio de inyectividad* a

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

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 $\text{inj}_g \mathcal{M}_1 = \infty$, pues todo el espacio comparte un único mapa de coordenadas global. Si le quitamos un punto, $\mathcal{M}_2 = \mathcal{M}_1 - \{p\}$ entonces $\text{inj}_g \mathcal{M}_2 = 0$ (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. Lo ventajoso de que h esté por debajo del radio de inyectividad, será que al integrar sobre las bolas que soportan el núcleo K alrededor de cada $p \in \mathcal{M}$, la densidad se podrá integrar - luego de una transformación - en $\mathbb{R}^d$, donde nuestras herramientas tradicionales han sido afinadas.

definido en **??** con ancho de banda h que satisface la condición **??**. Luego, existe una constante C_f tal que

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

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

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

Pelletier considera la convergencia en $L^2(\mathcal{M})$, ¿esto qué tipo de consistencia sería? ¿débil? ¿Qué diferencia hay con la que estudian 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

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

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

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

[**?**] se apalancan sobre los resultados previos proponiendo un clasificador binario ($M = 2$) basado en núcleos, para e.a. soportados sobre variedades compactas y cerradas de Riemann. Recordemos que en **??** propusimos un clasificador suave que asignase a cada clase, una probabilidad de pertenencia

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

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

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

y su estimador muestral

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

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, adaptado para M clases. Considerando como función objetivo a minimizar la misma probabilidad de error de clasificación que vimos con [**?**],

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

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 [**?**, §6 Consistencia], 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.

8.8. Variedades desconocidas

Los resultados combinados de **??** 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 $\mathcal{M} = \text{sop}X$ implica desconocer:

- su dimensión intrínseca $d_{\mathcal{M}}$,
- la distancia geodésica d_g,
- y la función de densidad de volumen θ_p,

aún *antes* de estimar la densidad $f : \mathcal{M} \to \mathbb{R}^+$, que nos trajo hasta aquí. Por partes o juntas, tendremos que *aprenderlas de los datos* de alguna manera.

8.8.1. La distancia geodésica d_g

Dada la naturaleza localmente euclídea de las variedades, para puntos "vecinos" entre sí, la distancia en $\mathbb{R}^{d_x}$ (en el espacio «ambiente» en que está «embebida» $\mathcal{M}$) será una aproximación razonable a la distancia geodésica en la variedad. 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.

Esta inocente observación es el núcleo de la innovación de Isomap[33], algoritmo presentado en [**?**, **?**] con el objetivo de "aprender la geometría global subyacente de un dataset, usando información métrica local fácilmente medible", de entre un conjunto amplio de variedades no-lineales. Su tarea central, consiste en aproximar adecuadamente las distancias geodésicas en la variedad $d_g(p,q)$ entre puntos alejados, conociendo únicamente las distancias euclídeas en la muestra $\|p-q\|$.

El algoritmo completo, consta de tres pasos principales [**?**, Tabla 1]:

1. **Constrúyase un grafo de vecinos muestrales NN** $= (\{\mathbf{X}\}, E)$ sobre el dataset completo, donde la arista $x \leftrightarrow y$ está incluida si $\|x-y\|_{d_x} < \epsilon$ ("$\epsilon-$Isomap"), o si y es uno de los K vecinos más cercanos de x («$K-$isomap») . 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 G. Los costos de los caminos mínimos $d_{\mathbf{NN}}(x,y)$ constituyen una aproximación de las distancias geodésicas $d_g(x,y)$.

3. **Constrúyase un *embedding* d$-$dimensional.** Utilizando escalamiento multidimensional[34], un algoritmo de reducción de dimensionalidad), crear una representación («embedding») en el espacio euclídeo $\mathbb{R}^d$ que minimice una métrica de discrepancia denominada «estrés», entre las distancias $d_{\mathbf{NN}}$ antes computadas con las distancias en la representación a construir $\|\cdot - \cdot\|_{\mathbb{R}^d}$.

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 de la que estamos sampleando.

Allende los costos computacionales, hay dos parámetros a fijar en este algoritmo. El primero es el parámetro de "vecindad" ϵ ó K de (1), cuyo valor óptimo no es trivial determinar. 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 «cortocircuitar» la representación - incluir en G aristas $e \in E$ que cruzan el espacio ambiente $\mathbb{R}^{d_x}$ completamente por fuera de $\mathcal{M}$. Consideraciones análogas complican la elección de cantidad de vecinos en $K-$Isomap.

[33] **Iso**metric feature **map**ping, en inglés
[34] MDS, **M**ulti**d**imensional **S**caling

El otro parámetro de interés es la dimensión d del embedding euclídeo «óptimo». Inspeccionando el gráfico de "estrés" de MDS como función de la dimensión d escogida, se pueden buscar punto(s) de inflexión ("codos") 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 $d_{\mathcal{M}}$. Al menos en ejemplos sintéticos, al método del codo lo heurístico no le quita lo certero. La representación $d-$dimensional que produce MDS no es la variedad $\mathcal{M}$ que buscamos, pero sí es razonable que con suficientes datos, $d \approx d_{\mathcal{M}}$.

8.8.2. La dimensión intrínseca $d_{\mathcal{M}}$

La literatura que intenta estimar directamente $d_{\mathcal{M}}$, compensa su escasez con creatividad. [**?**] se propone no sólo estimar $d_{\mathcal{M}}$, sino además ofrecer un algoritmo para proveer un «atlas» [35] de $\mathcal{M}$, una representación harto útil.

Llamemos $n(r)$ a la «función de conteo» que indica cuántos puntos de $\{\mathbf{X}\}$ se encuentran dentro de una bola en $\mathbb{R}^{d_{\mathcal{M}}}$centrada en un punto $p \in \mathcal{M}$. $n\,(r)$ debería crecer a tasa $r^{d_{\mathcal{M}}}$, pero únicamente en la escala en la que la variedad es efectivamente localmente lineal. Si hay ruido en la medición en $\mathbb{R}^{d_x}$, en la mínima escala los puntos se encontrarán en toda dirección y $n\,(r)$ crecerá a tasa r^{d_x}; en escalas mayores a la localmente lineal, la tasa de crecimiento de $n\,(r)$ también será mayor, pues la variedad ya no es perpendicular a la superficie de la bola, y la curvatura hace que r no deba crecer tan rápido para incorporar nuevos puntos. Un cuidadoso análisis de la tasa de crecimiento de $n\,(r)$ permitiría identificar la dimensión más probable de la variedad. Aunque teóricamente llamativo, el resultado es costoso de computar y no tan obvio de interpretar para datasets «naturales» o sintéticos pero de pequeño N.

[**?**] presenta una estrategia más directa. El vecindario local de un punto $p \in \mathcal{M}$ debería encontrar a sus vecinos en un subespacio lineal de dimensión $d_{\mathcal{M}}$, y espacio ambiente vacío en las demás dimensiones. De computar PCA[36] para el vecindario de p, esperaríamos encontrar $d_{\mathcal{M}}$ direcciones principales con autovalores ordenados $\lambda_1, \ldots, \lambda_{d_{\mathcal{M}}}$ significativos, y $\lambda_i \approx 0,\ i > d_{\mathcal{M}}$. A partir de esta observación, proponen esencialmente un estimador **??** con una $\mathbf{H}_i$ elegida específicamente para cada X_i en función de su vecindario «suave» o «duro»[37], añadiendo $\sigma^2\mathbf{I}$ a las las $d_{\mathcal{M}}$ direcciones principales (a escala). El regularizador $\sigma^2\mathbf{I}$ provee dos ventajas: evita tener que guardar las d_x componenetes principales

[35] El par (U, φ) compuesto por un conjunto abierto U medible en $\mathcal{M}$, y un homeomorfismo $\varphi : U \to A \subset \mathbb{R}^{d_{\mathcal{M}}}$ también abierto se denomina «carta» («chart»), o «entorno coordenado». Un conjunto de cartas «compatibles» entre sí cuya unión sea la variedad $\mathcal{M}$ es un «atlas», exactamente como llamamos en cartografía a un conjunto de mapas - euclídeos en $\mathbb{R}^2$- cuya unión representa la superficie terrestre S^2. El trabajo de [**?**] es sumamente interesante, aunque queda algo por fuera del ya extenso paseo bibliográfico. del trabajo de [**?**, §3.1 "Variedades Diferenciables"@] provee los preliminares necesarios para entenderlo.

[36] Por «consideraciones prácticas», los autores no implementan PCA sino la descomposición en valores singulares, SVD, y toman los $d_{\mathcal{M}}$ mayores valores singulares en lugar de los respectivos autovalores. Por qué han de hacerlo así no me queda del todo claro.

[37] Respectivamente, ponderando la contribución a la matriz de covarianza de cada vecino según un núcleo gaussiano (vecindario «suave»), o calculándola tradicionalmente sobre los K vecinos más cercanos («duro»)

para cada punto, y nos asegura que $\mathbf{H}_i$ siempre esté bien condicionada, aún cuando el vecindario tiene sólo $K < d_x$ vecinos. Cuando $d_{\mathcal{M}}$ no se conoce de antemano, una heurística como el «método del codo» antedicho o «tantos autovalores como sean necesarios para explicar $X\,\%$ de la varianza» debería funcionar razonablemente bien.

8.8.3. La densidad de volumen $\theta_p(q)$ - TBD

8.9. Distancias basadas en densidad

8.9.1. De Isomap al presente

Al núcleo de [**?**, **?**] y otros, está la idea de considerar el grafo de vecinos más cercanos **NN** de $\{\mathbf{X}\}$ como aproximación a la estructura de $\mathcal{M}$, y asumir que en los vecindarios de cada punto la distancia euclídea aún es representativa. Cuando $\mathcal{M}$ está ralamente muestreado, o tiene una curvatura considerable, aún este supuesto relativamente benigno puede resultar fatal.

[**?**][38] ya sugiere una alternativa heurística en el contexto de clustering: construir un grafo con aristas pesadas sobre $\{\mathbf{X}\}$ con pesos iguales *al cuadrado* de la distancia (euclídea) entre sus extremos y tomar como distancia entre vértices el costo de camino mínimo correspondiente. Esencialmente, lo mismo que Isomap pero con costo $\|x-y\|^2$en el primer paso. El cuadrado castiga más severamente los saltos entre puntos alejados, y favorecerá caminos mínimos que pasen por regiones de alta densidad. El trabajo de [**?**] ya habla explícitamente de «distancias basadas en densidad»[39], definidas a partir de transformaciones g monótonamente decrecientes en la densidad f. Más aún, en el caso de la familia $g(f|r) = f^{-r}$ que resulta de pesar el grafo según $\|x-y\|^q$ donde $q = rd+1$, considera su estimación empírica en el grafo completo de vértices $\{\mathbf{X}\}$. La dimensión intrínseca d de la variedad a estimar casi nunca es conocida de antemano, pero esto no es obstáculo para aplicar esta familia: podemos elegir - por validación cruzada, por ejemplo - q directamente, y atrapar en dicho parámetro r, d a la vez.

Observación 22. [40]Cuando la densidad es efectivamente uniforme en la variedad, f es constante en $\mathcal{M}$ y g también, así que medir la distancia entre puntos según $\|x-y\|$es óptimo. Lamentablemente, las densidades que buscamos estimar nunca son uniformes.

Observación 23. Para Isomap, el parámetro de vecindad es clave, en tanto «esculpe» la estructura local del grafo completo. Al usar una distancia basada en densidad, tal restricción ya no es necesaria. Se puede elegir un k, ϵ pequeño por consideraciones computacionales, pero en principio las distancias basadas en densidad sólo se benefician al agrandar los vencindarios a considerar.

La década de los 2010 fue fructífera para las distancias basadas en densidad, y hacia fines de ella se dan múltiples resultados sólidos casi en paralelo. [**?**]

[38] [**?**] lo cita, pero no me resultó posible encontrar el PDF del trabajo original. Atención a la fecha: 2003, hace dos décadas, probablemente por el trabajo de [**?**] fresquito en la memoria.
[39] DBDs, density-based distances.
[40] [**?**, §3]

prueba una relación sorprendente entre la distancia de vecinos más cercanos y la de «aristas cuadradas»

Definición 24. [**?**, Definición 1.1] Dada una función de costo continua $c : \mathbb{R}^d \to \mathbb{R}$ definimos el costo «basado en densidad» de un sendero γ relativo a c como $c(\gamma) = \int_0^1 c(\gamma(t)) \|\gamma'(t)\| dt$, donde el sendero γ es un mapa continuo $\gamma : [0,1] \to \mathbb{R}^d$. Sea senderos (p,q) el conjunto de senderos C^1 de a tramos[41]. Definimos la DBD entre dos puntos $p, q \in \mathbb{R}^d$ como

$$d_c(p,q) = \inf_{\gamma \in \text{senderos}(p,q)} c(\gamma)$$

Definición 25. [**?**, Definición 1.2]Sea $Q \subseteq \mathbb{R}^d$ un conjunto finito. Definiremos la métrica de vecino más cercano, $\mathbf{r}_Q(a) = 4\,\text{mín}_{q \in Q} \|a - q\|$ y la distancia asociada como

$$d_{\mathbf{N}}(a,b) \coloneqq d_{\mathbf{r}_Q} \forall\ a, b \in \mathbb{R}^d$$

.

Definición 26. [**?**, Definición 1.3]Para un conjunto de puntos $Q \in \mathbb{R}^d$,la distancia de aristas cuadradas $\forall\ p, q \in Q$ es

$$d_{\mathbf{2}}(a,b) = \inf_{(q_0,\ldots,q_k)} \sum_{i=1}^{k} \|q_i - q_{i-1}\|^2$$

donde el ínfimo es sobre secuencia de k puntos con $q_0 = a$ y $q_k = b$.

Teorema 27. *[**?**, Definición 1.3] La métria de vecino más cercano y la métrica de aristas cuadradas son equivalentes para cualquier conjunto finito de puntos Q en dimensión arbitraria*[42].

Contar con una aproximación en un grafo finito para computar una distancia sobre senderos arbitrarios en el espacio ambiente es un resultado muy poderoso. Sin embargo, ya [**?**, §2] mencionaba que la construcción de un estimador $\hat{f}$ basado en la métrica de vecino más cercano es insesgado para la *mediana*, pero no es consistente, pues su varianza permanece constante aún cuando $N \to \infty$.

8.9.2. Distancia de Fermat

El estudio de las distancias correspondientes a las funciones de costo $g = f^{-r}$, equivalentes continuos a la distancia de camino mínimo en el grafo pesado por $\|x - y\|^q$, es estudiado por [**?**, **?**, **?**], con diferencias de notación y aplicación, pero

[41] i.e., con primera derivada continua. Es el conjunto de senderos sobre los que es factible computar $c(\gamma)$

[42] El resultado es aún más fuerte: establece que $d_{\mathbf{N}} \equiv d_{\mathbf{2}}$ para todo Q sea una colección finita de conjuntos compactos conectados por senderos. Es decir, si reemplazamos los puntos por «regiones compactas» del espacio - que no tenga costo atravesar -, la equivalencia aguanta.

no sustantivas. [?, Def. 2.1] considera la generalización $\mathbf{d}_\alpha$ de $\mathbf{d_2}$, que llaman «distancia de Fermat muestral»[43]

$$d_{Q,\alpha}(a,b) = \inf_{(q_0,\ldots,q_K)} \sum_{i=1}^{K} \|q_i - q_{i-1}\|^\alpha, \ \alpha \geq 1 \tag{1}$$

Los autores definen esta distancia muestral para conjuntos arbitrarios Q, pero en general 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 se preste a confusión, omitiremos la dependencia en Q, α. A continuación, se define la versión *macroscópica* de la distancia de Fermat muestral.

Definición 28 (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 macroscópica*[44] $\mathcal{D}_{f,\beta}(s,t)$ 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:

Teorema 29. [?, Teorema 2.7] Sea $\mathcal{M}$ una variedad d-dimensional, isométrica y C^1 embebida en $\mathbb{R}^D$[45]. 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.[46]

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.

[43] [?, ?] consideran una versión «normalizada» de la distancia de Fermat, $(d_\alpha(a,b))^{1/\alpha}$. Donde éstos últimos autores consideran una f definida en la unión de variedades disjuntas y usan la distancia resultante para «clustering espectral», los primeros consideran una única variedad compacta y usan la distancia en clustering por $K-$medoides. Para evitar estirar este de por sí extenso censo del arte, la exposición posterior corresponde únicamente a [?]

[44] O *distancia de Fermat*, a secas.

[45] 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 $\phi(\bar{S}) = \mathcal{M}$. En aplicaciones reales se espera que $d \ll D$, pero no es necesario.

[46] Debería unificar la notación de [?, ?, ?], creo que la de Chu es la más amena, pero estaría bueno revisarlas.

9. Propuesta

Hemos repasado en detalle la historia y motivación por detrás de un método eficiente y sumamente estudiado para responder al **??** en dominios de alta dimensionalidad: la estimación de densidad por núcleos (KDE), hasta llegar a definirla en variedades de Riemann. Notamos que de los tres parámetros a elegir - el núcleo, el ancho de banda y la distancia - tanto el ancho de banda como la distancia son problemáticos en alta dimensiones. Para KDE, la elección del ancho de banda el tratamiento encontrado en la literatura es extenso y exhaustivco; no así para la elección de la distancia. Nos proponemos elucidar si es posible mejorar la performance de **??** usando una noción de distancia basada en densidades de desarrollo reciente, la distancia muestral de Fermat. Más específicamente, construiremos

- un **??** en variedades según [**?**]
- con matriz de suavización $\mathbf{H}_i$ individualmente orientada en cada elemento muestral según [**?**]
- y distancia varietal aprendida según [**?**] por validación cruzada de α

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.

10. 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?

10.0.1. 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.

10.0.2. 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.

10.0.3. 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.

10.0.4. 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.

10.0.5. 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.

11. 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?

12. Análisis experimentale

13. Cuentita

14. Conclusiones