

Índice

1	TODOs	1
2	Vocabulario y Notación	1
3	Preliminares	3
3.1	El problema de clasificación	3
3.2	Estimación de densidad por núcleos	5
3.3	Estimación de densidad multivariada	8
3.4	Variedades de Riemann	13
3.5	Clasificación en variedades	29
3.6	Aprendizaje de distancias	30
4	Propuesta Original	49
4.1	Evaluación	50
5	Resultados	54
5.1	Chequeo de sanidad: <code>blobs</code>	54
6	Glosario	75
7	Listados	75
8	Tablas	77
9	Código	77
Bibliografía	77	
9.1	Slides sobre diseño experimental	80

1 TODOs

- [] Ponderar por n^β
- [] Evitar coma entre sujeto y predicado

2 Vocabulario y Notación

A lo largo de esta monografía tomaremos como referencia enclopédica al *Elements of Statistical Learning* (Hastie, Tibshirani y Friedman, 2009), de modo que en la medida de lo posible, basaremos nuestra notación en la suya también.

Típicamente, denotaremos a las variables independientes¹ con X . Si X es un vector, accederemos a sus componentes con subíndices, X_j . En el contexto del problema de clasificación, la variable *cualitativa* dependiente² será G (de *Grupo*). Usaremos letras mayúsculas como X, G para referirnos a los aspectos genéricos de una variable. Los valores *observados* se escribirán en minúscula, de manera que el i -ésimo valor observado de X será x_i (de nuevo, x_i puede ser un escalar o un vector).

¹También conocidas como predictoras, o *inputs*

²También conocida como variable respuesta u *output*

Representaremos a las matrices con letras mayúsculas en negrita, \mathbf{X} e.g.: el conjunto de de N vectores p -dimensionales $\{x_i, i \in \{1, \dots, N\}\}$ será representado por la matrix \mathbf{X} de dimensión $N \times p$.

En general, los vectores *no* estarán en negrita, excepto cuando tengan N componentes; esta convención distingue el p -vector de *inputs* para la i -ésima observación, x_i , del N -vector \mathbf{x}_j con todas las observaciones de la variable X_j . Como todos los vectores se asumen vectores columna, la i -ésima fila de \mathbf{X} es x_i^T , la traspuesta de la i -ésima observación x_i .

A continuación, algunos símbolos y operadores utilizados a lo largo del texto:

\mathbb{R} los números reales; \mathbb{R}_+ denotará los reales estrictamente positivos.

d_x

\mathbb{R}^{d_x}

$[k]$ el conjunto de los k números enteros, $\{1, \dots, k\}$

\mathcal{M}

\mathbf{H}

$\|\cdot\|$

$\{\mathbf{X}\}$

$X_{i,j}$

$\mathbb{1}(x)$ la función indicadora, $\mathbb{1}(x) = \begin{cases} 1 & \text{si } x \text{ es verdadero} \\ 0 & \text{si no} \end{cases}$

$\Pr(x)$ función de probabilidad,

$\mathbb{E}(x)$ esperanza,

$\text{Var}(x)$ varianza,

i.i.d. independiente e idénticamente distribuido (suele aplicar a una muestra \mathbf{X})

\emptyset el conjunto vacío

\overline{S} la *clausura* de S ; la unión de S y sus puntos límites.

$\lambda(x)$ la medida de Lebesgue de x en \mathbb{R}^d

$a \mapsto b$ la función que «toma» a y «devuelve» b en [notación de flechas](#)

$y \propto x$ « y es proporcional a x », existe una constante $c : y = c \times x$

c.s. «casi seguramente», al referirse a convergencia de v.v.a.a.

3 Preliminares

3.1 El problema de clasificación

3.1.1 Definición y vocabulario

El *aprendizaje estadístico supervisado* busca estimar (aprender) una variable *respuesta* a partir de cierta(s) variable(s) *predictora(s)*. Cuando la *respuesta* es una variable *cualitativa*, el problema de asignar cada observación X a una clase $G \in \mathcal{G} = \{\mathcal{G}^1, \dots, \mathcal{G}^K\}$ se denomina *de clasificación*. En general, reemplazaremos los nombres o «etiquetas» de clases g_i por los enteros correspondientes, $G \in [K]$. En esta definición del problema, las clases son mutuamente excluyentes y conjuntamente exhaustivas:

- mutuamente excluyentes: cada observación X_i está asociada a lo sumo a una clase
- conjuntamente exhaustivas: cada observación X_i está asociada a alguna clase.

Definición 3.1.1.1 (clasificador): Un *clasificador* es una función $\hat{G}(X)$ que para cada observación intenta aproximar su verdadera clase G por \hat{G} («ge sombrero»).

Para construir \hat{G} , contaremos con una muestra o *conjunto de entrenamiento* \mathbf{X}, \mathbf{g} , de pares $(x_i, g_i), i \in \{1, \dots, N\}$ conocidos.

Para discernir cuán bien se «ajusta» un clasificador a los datos, la teoría requiere de una función de *pérdida* $L(G, \hat{G}(X))$.³ Será de especial interés la función de clasificación f que minimiza la *esperanza de predicción errada* EPE:

$$\hat{G} = \arg \min_f \text{EPE}(f) = \arg \min_f \mathbb{E}(L(G, f(X))) \quad (1)$$

donde la esperanza es contra la distribución conjunta X, G). Por la ley de la probabilidad total, podemos condicionar a X y expresar el EPE como

$$\begin{aligned} \text{EPE}(f) &= \mathbb{E}(L(G, \hat{G}(X))) \\ &= \sum_{k \in [K]} \mathbb{E}(L(\mathcal{G}_k, \hat{G}(X)) \Pr(\mathcal{G}_k | X)) \mathbb{E}(X) \\ &= \mathbb{E}(X) \sum_{k \in [K]} \mathbb{E}(L(\mathcal{G}_k, g) \Pr(\mathcal{G}_k | X)) \end{aligned} \quad (2)$$

Y basta con minimizar punto a punto para obtener una expresión computable de \hat{G} :

³*loss function* en inglés. A veces también «función de riesgo» - *risk function*.

$$\begin{aligned}\hat{G}(x) &= \arg \min_{g \in \mathcal{G}} \sum_{k \in [K]} L(\mathcal{G}_k, g) \Pr(\mathcal{G}_k \mid X = x) \\ &= \arg \min_{g \in \mathcal{G}} \sum_{k \in [K]} L(\mathcal{G}_k, g) \Pr(\mathcal{G}_k \mid X = x)\end{aligned}\tag{3}$$

Con la *pérdida 0-1*⁴, la expresión se simplifica a

$$\begin{aligned}\hat{G}(x) &= \arg \min_{g \in \mathcal{G}} \sum_{k \in [K]} \mathbb{1}(\mathcal{G}_k \neq g) \Pr(\mathcal{G}_k \mid X = x) \\ &= \arg \min_{g \in \mathcal{G}} [1 - \Pr(g \mid X = x)] \\ &= \arg \max_{g \in \mathcal{G}} \Pr(g \mid X = x)\end{aligned}\tag{4}$$

Esta razonable solución se conoce como el *clasificador de Bayes*, y sugiere que clasifiquemos a cada observación según la clase modal condicional a su distribución conjunta $\Pr(G \mid X)$. Su error esperado de predicción EPE se conoce como la *tasa de Bayes*. Un aproximador directo de este resultado es el clasificador de k vecinos más cercano⁵

Definición 3.1.1.2 (clasificador de k-vecinos-más-cercanos): Sean $x^{(1)}, \dots, x^{(k)}$ los k vecinos más cercanos a x , y $\mathcal{G}^{(1)}, \dots, \mathcal{G}^{(k)}$ sus respectivas clases. El clasificador de k-vecinos-más-cercanos le asignará a x la clase más frecuente entre $\mathcal{G}^{(1)}, \dots, \mathcal{G}^{(k)}$. Más formalmente:

$$\begin{aligned}\hat{G}_{\text{kNN}}(x) &= \mathcal{G}_{\text{kNN}} = \arg \max_{g \in \mathcal{G}} \sum_{i \in [k]} \mathbb{1}(\mathcal{G}^{(i)} = g) \\ \Leftrightarrow \#\{i : \mathcal{G}^{(i)} = \mathcal{G}_{\text{kNN}}, i \in [k]\} &= \max_{g \in \mathcal{G}} \#\{i : \mathcal{G}^{(i)} = g, i \in [k]\}\end{aligned}\tag{5}$$

3.1.2 Clasificador de Bayes empírico

La *Regla de Bayes*,

$$\Pr(G \mid X) = \frac{\Pr(X \mid G) \times \Pr(G)}{\Pr(X)}\tag{6}$$

nos sugiere una reescritura de \hat{G} :

$$\begin{aligned}\hat{G}(x) &= g = \arg \max_{g \in \mathcal{G}} \Pr(g \mid X = x) \\ \Leftrightarrow \Pr(g \mid X = x) &= \max_{g \in \mathcal{G}} \Pr(g \mid X = x) \\ \Leftrightarrow \Pr(g \mid X = x) &= \max_{g \in \mathcal{G}} \Pr(X = x \mid g) \times \Pr(g) \\ \Leftrightarrow \Pr(\mathcal{G}_k \mid X = x) &= \max_{k \in [K]} \Pr(X = x \mid \mathcal{G}_k) \times \Pr(\mathcal{G}_k)\end{aligned}\tag{7}$$

⁴que no es otra cosa que la función indicadora de un error en la predicción,
 $\mathbf{01}(\hat{G}, G) = \mathbb{1}(\hat{G} \neq G)$

⁵*k-nearest-neighbors classifier*

A las probabilidades «incondicionales» de clase $\Pr(\mathcal{G}_k)$ se las suele llamar su «distribución a priori», y notarlas por $\pi = (\pi_1, \dots, \pi_K)^T$, con $\pi_k = \Pr(\mathcal{G}_k) \forall k \in [K]$, $\sum \pi_k = 1$. Una aproximación razonable, si es que el conjunto de entrenamiento se obtuvo por muestreo aleatorio simple⁶, es tomar las proporciones muestrales

$$\begin{aligned} \forall k \in [K], \quad \hat{\pi}_k &= N^{-1} \sum_{i \in [N]} \mathbb{1}(g_i = \mathcal{G}_k) \\ &= \frac{\#\{g_i : g_i = \mathcal{G}_k, i \in [N]\}}{N} \end{aligned} \tag{8}$$

Resta hallar una aproximación $\widehat{\Pr}(X = x | \mathcal{G}_k)$ a las probabilidades condicionales $X | \mathcal{G}_k$ para cada clase.

3.2 Estimación de densidad por núcleos

De conocer las K densidades $f_{X|\mathcal{G}_k}$, el cómputo de las mentadas probabilidades es directo. Tal vez la metodología más estudiada a tales fines es la *estimación de densidad por núcleos*, comprensivamente reseñada en (Hastie, Tibshirani y Friedman, 2009, §6.6). Al estimador resultante, sobre todo en el caso unidimensional, se lo conoce con el nombre de Parzen-Rosenblatt, por sus contribuciones fundacionales en el área (Rosenblatt, 1956; Parzen, 1962).

3.2.0.1 Estimación unidimensional

Para fijar ideas, asumamos que $X \in \mathbb{R}$ y consideremos la estimación de densidad en una única clase para la que contamos con N ejemplos $\{x_1, \dots, x_N\}$. Una aproximación \hat{f} directa sería

$$\hat{f}(x_0) = \frac{\#\{x_i \in \mathcal{N}(x_0)\}}{N \times h} \tag{9}$$

donde \mathcal{N} es un vecindario métrico de x_0 de diámetro h .

Esta estimación es irregular, con saltos discretos en el numerador, por lo que se prefiere el estimador «suavizado por núcleos» de Parzen-Rosenblatt. Pero primero: ¿qué es un núcleo?

Definición 3.2.0.1.1 (función núcleo o *kernel*):

Se dice que $K(x) : \mathbb{R} \rightarrow \mathbb{R}$ es una *función núcleo* si cumple que

1. toma valores reales no negativos: $K(u) \geq 0$,
2. está normalizada: $\int K(u) du = 1$,
3. es simétrica: $K(u) = K(-u)$ y
4. alcanza su máximo en el centro: $\max_u K(u) = K(0)$

⁶simple random sampling, o s.r.s.

Observación: Todas las funciones de densidad simétricas centradas en 0 son núcleos; en particular, la densidad «normal estándar» $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ lo es.

Observación: Si $K(u)$ es un núcleo, entonces $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$ también lo es.

Ahora sí estamos en condiciones de enunciar el

Definición 3.2.0.1.2 (estimador de densidad por núcleos):

Sea $\mathbf{x} = (x_1, \dots, x_N)^T$ una muestra i.i.d. de cierta variable aleatoria escalar $X \in \mathbb{R}$ con función de densidad f . Su estimador de densidad por núcleos, o estimador de Parzen-Rosenblatt es

$$\hat{f}(x_0) = \frac{1}{N} \sum_{i=1}^N K(x_0, x_i) \quad (10)$$

donde K es una [Definición 3.2.0.1.1](#)

Observación: La densidad de la distribución uniforme centrada en 0 de diámetro 1, $U(x) = \mathbb{1}\left(\frac{1}{2} < x \leq \frac{1}{2}\right)$ es un núcleo. Luego,

$$U_h(x) = \frac{1}{h} \mathbb{1}\left(-\frac{h}{2} < x < \frac{h}{2}\right) \quad (11)$$

también es un núcleo válido, y por ende el estimador de Ecuación 9 es un caso particular del estimador de [Definición 3.2.0.1.2](#):

$$\begin{aligned} \hat{f}(x_0) &= \frac{\#\{x_i \in \mathcal{N}(x_0)\}}{N \times h} \\ &= \frac{1}{N} \sum_{i \in [N]} \frac{1}{h} U\left(\frac{x_i - x_0}{h}\right) \\ &= \frac{1}{N} \sum_{i=1}^N U_h(x_i - x_0) \end{aligned} \quad (12)$$

3.2.1 Clasificador de densidad por núcleos

Si $\hat{f}_k, k \in [K]$ son estimadores de densidad por núcleos⁷ de cada una de las K densidades condicionales $X|\mathcal{G}_k$ según [Definición 3.2.0.1.2](#), podemos construir el siguiente clasificador

⁷KDEs ó *Kernel Density Estimators*, por sus siglas en inglés

Definición 3.2.1.1 (clasificador de densidad por núcleos): Sean $\hat{f}_1, \dots, \hat{f}_K$ estimadores de densidad por núcleos según [Definición 3.2.0.1.2](#). Sean además $\hat{\pi}_1, \dots, \hat{\pi}_K$ las estimaciones de la probabilidad incondicional de pertenecer a cada grupo $\mathcal{G}_1, \dots, \mathcal{G}_k$. Luego, la siguiente regla constituye un clasificador de densidad por núcleos:

$$\begin{aligned}\hat{G}_{\text{KD}}(x) = g &= \arg \max_{i \in [K]} \widehat{\Pr}(\mathcal{G}_i \mid X = x) \\ &= \arg \max_{i \in [K]} \widehat{\Pr}(X = x \mid \mathcal{G}_i) \times \widehat{\Pr}(\mathcal{G}_i) \\ &= \arg \max_{i \in [K]} \hat{f}_i(x) \times \hat{\pi}_i\end{aligned}\quad (13)$$

3.2.2 Clasificadores duros y suaves

Un clasificador que asigna a cada observación *una clase* - la más probable, se suele llamar *clasificador duro*. Un clasificador que asigna a cada observación *una distribución de probabilidades de clase* $\hat{\gamma}$ ⁸ se suele llamar *clasificador blando*. Dado un clasificador blando \hat{G}_{Blando} , es trivial construir el clasificador duro asociado \hat{G}_{Duro} :

$$\hat{G}_{\text{Duro}}(x_0) = \arg \max_i \hat{G}_{\text{Blando}}(x_0) = \arg \max_i \hat{\gamma}_i \quad (14)$$

Observación: El clasificador de [Definición 3.2.1.1](#) es en realidad la versión dura de un clasificador blando $\hat{G}_{\text{KD}}(x) = \hat{\gamma}$, donde

$$\hat{\gamma}_i = \frac{\hat{f}_i(x) \times \hat{\pi}_i}{\sum_{i \in [K]} \hat{f}_i(x) \times \hat{\pi}_i} \quad (15)$$

Observación: Algunos clasificadores sólo pueden ser duros, como $\hat{G}_{\text{1-NN}}$, el clasificador de [Definición 3.1.1.2](#) con $k = 1$.

Dos clasificadores *blandos* pueden tener la misma pérdida 0 – 1, pero «pintar» dos panoramas muy distintos respecto a cuán «seguros» están de cierta clasificación. Por caso,

$$\begin{aligned}\hat{G}_{\text{C(onfiado)}}(x_0) : \widehat{\Pr}(\mathcal{G}_i \mid X = x_0) &= \begin{cases} 1 - \varepsilon \times (K - 1) & \text{si } i = 1 \\ \varepsilon & \text{si } i \neq 1 \end{cases} \\ \hat{G}_{\text{D(udoso)}}(x_0) : \widehat{\Pr}(\mathcal{G}_i \mid X = x_0) &= \begin{cases} \frac{1}{K} + \varepsilon \times (K - 1) & \text{si } i = 1 \\ \frac{1}{K} - \varepsilon & \text{si } i \neq 1 \end{cases}\end{aligned}\quad (16)$$

\hat{G}_C está «casi seguro» de que la clase correcta es \mathcal{G}_1 , mientras que \hat{G}_D está prácticamente indeciso entre todas las clases. Para el entrenamiento y análisis de clasificadores blandos como el de densidad por núcleos, será

⁸Todas las restricciones habituales aplican: dado $\hat{\gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_K)^T$, todas sus componentes deben pertenecer al intervalo $[0, 1]$ y su suma ser exactamente 1.

relevante encontrar funciones de pérdida que recompensen y penalicen adecuadamente esta capacidad.

3.3 Estimación de densidad multivariada

3.3.1 Naive Bayes

Una manera «ingenua» de adaptar el procedimiento de estimación de densidad ya mencionado a X multivariadas, consiste en sostener el desde-luego-falso-pero-útil supuesto de que sus componentes X_1, \dots, X_p son independientes entre sí. De este modo, la estimación de densidad conjunta se reduce a la estimación de p densidades marginales univariadas. Dada cierta clase j , podemos escribir la densidad condicional $X|j$ como

$$f_j(X) = \prod_{k=1}^p f_{jk}(X_k) \quad (17)$$

A este procedimiento se lo conoce como «Naive Bayes», y a pesar de su aparente ingenuidad es competitivo contra algoritmos mucho más sofisticados en un amplio rango de tareas. En términos de cómputo, permite resolver la estimación con $K \times p$ KDE univariados. Además, permite que en X se combinen variables cuantitativas y cualitativas: basta con reemplazar la estimación de densidad para los X_k cualitativos por su correspondiente histograma.

3.3.2 KDE multivariado

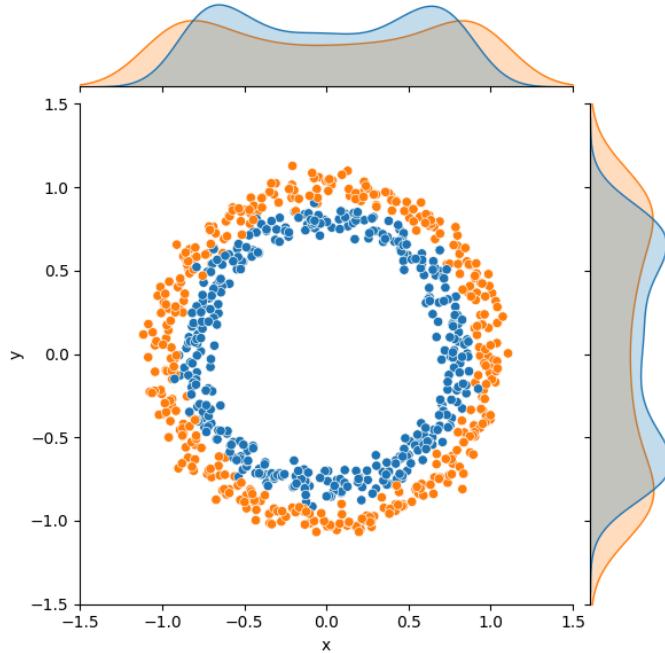


Figura 1: Dos círculos concéntricos y sus KDE marginales por clase: a pesar de que la frontera entre ambos grupos de puntos es muy clara, es casi imposible distinguirlas a partir de sus densidades marginales.

En casos como este, el procedimiento de Naive Bayes falla miserablemente, y será necesario adaptar el procedimiento de KDE unidimensional a $d \geq 2$ sin basarnos en el supuesto de independencia de las X_1, \dots, X_k . A lo largo de las cuatro décadas posteriores a las publicaciones de Parzen y Rosenblatt, el estudio de los estimadores de densidad por núcleos avanzó considerablemente, de manera que ya para mediados de los “90 existen minuciosos libros de referencia como «Kernel Smoothing» (Wand y Jones, 1995), que seguiremos en la presente sección

Definición 3.3.2.1 (KDE multivariada, (Wand y Jones, 1995, §4)):
En su forma más general, estimador de densidad por núcleos d – variado es

$$\hat{f}(x; \mathbf{H}) = N^{-1} \sum_{i=1}^N K_{\mathbf{H}}(x - x_i) \quad (18)$$

donde

- $\mathbf{H} \in \mathbb{R}^{d \times d}$ es una matriz simétrica def. pos. análoga a la ventana $h \in \mathbb{R}$ para $d = 1$,
- $K_{\mathbf{H}}(t) = |\det \mathbf{H}|^{-\frac{1}{2}} K(\mathbf{H}^{-\frac{1}{2}} t)$
- K es una función núcleo d -variada tal que $\int K(\mathbf{x}) d\mathbf{x} = 1$

Típicamente, K es la densidad normal multivariada

$$\Phi(x) : \mathbb{R}^d \rightarrow \mathbb{R} = (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right) \quad (19)$$

3.3.3 La elección de \mathbf{H}

Sean las clases de matrices $\mathbb{R}^{d \times d}$...

- \mathcal{F} , de matrices simétricas definidas positivas,
 - \mathcal{D} , de matrices diagonales definidas positivas ($\mathcal{D} \subseteq \mathcal{F}$) y
 - \mathcal{S} , de múltiplos escalares de la identidad: $\mathcal{S} = \{h^2 \mathbf{I} : h > 0\} \subseteq \mathcal{D}$
- Aún tomando una única \mathbf{H} para *toda* la muestra, $\mathbf{H} \in \dots$, la elección de \mathbf{H} en dimensión d requiere definir...
- $\binom{d}{2} = (d^2 - d)/2$ parámetros de ventana si $\mathbf{H} \in \mathcal{F}$,
 - d parámetros si $\mathbf{H} \in \mathcal{D}$ y
 - un único parámetro h si $\mathbf{H} = h^2 \mathbf{I}$.

La evaluación de la conveniencia relativa de cada parametrización se vuelve muy compleja, muy rápido. (Wand y Jones, 1993) proveen un análisis detallado para el caso $d = 2$, y concluyen que aunque cada caso amerita su propio estudio, $\mathbf{H} \in \mathcal{D}$ suele ser «adecuado». Sin embargo, este no es un gran consuelo para valores de d verdaderamente altos, en cuyo caso existe aún un problema más fundamental.

3.3.4 La maldición de la dimensionalidad

Uno estaría perdonado por suponer que el problema de estimar densidades en alta dimensión se resuelve con una buena elección de \mathbf{H} , y una muestra «lo suficientemente grande». Considérese, sin embargo, el siguiente ejercicio, adaptado de para ilustrar ese «suficientemente grande»:

Sean $X_i \stackrel{\text{iid}}{\sim} \text{Uniforme}([-1, 1]^d)$, $i \in [N]$, y consideremos la estimación de la densidad en el origen, $\hat{f}(\mathbf{0})$. Suponga que el núcleo $K_{\mathbf{H}}$ es un «núcleo producto» basado en la distribución univariada Uniforme($-1, 1$), y $\mathbf{H} = h^2 \mathbf{I}$. Derive una expresión para la proporción esperada de puntos incluidos dentro del soporte del núcleo $K_{\mathbf{H}}$ para h, d . arbitrarios.

— adaptado de (Wand y Jones, 1995, §4.9 ej 4.1)

El «núcleo producto» multivariado basado en la ley Uniforme($-1, 1$) evaluado alrededor del origen es:

$$K(x - 0) = K(x) = \prod_{i=1}^d \mathbb{1}(-1 \leq x_i \leq 1) = \mathbb{1}\left(\bigcap_{i=1}^d |x_i| \leq 1\right) \quad (20)$$

De la [Definición 3.3.2.1](#) y el hecho de que $\det \mathbf{H} = h^{2d}$; $\mathbf{H}^{-\frac{1}{2}} = h^{-1} \mathbf{I}$, se sigue que

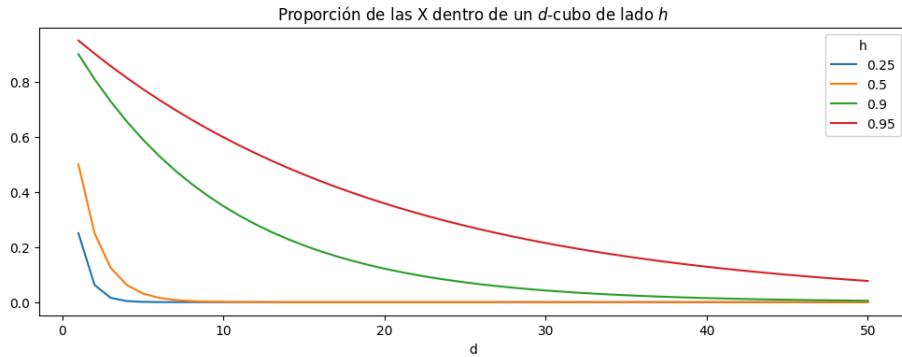
$$\begin{aligned} K_{\mathbf{H}}(x) &= |h^{2d}|^{-\frac{1}{2}} K(h^{-1} \mathbf{I} x) = h^{-d} K\left(\frac{x}{h}\right) \\ &= h^{-d} \mathbb{1}\left(\bigcap_{i=1}^d \left|\frac{x_i}{h}\right| \leq 1\right) = h^{-d} \mathbb{1}(\bigcap_{i=1}^d |x_i| \leq h) \quad (21) \\ &= h^{-d} \mathbb{1}(x \in [-h, h]^d) \end{aligned}$$

De modo que $\text{sop } K_{\mathbf{H}} = [-h, h]^d$, y ahora nos resta encontrar la esperanza. Como las componentes de una ley uniforme multivariada son independientes entre sí,

$$\begin{aligned} \Pr(X \in [-h, h]^d) &= \prod_{i=1}^d \Pr(X_i \in [-h, h]) \\ &= \Pr(-h \leq X_1 \leq h)^d \quad (22) \\ &= \left[\frac{h - (-h)}{1 - (-1)}\right]^d = h^d \quad \square \end{aligned}$$

Para $h = 0.5, d = 20$, $\Pr(X \in [-0.5, 0.5]^{20}) = 0.5^{-20} \approx 0.00000095$, ¡menos de uno en un millón! En general, la caída es muy rápida, aún para valores altos de h . Si X representa un segundo de audio respeta el estandar *mínimo* de llamadas telefónicas⁹

⁹De Wikipedia: La tasa [DS0](#), o *Digital Signal 0*, fue introducida para transportar una sola llamada de voz «digitizada». La típica llamada de audio se digitiza a 8 kHz, o a razón de 8.000 veces por segundo. se



tiene $d = 8000$. En tal espacio ambiente, aún con $h = 0.999$, $\Pr(\cdot) \approx 0.000334$, o 1:3.000.

3.3.5 La hipótesis de la variedad («manifold hypothesis»)

Ahora, si el espacio está *tan*, pero *tan* vacío en alta dimensión, ¿cómo es que el aprendizaje supervisado *sirve de algo*? La reciente explosión en capacidades y herramientas de procesamiento (¡y generación!) de formatos de altísima dimensión¹⁰ pareciera ser prueba fehaciente de que la tan mentada *maldición de la dimensionalidad* no es más que un cuento de viejas.

Pues bien, el ejemplo del segundo de audio antedicho *es* sesgado, ya que simplemente no es cierto que si X representa 1s de voz humana, su ley sea uniforme 8000 dimensiones¹¹: si uno muestreara un segundo de audio siguiendo cualquier distribución en la que muestras consecutivas no tengan ninguna correlación, obtiene *ruido blanco*. La voz humana, por su parte, tiene *estructura*, y por ende correlación instante-a-instante. Cada voz tiene un *timbre* característico, y las palabras enunciadas posibles están ceñidas por la *estructura fonológica* de la lengua locutada.

Sin precisar detalles, podríamos postular que las realizaciones de la variable de interés X (el habla), que registramos en un soporte $\mathcal{S} \subseteq \mathbb{R}^d$ de alta dimensión, en realidad se concentran en cierta *variedad*¹² $\mathcal{M} \subseteq \mathcal{S}$ potencialmente de mucha menor dimensión $\dim(M) = d_{\mathcal{M}} \ll d$, en la que noción de distancia entre observaciones aún conserva significado. A tal postulado se lo conoce como «la hipótesis de la variedad», o *manifold hypothesis*.¹³

¹⁰audio, video, texto y data genómica por citar sólo algunos

¹¹El audio se digitaliza usando 8 bits para cada muestra, así que más precisamente, sop $X = [2^8]^{8000}$ o 64 kbps, kilobits-por-segundo.

¹²Término que ya precisaremos. Por ahora, \mathcal{M} es el *subespacio de realizaciones posibles* de X

¹³Para el lector curioso: (Rifai *et al.*, 2011) ofrece un desglose de la hipótesis de la variedad en tres aspectos complementarios, de los cuales el aquí presentado sería el segundo, la «hipótesis de la variedad no-supervisada». El tercero, «la hipótesis de

La hipótesis de la variedad no es exactamente una hipótesis contrastable en el sentido tradicional del método científico; de hecho, ni siquiera resulta obvio que de existir, sean bien definibles las variedades en las que existen los elementos del mundo real: un dígito manuscrito, el canto de un pájaro, o una flor. Y de existir, es de esperar que sean altamente no-lineales.



Figura 2: Ejemplos de variedades en el mundo físico: tanto la hoja de un árbol como una bandera flameando al viento tienen dimensión intrínseca

$$d_{\mathcal{M}} = 2, \text{ están embedidas en } \mathbb{R}^3, \text{ y son definitivamente no-lineales.}$$

Más bien, corresponde entenderla como un modelo mental, que nos permite aventurar ciertas líneas prácticas de trabajo en alta dimensión¹⁴. Pero antes de profundizar en esta línea, debemos platearnos algunas preguntas básicas:

¿Qué es, exactamente, una variedad?

¿Es posible construir un KDE con soporte en cierta variedad *conocida*?

¿Sirve de algo todo esto si *no conocemos* la variedad en cuestión?

3.4 Variedades de Riemann

Adelantando la respuesta a la segunda pregunta, resulta ser que si el soporte de X es una «variedad de Riemann», bajo ciertas condiciones razonables sí es posible estimar su densidad por núcleos en la variedad (Pelletier, 2005).

la variedad para clasificación», dice que «puntos de distintas clases se concentrarán sobre variedades disjuntas separadas por regiones de muy baja densidad, lo asumimos implícitamente a la hora de construir un clasificador.

¹⁴TODO: (Gallese, 2003) : shared manifold hypothesis y (Bengio, 2019)

A continuación, damos un recorrido sumario e idiosincrático por ciertos conceptos básicos de topología y variedades que consideramos necesarios para motivar la definición de variedades Riemannianas, que de paso precisarán la respuesta a la primer pregunta - ¿qué es una variedad? - en el contexto que nos interesa. A tal fin, seguimos la exposición de la monografía *Estimación no paramétrica de la densidad en variedades Riemannianas* (Muñoz, 2011), que a su vez sigue, entre otros, el clásico *Introduction to Riemannian Manifolds* (Lee, 2018).

3.4.1 Variedades Diferenciables

Definición 3.4.1.1 (espacio topológico (TODO: ARROBA CITA WIKIPEDIA)):

Formalmente, se llama espacio topológico al par ordenado (X, T) formado por un conjunto X y una *topología* T sobre X , es decir una colección de subconjuntos de X que cumple las siguientes tres propiedades:

1. El conjunto vacío y X están en T : $\emptyset \in T, X \in T$
2. La intersección de cualquier subcolección *finita* de T está en T :

$$X \in T, Y \in T \Rightarrow X \cap Y \in T \quad (23)$$

La unión de *cualquier* subcolección de conjuntos de T está en T :

$$\forall S \subset T, \bigcup_{O \in S} O \in T \quad (24)$$

A los conjuntos pertenecientes a la topología T se les llama conjuntos abiertos o simplemente abiertos de (X, T) ; y a sus complementos en X , conjuntos cerrados.

Definición 3.4.1.2 (entorno (TODO arroba wikipedia)): Si (X, T) es un espacio topológico y p es un punto perteneciente a X , un *entorno*¹⁵ del punto p es un conjunto V en el que está contenido un conjunto abierto U que incluye al propio p : $p \in U \subseteq V$.

¹⁵También se los conoce como «vecindarios» - por *neighborhoods*, su nombre en inglés.

Definición 3.4.1.3 (espacio de Hausdorff (TODO: ARROBA CITA WIKIPEDIA)):

Sea (X, T) un espacio topológico. Se dice que dos puntos $p, q \in X$ cumplen la propiedad de Hausdorff si existen dos entornos U_p de p y U_q de q tales que $U_p \cap U_q = \emptyset$ (i.e., son disjuntos).

Se dice que un espacio topológico es un espacio de Hausdorff¹⁶ si todo par de puntos distintos del espacio verifican la propiedad de Hausdorff.

En términos coloquiales, un espacio de Hausdorff es aquél donde todos sus puntos están «bien separados».

Definición 3.4.1.4 (variedad topológica (Muñoz, 2011, Def. 3.1.1), (Lee, 2018, Apéndice A)): Una variedad topológica de dimensión $d \in \mathbb{N}$ es un espacio topológico (\mathcal{M}, T) de Hausdorff, de base numerable, que es localmente homeomorfo a \mathbb{R}^d . Es decir, para cada $p \in \mathcal{M}$ existe un abierto $U \in T$ y un abierto $A \subseteq \mathbb{R}^d$, tal que $p \in U$ (U es un entorno de p) y existe un homemorfismo $\varphi : U \rightarrow A$.

Observación (Sobre variedades con y sin frontera): Toda n – variedad¹⁷ tiene puntos interiores, pero algunas además tienen una *frontera*; esta frontera es a su vez una variedad *sin* frontera de dimensión $n - 1$. Por caso: un disco en el plano euclídeo \mathbb{R}^2 es una 2 – variedad *con* frontera, cuya frontera es una variedad de dimensión $2 - 1 = 1$ sin frontera: el círculo S^1 ; una pelota de tenis es una 3 – variedad con frontera dada por su superficie, la variedad sin frontera S^2 . De aquí en más, cuando hablamos de variedades topológicas, nos referiremos a variedades *sin* frontera.

En una variedad topológica, cobra sentido cierto concepto de cercanía - pero no necesariamente de *distancia*, y es posible definir funciones continuas y límites.

Un *homeomorfismo*¹⁸ es una función phi entre dos espacios topológicos si es biyectiva y tanto ella como su inversa son continuas. El par ordenado (U, φ) es una *carta*¹⁹ alrededor de p .

A un conjunto numerable de tales cartas que cubran completamente la variedad se lo denomina «atlas». Simbólicamente, $\mathcal{A} = \{(U_\alpha, \varphi_\alpha) : \alpha \in \mathcal{I}\}$ es un atlas sí y sólo si $\mathcal{M} = \cup_\alpha U_\alpha$. Al conjunto de entornos $\{\mathbf{U}_\alpha\} = \{U_\alpha : (U_\alpha, \varphi_\alpha) \in \mathcal{A}\}$ que componen un atlas se lo denomina «cobertura» de \mathcal{M} .

¹⁶O que verifica la propiedad de Hausdorff, o que es separado o que es \mathbf{T}_2

¹⁷i.e. variedad de dimensión n

¹⁸del griego *homo-*: igual, *-morpho*: forma; de igual forma

¹⁹*chart* en inglés

Cuando un homeomorfismo - y su inversa - es r –veces diferenciable, se le llama C^r -difeomorfismo, o simplemente difeomorfismo²⁰. En particular, un C^∞ –difeomorfismo es un difeomorfismo *suave*.

Definición 3.4.1.5:

Sean (\mathcal{M}, T) una variedad topológica de dimensión d y sean $(U, \varphi), (V, \psi)$ dos cartas. Diremos que son *suavemente compatibles*²¹ si $U \cap V = \emptyset$ o bien si la función cambio de coordenadas restringida a $U \cap V$ es un difeomorfismo.

La compatibilidad requiere que la transición entre mapas no sea sólo continua, sino también *suave*. El motivo de esta condición es asegurar que el concepto de *suavidad* esté bien definido en toda la variedad \mathcal{M} , independientemente de qué carta se use: si una función es diferenciable vista a través de una carta, también lo será al analizarla desde cualquier carta compatible.

Definición 3.4.1.6 (estructura diferenciable (Muñoz, 2011, Def. 3.1.3)): Un atlas $\mathcal{A} = \{(U_\alpha, \varphi_\alpha) : \alpha \in \mathcal{I}\}$ es diferenciable si sus cartas son compatibles entre sí. Si un atlas diferenciable \mathcal{D} es *maximal* lo llamaremos una *estructura diferenciable de la variedad* \mathcal{M} . Con maximal queremos decir lo siguiente: Si (U, φ) es una carta de \mathcal{M} que es compatible con todas las cartas de \mathcal{D} , entonces $(U, \varphi) \in \mathcal{D}$ ²²

Definición 3.4.1.7 (variedad diferenciable (Muñoz, 2011, Def. 3.1.4)): Una variedad diferenciable de dimensión d es una terna $(\mathcal{M}, \tau, \mathcal{D})$ donde (\mathcal{M}, τ) es una variedad topológica de dimensión d y \mathcal{D} una estructura diferenciable.

Una variedad diferenciable entonces, es aquella en la que la operación de diferenciación tiene sentido no sólo punto a punto, sino globalmente. Nótese que de no poder diferenciar, tampoco podremos tomar integrales, y no sólo la *estimación* de la densidad por núcleos sería imposible, sino que ni siquiera tendría sentido plantear una función densidad.

Sobre una variedad diferenciable, cobra sentido plantear el concepto de *métrica*. En particular, toda variedad diferenciable admite una «métrica de Riemann» (TODO arroba do carmo, Proposición 2.10).

²⁰Luego, un homeomorfismo es un C^0 –difeomorfismo

²¹*smoothly compatible* según (Lee, 2018, § «Smooth Manifolds and Smooth Maps»). (Muñoz, 2011) lo denomina *compatible* a secas.

²²i.e., no existe otro atlas diferenciable que contenga propiamente a \mathcal{D} , lo cual desambigua la referencia.

Definición 3.4.1.8 («métrica Riemanniana TODO at Do carmo Def 2.1»): Sea $T_p\mathcal{M}$ el *espacio tangente* a un punto $p \in \mathcal{M}$. Una métrica Riemanniana - o estructura Riemanniana - en una variedad diferenciable \mathcal{M} es una correspondencia que asocia a cada punto $p \in \mathcal{M}$ un producto interno $\langle \cdot, \cdot \rangle$ (i.e., una forma bilineal simétrica positiva definida) en el espacio tangente $T_p\mathcal{M}$ que «varía diferencialmente»²³ en el entorno de p .

A dicho producto interno se lo denomina g_p e induce naturalmente una norma: $\|v\|_p = \sqrt{g_p(v, v)} = \sqrt{\langle v, v \rangle}$. Decimos entonces que g_p es una métrica Riemanniana y el par (\mathcal{M}, g) es una variedad de Riemann.

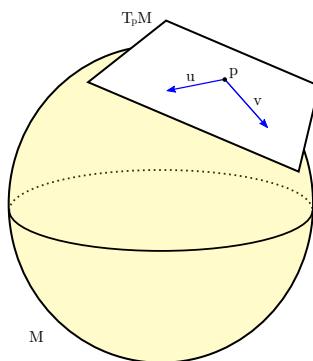


Figura 3: Espacio tangente $T_p\mathcal{M}$ a una esfera $\mathcal{M} = S^2$ por p . Nótese que el espacio tangente varía con p , pero siempre mantiene la misma dimensión ($d = 2$) que \mathcal{M}

Observación (según TODO at do carmo Prop. 2.10): Toda variedad diferenciable admite una métrica Riemanniana, que se puede construir componiendo las métricas Riemannianas locales a cada carta de su estructura diferenciable según la «partición de la unidad»²⁴ $\{f\} = \{f_\alpha : \alpha \in \mathcal{I}\}$ subordinada a su cobertura.

Es claro que podemos definir una métrica Riemanniana $\langle \cdot, \cdot \rangle^\alpha$ en cada V_α : la métrica inducida por el sistema de coordenadas locales. Sea entonces el conjunto:

²³para el lector curioso, do Carmo Def 2.1 define precisamente el sentido de esta expresión

²⁴La definición formal de «partición de la unidad» la da - sin prueba de existencia - TODO at do carmo §0.5, p. 30. Intuitivamente, da una base funcional de \mathcal{M} , en la que a cada entorno de la cobertura de \mathcal{M} se le asigna una función f_α de manera que $\sum_\alpha f_\alpha(p) = 1 \forall p \in \mathcal{M}$. para es una técnica que pondera con pesos que suman 1 las métricas locales a cada carta para obtener un resultado global coherente

$$\langle u, v \rangle_p = \sum_{\alpha} f_{\alpha}(p) \langle u, v \rangle_p^{\alpha} \quad \forall p \in \mathcal{M}, u, v \in T_p \mathcal{M} \quad (25)$$

es posible verificar que esta construcción define una métrica Riemanniana en todo \mathcal{M} .

Observación: Cuando $\mathcal{M} = \mathbb{R}^d$, el espacio es constante e idéntico a la variedad: $\forall p \in \mathbb{R}^d, T_p \mathbb{R}^d = \mathbb{R}^d$. La base canónica de $T_p \mathbb{R}^d = \mathbb{R}^d$ formada por las columnas de \mathbf{I}_d es una matriz positiva definida que da lugar al pructo interno «clásico» $\langle u, v \rangle = u^T \mathbf{I}_d v = \sum_{i=1}^d u_i v_i$ es una métrica Riemanniana que induce la norma euclídea $\|v\| = \sqrt{v^T v}$ y la distancia $d(x, y) = \|x - y\|$.

3.4.1.1 Geodésicas y mapa exponencial

Dado este andamiaje, podemos reconstruir algunos conceptos básicos, como longitud, distancia y geodésica. Sea $\gamma : [a, b] \rightarrow \mathcal{M}$ una curva diferenciable en \mathcal{M} , y γ' su derivada. La longitud de γ está dada por

$$L(\gamma) = \int_a^b \|\gamma'(t)\| dt = \int_a^b \sqrt{g_{\gamma(t)}(\gamma'(t), \gamma'(t))} dt \quad (26)$$

Definición 3.4.1.1.1 (distancia en variedades de Riemann): Sea (\mathcal{M}, g) una variedad de Riemann, y $p, q \in \mathcal{M}$ dos puntos. Definimos la distancia entre ellos inducida por la métrica g como

$$d_g(p, q) = \inf_{\gamma} \{L(\gamma) : \gamma : [0, 1] \rightarrow \mathcal{M}, \gamma(0) = p, \gamma(1) = q\} \quad (27)$$

A la curva γ que minimiza la distancia entre p y q se la denomina *geodésica*, una generalización de la «línea recta» en la geometría euclídea.

En efecto, considérese la siguiente analogía: en la física clásica, un objeto que no es sujeto a ninguna fuerza (no recibe *aceleración* alguna), estará o quieto (con velocidad nula) o en movimiento rectilíneo uniforme («MRU»). En variedades diferenciables, la geodésicas son exactamente eso: curvas parametrizables sin aceleración ($\gamma''(t) = 0 \forall t$). En esta línea «intuitiva», lo que sigue es una adaptación de «El flujo geodésico» TODO at docarmo §3.2.

Sea $\gamma : [0, 1] \rightarrow \mathcal{M}, \gamma(0) = p, \gamma(1) = q$ una curva parametrizable. Su derivada en el origen - su *velocidad inicial* - $\gamma'(0)$ es necesariamente tangente a $\gamma(0) = p \in \mathcal{M}$, o sea que $\gamma'(0) \in T_p \mathcal{M}$: el espacio tangente $T_p \mathcal{M}$ contiene todas las *velocidades* posibles desde p . Dada una velocidad $v \in T_p \mathcal{M}$, podemos descomponerla en su *magnitud* $\|v\|$ y su *dirección* $\frac{v}{\|v\|}$. Como la geodésica es una curva sin aceleración, $\gamma''(t) = 0 \forall t \in [0, 1]$, y luego $\gamma'(t) = \gamma'(0) = v \in T_p \mathcal{M} \forall t \in [0, 1]$. La geodésica de p a q es la única curva $\gamma : [0, 1] \rightarrow \mathcal{M}, \gamma(0) = p$ con velocidad inicial $\gamma'(0) = v \in T_p \mathcal{M}$, de modo que $L(\gamma) = \|v\| = d_g(p, q)$ y luego de «una unidad de tiempo», $\gamma(1) = q$.

Esta relación, entre vectores de $T_p\mathcal{M}$ y geodésicas de \mathcal{M} con origen en p , nos permite relacionar una «bola» en $T_p\mathcal{M}$ con su análogo en \mathcal{M} .

Definición 3.4.1.1.2 (mapa exponencial): Sean $p \in \mathcal{M}, v \in T_p\mathcal{M}$.

Se conoce como *mapa exponencial* a la función

$$\exp_p(v) : T_p\mathcal{M} \rightarrow \mathcal{M} = \gamma_{p,v}(1) \quad (28)$$

donde $\gamma_{p,v}(t)$ es la única geodésica que en el instante $t = 0$ pasa por p con velocidad v .

Definición 3.4.1.1.3 (bola normal): Sea $B_\varepsilon(x) \subset \mathbb{R}^d$ la bola cerrada de radio ε centrada en x :

$$B_\varepsilon(x) = \{y \in \mathbb{R}^d : d_g(x, y) = \|x - y\| \leq \varepsilon\} \quad (29)$$

Si \exp_p es un difeomorfismo en un vecindario (entorno) V del origen en $T_p\mathcal{M}$, su imagen $\overline{U} = \exp_p(V)$ es un «vecindario normal» de p .

Si $B_\varepsilon(0)$ es tal que $\overline{B_\varepsilon(0)} \subset V$, llamamos a $\exp_p B_\varepsilon(0) = B_\varepsilon(p)$ la *bola normal* – o «bola geodésica» - con centro p y radio ε .

La frontera de $B_\varepsilon(p)$ es una «subvariedad» de \mathcal{M} ortogonal a las geodésicas que irradian desde p . Una concepción intuitiva de qué es una bola normal, es «un entorno de p en el que las geodésicas que pasan por p son minimizadoras de distancias». El siguiente concepto es útil para entender «cuán lejos vale» la aproximación local a un espacio euclídeo en la variedad.

Definición 3.4.1.1.4 (radio de inyectividad²⁵): Sea (\mathcal{M}, g) una d – variedad Riemanniana. Llamamos «radio de inyectividad en p » a

$$\text{iny}_p\mathcal{M} = \sup\{s \in \mathbb{R} > 0 : B_s(p) \text{ es una bola normal}\} \quad (31)$$

El ínfimo de los radios de inyectividad «puntuales», es el radio de inyectividad de la variedad \mathcal{M} .

$$\text{iny}\mathcal{M} = \inf_{p \in \mathcal{M}} \text{iny}_p\mathcal{M} \quad (32)$$

²⁵Basado en (Muñoz, 2011, Def. 3.3.16) Una definición a mi entender más esclarecedora se encuentra en TODO at do carmo, §13.2, *The cut locus*, que introducimos aquí informalmente. El *cut locus* o *ligne de partage* $C_m(p)$ - algo así como la línea de corte - de un punto p es la unión de todos los puntos de corte: los puntos a lo largo de las geodésicas que irradian de p donde éstas dejan de ser minimizadoras de distancia. El ínfimo de la distancia entre p y su línea de corte, es el radio de inyectividad de \mathcal{M} en p , de modo podemos escribir

$$\text{iny } \mathcal{M} = \inf_{p \in \mathcal{M}} d(p, C_m(p)) \quad (30)$$

donde la distancia de un punto a una variedad es el ínfimo de la distancia a todos los puntos de la variedad.

Observación: Si $\mathcal{M} = \mathbb{R}^d$ con la métrica canónica entonces $\text{iny } \mathcal{M} = \infty$. Si $\mathcal{M} = \mathbb{R}^d - \{p\}$, con la métrica usual, entonces existe un punto arbitrariamente cerca de p en el que la geodésica que irradia en dirección a p se corta inmediatamente: entonces el radio de inyectividad es cero. Si $\mathcal{M} = S^1$ con radio unitario y la métrica inducida de \mathbb{R}^2 , el radio de inyectividad es π , puesto que si tomamos «el polo norte» p_N como origen de un espacio tangente $T_{p_N}S^1$, todas (las dos) geodésicas que salen de él llegan al polo sur p_S «al mismo tiempo» π , y perdemos la inyectividad.

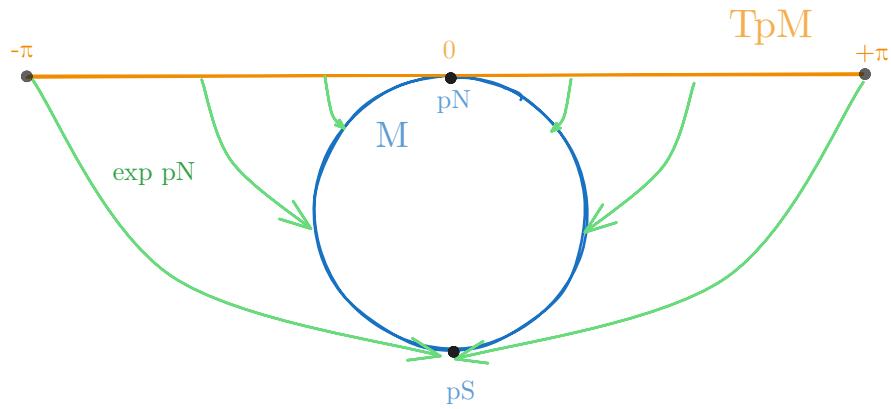


Figura 4: Espacio tangente y mapa exponencial para $p_N \in S^1$. Nótese que $\text{iny } S^1 = \pi$. Prolongando una geodésica $\gamma(t)$ más allá de $t = \pi$, ya no se obtiene un camino mínimo, pues hubiese sido más corto llegar por $-\gamma(s)$, $s = t \bmod \pi$.

Agregamos una última definición para restringir la clase de variedades de Riemann que nos interesarán:

Definición 3.4.1.1.5 (variedad compacta): Decimos que una variedad es *acotada* cuando $\sup_{(p,q) \in \mathcal{M}^2} d_g(p,q) = \bar{d} < \infty$ - no posee elementos distanciados infinitamente entre sí. Una variedad que incluya todos sus «puntos límite» es una variedad *cerrada*. Una variedad cerrada y acotada se denomina *compacta*.

Observación: Un círculo en el plano, $S^1 \subset \mathbb{R}^2 = \{(x,y) : x^2 + y^2 = 1\}$ es una variedad compacta: es acotada - ninguna distancia es mayor a medio gran círculo, π - y cerrada. \mathbb{R}^2 es una variedad cerrada pero no acotada. El «disco sin borde» $\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ es acotado pero no cerrado - pues no incluye su borde S^1 . El «cilindro infinito» $\{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 < 1\}$ no es ni acotado ni compacto.

Ahora sí, hemos arribado a un objeto lo suficientemente «bien portado» para soportar funciones diferenciables, una noción de distancia y todo

aquellos que precisamos para definir elementos aleatorios: la variedad de Riemann compacta sin frontera. Cuando hablemos de una variedad de Riemann sin calificarla, nos referiremos a ésta.

3.4.2 Probabilidad en Variedades

Hemos definido una clase bastante general de variedades - las variedades de Riemann - que podrán soportar funciones de densidad y sus estimaciones (Pelletier, 2005). Estos desarrollos relativamente modernos²⁶, no constituyen sin embargo el origen de la probabilidad en variedades. Mucho antes de su sistematización, ciertos casos particulares habían sido bien estudiados y allanaron el camino para el interés en variedades más generales. Probablemente la referencia más antigua a un elemento aleatorio en una variedad distinta a \mathbb{R}^d , se deba a Richard von Mises, en *Sobre la naturaleza entera del peso atómico y cuestiones relacionadas* (von Mises, 1918)²⁷. En él, von Mises se plantea la pregunta explícita de si los pesos atómicos - que empíricamente se observan siempre muy cercanos a la unidad para los elementos más livianos - son enteros con un cierto error de medición, y argumenta que para tal tratamiento, el «error gaussiano» clásico es inadecuado:

(dots) Pues no es evidente desde el principio que, por ejemplo, para un peso atómico de 35,46 (Cl), el error sea de +0,46 y no de -0,54: es muy posible que se logre una mejor concordancia con ciertos supuestos con la segunda determinación. A continuación, se desarrollan los elementos — esencialmente muy simples — de una «teoría del error cíclico», que se complementa con la teoría gaussiana o «lineal» y permite un tratamiento completamente inequívoco del problema de la «enteridad» y cuestiones similares.

— traducido de (von Mises, 1918)

²⁶del siglo XXI, al menos

²⁷«Über die “ganzzahligkeitwider” atomgewichte und verwandte fragen». en el original

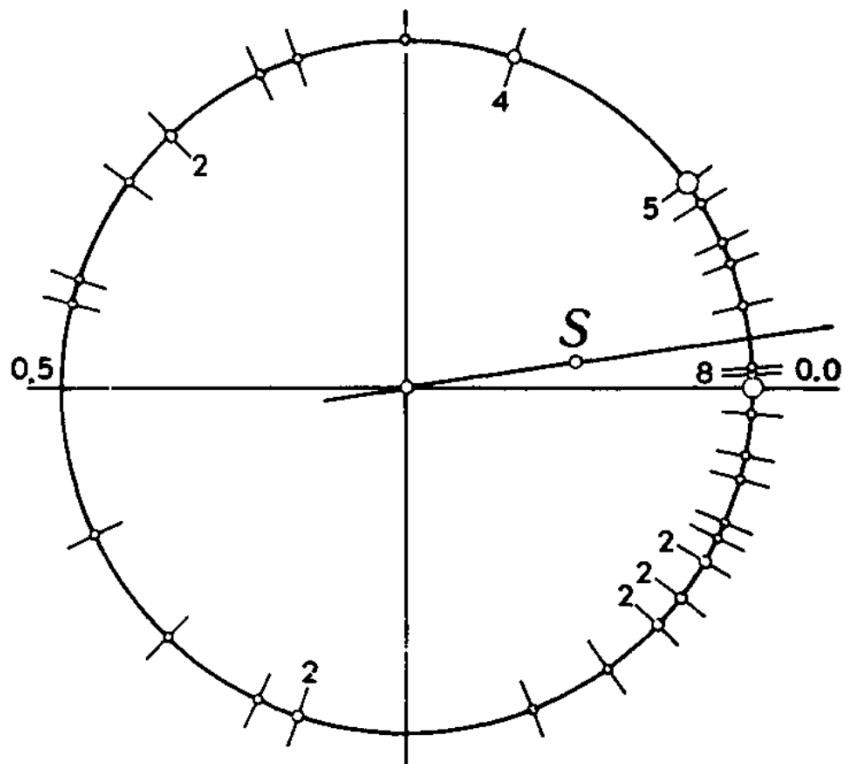


Fig. 6.

Figura 5: Pretendido «error» - diferencia módulo 1 - de los pesos atómicos medidos para ciertos elementos, sobre S^1 . Nótese como la mayoría de las mediciones se agrupan en torno al 0.0.

Motivado también por un problema del mundo físico - las mediciones de posición en una esfera «clásica» $S^2 \subset \mathbb{R}^3$, Ronald Fisher escribe «Dispersiones en la esfera» (Fisher, 1957), donde desarrolla una forma de teoría que parece ser apropiada para mediciones de posición en una esfera²⁸ y los ilustra utilizando mediciones de la dirección de la magnetización remanente de flujos de lava directa e inversamente magnetizados en Islandia.

Dos décadas más tarde, los casos particulares de von Mises (S^1) y Fisher (S^2) estaban integrados en el caso más general S^n en lo que se conocería como «estadística direccional»²⁹. En 1975 se habla ya de *teoría de la distribución* para la distribución von Mises - Fisher (Mardia, 1975),

²⁸y como era de esperar del padre del test de hipótesis, también un test de significancia análogo al t de Student.

²⁹ya que la n -esfera S^n de radio 1 con centro en 0 contiene exactamente a todos los vectores unitarios, i.e. a todas las *direcciones* posibles de un vector en su espacio ambiente \mathbb{R}^{n+1}

la «más importante en el análisis de datos direccionales»; a fines de los “90 Jupp y Mardia plantean «una visión unificada de la teoría de la estadística direccional» (Jupp y Mardia, 1989) , relacionándola con conceptos claves en el «caso euclídeo» como las familias exponenciales y el teorema central del límite, entre otros.

Aunque el caso particular de la n –esfera sí fue bien desarrollado a lo largo del siglo XX, el tratamiento más general de la estadística en variedades riemannianas conocidas pero arbitrarias aún no se hacía presente.

3.4.3 KDE en variedades de Riemann

Un trabajo sumamente interesante a principios del siglo XXI es el de Bruno Pelletier, que se propone una adaptación directa del estimador de densidad por núcleos de [Definición 3.3.2.1](#) en variedades de Riemann compactas sin frontera (Pelletier, 2005). Lo presentamos directamente y ampliamos los detalles a continuación

Definición 3.4.3.1 (KDE en variedades de Riemann (Pelletier, 2005, Ecuación 1)): Sean

- (\mathcal{M}, g) una variedad de Riemann compacta y sin frontera de dimensión d , y d_g la distancia de Riemann,
- K un *núcleo isotrópico* en \mathcal{M} soportado en la bola unitaria en \mathbb{R}^d
- dados $p, q \in \mathcal{M}$, $\theta_p(q)$ la *función de densidad de volumen en \mathcal{M}*
- Sea \mathbf{X} una muestra de N observaciones de una variable aleatoria X con densidad f soportada en \mathcal{M}

Luego, el estimador de densidad por núcleos para X es la $\hat{f} : \mathcal{M} \rightarrow \mathbb{R}$ que a cada $p \in \mathcal{M}$ le asocia el valor

$$\begin{aligned}\hat{f}(p) &= N^{-1} \sum_{i=1}^N K_h(p, X_i) \\ &= N^{-1} \sum_{i=1}^N \frac{1}{h^d} \frac{1}{\theta_{X_i}(p)} K\left(\frac{d_g(p, X_i)}{h}\right)\end{aligned}\tag{33}$$

con la restricción de que la ventana $h \leq h_0 \leq \text{iny } \mathcal{M}$, el *radio de inyectividad* de \mathcal{M} .³⁰ El autor prueba la convergencia en $L^2(\mathcal{M})$:

³⁰Esta restricción no es catastrófica. Para toda variedad compacta, el radio de inyectividad será estrictamente positivo (Muñoz, 2011, Prop. 3.3.18). Como además h es en realidad una sucesión $\{h_n\}_{n=1}^N$ decreciente como función del tamaño muestral, siempre existirá un cierto tamaño muestral a partir del cual $h_n < \text{iny } \mathcal{M}$.

Teorema 3.4.3.1 (convergencia de \hat{f} en L^2 (Pelletier, 2005, §3 Teorema 5)): Sea f una densidad de probabilidad dos veces diferenciable en \mathcal{M} con segunda derivada covariante acotada. Sea \hat{f}_n el estimador de densidad definido en [Definición 3.4.3.1](#) con ventana $h_n < h_0 < \text{iny } \mathcal{M}$. Luego, existe una constante C_f tal que

$$\mathbb{E} \|\hat{f}_n - f\|_{L^2(\mathcal{M})}^2 \leq C_f \left(\frac{1}{nh^d} + r^4 \right). \quad (34)$$

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

$$\mathbb{E} \|\hat{f}_n - f\|_{L^2(\mathcal{M})}^2 = O(n^{-\frac{4}{d+4}}) \quad (35)$$

Nótese que esta formulación revela una buena sugerencia de en qué orden comenzar la búsqueda de h . (Henry y Rodriguez, 2009, Teorema 3.2) prueba la consistencia fuerte de \hat{f} : bajo los mismos (Pelletier, 2005), obtienen que

$$\sup_{p \in \mathcal{M}} |\hat{f}_{n(p)} - f(p)| \xrightarrow{\text{c.s.}} 0 \quad (36)$$

[Definición 3.4.3.2](#) (núcleo isotrópico): Sea $K : \mathbb{R}_+ \rightarrow \mathbb{R}$ un mapa no-negativo tal que:

$$\int_{\mathbb{R}^d} K(\|x\|) d\lambda(x) = 1 \quad K \text{ es función de densidad en } \mathbb{R}^d$$

$$\int_{\mathbb{R}^d} xK(\|x\|) d\lambda(x) = 0 \quad \text{Si } Y \sim K, \mathbb{E}Y = 0$$

$$\int_{\mathbb{R}^d} \|x\|^2 K(\|x\|) d\lambda(x) < \infty \quad \text{Si } Y \sim K, \text{Var } Y = 0$$

$$\text{sop } K = [0, 1]$$

$$\sup_x K(x) = K(0) \quad K \text{ se maximiza en el origen}$$

Decimos entonces que el mapa $\mathbb{R}^d \ni x \rightarrow K(\|x\|) \in \mathbb{R}$ es un «núcleo isotrópico» en \mathbb{R}^d soportado en la bola unitaria.

Observación: Todo núcleo válido en [Definición 3.3.2.1](#) también es un núcleo isotrópico. A nuestros fines, continuaremos utilizando el núcleo normal.

Definición 3.4.3.3 (función de densidad de volumen TODO at besse 78 §6.2): Sean $p, q \in \mathcal{M}$; le llamaremos *función de densidad de volumen* en \mathcal{M} a $\theta_p(q)$ definida como

$$\theta_p(q) : q \mapsto \theta_p(q) = \frac{\mu_{\exp_p^* g}}{\mu_{g_p}}(\exp_p^{-1}(q)) \quad (37)$$

es decir, el cociente de la medida canónica de la métrica Riemanniana \exp_p^* sobre $T_p\mathcal{M}$ (la métrica *pullback* que resulta de transferir g de \mathcal{M} a $T_p\mathcal{M}$ a través del mapa exponencial \exp_p), por la medida de Lebesgue de la estructura euclídea en $T_p\mathcal{M}$.

Observación:

$\theta_p(q)$ está bien definida «cerca» de p : por ejemplo, es idénticamente igual a 1 en el entorno U localmente «plano» de p donde las geodésicas $\gamma \subset \mathcal{M}$ coinciden con sus representaciones en $T_p\mathcal{M}$, coinciden con su representación. Ciertamente está definida para todo q dentro del radio de inyectividad de p , $d_g(p, q) < \text{iny}_p \mathcal{M}$ ³¹. Con N «suficientemente grande», siempre podremos elegir $h_N < \text{iny}_p \mathcal{M}$ que mapee «suficientes» observaciones al soporte de K , $[0, 1]$ en las que el cálculo de $\theta_p(q)$ sea factible, y las más lejanas queden por fuera, de modo que su cálculo *no sea necesario*.

El mapa exponencial alrededor de p , $\exp_p : T_p\mathcal{M} \rightarrow \mathcal{M}$ es un difeomorfismo en cierta bola normal alrededor de p , así que admite una inversa continua y biyectiva al menos en tal bola; lo llamaremos $\exp_p^{-1} : \mathcal{M} \rightarrow T_p\mathcal{M}$. Así, $\exp_p^{-1}(q) \in T_p\mathcal{M}$ es la representación de q en las coordenadas localmente euclídeas del espacio tangente a p (o sencillamente «locales a p »). De esta cantidad $x = \exp_p^{-1}(q)$, queremos conocer el cociente entre dos medidas:

- la métrica *pullback* de g : la métrica inducida en $T_p\mathcal{M}$ por la métrica riemanniana g en \mathcal{M}
- la medida de lebesgue en la estructura euclídea de $T_p\mathcal{M}$.

En otras palabras, $\theta_p(q)$ representa cuánto se infla / encoge - el espacio en la variedad \mathcal{M} alrededor de p , relativo al volumen «natural» del espacio tangente. En general, su cómputo resulta sumamente complejo, salvo en casos particulares como las variedades «planas» o de curvatura constante. En un trabajo reciente, por ejemplo, se reseña:

Un problema restante a esta altura es el de entender cómo la *regularidad*³³ de \mathcal{M} afecta las tasas de convergencia de funciones

³¹su definición global es compleja y escapa al tema de esta monografía³².

³²Besse y Pelletier consideran factible extenderla a todo \mathcal{M} utilizando *campos de Jacobi* TODO besse pelletier

suaves (...). Luego, en el caso especial en que la dimensión de \mathcal{M} es conocida e igual a 1, podemos construir un estimador que alcanza la tasa [propuesta anteriormente]. Así, se establece que en dimensión 1 al menos, la regularidad de la variedad \mathcal{M} no afecta la tasa para estimar f aún cuando \mathcal{M} es desconocida. Sin embargo, la función de densidad de volumen $\theta_p(q)$ no es constante tan pronto como $d \geq 2$ y obtener un panorama global en mayores dimensiones es todavía un problema abierto y presumiblemente muy desafiante.

— (Berenfeld y Hoffmann, 2021, §1.2, «Resultados Principales»)

Para ganar en intuición, consideraremos $\theta_p(q)$ para algunas variedades profusamente estudiadas.

3.4.4 La densidad de volumen $\theta_p(p)$ en variedades «planas»

Observación: En el entorno de p en que el espacio es localmente análogo a \mathbb{R}^d , $\theta_p(q) = 1$.

En los espacios «planos» la métrica g es constante a través de toda la variedad g_p . El espacio euclídeo \mathbb{R}^d acompañado de la métrica habitual dotado de la métrica habitual tiene por distancia $d_I(x, y) = \sqrt{\|x - y\|} = \sqrt{(x - y)^T \mathbf{I}_d (x - y)}$. El espacio euclídeo con distancia d_Σ de Mahalanobis también es plano, sólo que con distancia $d_\Sigma(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)} = \sqrt{\|\Sigma^{-\frac{1}{2}}(x - y)\|}$. d_Σ no es «isotrópica»: en algunas direcciones cambia más rápido: tiene mayor *velocidad*.

El *tensor métrico* g es constante y de dimensión finita en ambos casos, así que esta «forma bilinear simétrica positiva definida» se puede representar con única matriz definida positiva $g = g_{ij}, g \in \mathbb{R}^{d \times d}$ que se conoce como *tensor métrico*. A la distancia «habitual» en \mathbb{R}^d le corresponde $g = \mathbf{I}_d$, a la distancia de mahalanobis $g = \Sigma$.

Al tener radio de inyectividad infinito, basta con una única carta para cubrir el espacio euclídeo, de manera que su atlas maximal será de la forma $A = \{(\mathbb{R}^d, \varphi)\}$. De todos los homeomorfismos φ posibles, resulta tal vez el más «conveniente» $\exp_p^{-1} : \mathcal{M} \rightarrow T_p \mathcal{M}$ el difeomorfismo inverso al mapa exponencial.

Nótese que la distancia cuadrada $d_\Sigma^2(p, q) = \|\Sigma^{-\frac{1}{2}}(q - p)\|$ no es más que la norma de $q - p$ luego de una transformación lineal $\Sigma^{-\frac{1}{2}}$, que «manda» los puntos

$$\mathcal{M} \ni (p, q) \mapsto (x, y) = \exp_p(p, q) = (0, \exp q) \in T_p \mathcal{M} \quad (38)$$

de la variedad $\mathcal{M} = \mathbb{R}^d$ a los puntos $(0, \exp_x y)$ del espacio tangente a \mathcal{M} en p , $T_p \mathcal{M} = \mathbb{R}^d$. Usamos (p, q) para referirnos a los puntos en \mathcal{M} y (x, y) para $T_p \mathcal{M}$.

³³En este contexto, se entiende que una variedad es más regular mientras menos varíe su densidad de volumen punto a punto

$\Sigma^{-\frac{1}{2}}$ no es otra cosa más que el mapa exponencial inverso, $\forall p \in \mathcal{M}$, $\exp_p^{-1} q = \Sigma^{-\frac{1}{2}}(q - p)$ y su «directo» es, entonces:

$$\exp_x y : T_p \mathcal{M} \rightarrow \mathcal{M} = \Sigma^{\frac{1}{2}}(y - x) \quad (39)$$

Habiendo obtenido $\Sigma^{\frac{1}{2}}(q - p) = \exp_p^{-1}(q)$, reemplazamos en la definición de densidad de volumen y obtenemos

$$\theta_p(q) = \frac{\mu_{\exp_p^* g}}{\mu_{g_p}}(\Sigma^{-\frac{1}{2}}(q - p)) \quad (40)$$

Consideremos $s = q - p$. El elemento de volumen según la estructura euclídea no es otro más que $\mu_{g_p}(\Sigma^{-\frac{1}{2}}s) = |\det \Sigma^{-\frac{1}{2}}| \|s\|$. La medida del *pullback* de g hacia el espacio tangente, resulta de

1. transportar s de $T_p \mathcal{M}$ con el mapa exponencial a \mathcal{M} , y
2. tomar la medida μ_{g_p} de $\exp s$

$$\mu_{\exp_p^* g}(\Sigma^{-\frac{1}{2}}s) = \mu_{g_p}(\exp_p(\Sigma^{-\frac{1}{2}}s)) = \mu_{g_p}(\mathbf{I}s) = \|s\| \quad (41)$$

de manera que para $p, q \in \mathcal{M}, s = \Sigma^{-\frac{1}{2}}(q - p)$,

$$\theta_p(q) = \frac{\mu_{\exp_p^* g}(s)}{\mu_{g_p}(s)} = \frac{\|s\|}{|\det \Sigma^{-\frac{1}{2}}| \|s\|} = |\det \Sigma|^{\frac{1}{2}} \quad (42)$$

para todo $p, q \in \mathcal{M}$. Recordemos de la definición de [Definición 3.3.2.1](#) que el estimador de densidad por núcleos multivariado con matriz de suavización \mathbf{H} es

$$\hat{f}(t; \mathbf{H}) = N^{-1} \sum_{i=1}^N |\det \mathbf{H}|^{-\frac{1}{2}} K(\mathbf{H}^{-\frac{1}{2}}(t - x_i)) \quad (43)$$

consideremos $\mathbf{H} = h^2 \Sigma$, $h \in \mathbb{R}$, $\Sigma \in \mathbb{R}^{d \times d}$:

$$\hat{f}(t; \mathbf{H}) = N^{-1} \sum_{i=1}^N h^{-d} |\det \Sigma|^{-\frac{1}{2}} K\left(\frac{\Sigma^{-\frac{1}{2}}(t - x_i)}{h}\right) \quad (44)$$

donde $|\det \Sigma|^{\frac{1}{2}} = \theta_p(q)$ del espacio euclídeo con métrica de Mahalanobis Σ y usábamos el núcleo normal $\Phi(x) : \mathbb{R}^d \rightarrow \mathbb{R} = (2\pi)^{-\frac{d}{2}} \exp(-\frac{\|x\|^2}{2})$ que depende de x sólo a través de su norma euclídea. Tomando la norma del argumento de $K(\cdot)$ vemos que

$$\left\| \frac{\Sigma^{-\frac{1}{2}}(t - x_i)}{h} \right\| = \frac{1}{|h|} \|\Sigma^{-\frac{1}{2}}(t - x_i)\| = \frac{d_\Sigma(t, x_i)}{h} \quad (45)$$

. De manera que K sólo depende de t a través de $d_\Sigma(t, x_i)/h$. Tomemos

$$\tilde{K}\left(\frac{d_\Sigma(t, x_i)}{h}\right) = K\left(h^{-1} \Sigma^{-\frac{1}{2}}(t - x_i)\right) \quad (46)$$

y recordemos que además $\theta_p(q) = |\det \Sigma|^{\frac{1}{2}}$ cuando $g = \Sigma$. Luego,

$$\hat{f}(t; \mathbf{H}) = N^{-1} \sum_{i=1}^N \frac{1}{h^d} \frac{1}{\theta_{X_i}(t)} \tilde{K}\left(\frac{d_g(t, x_i)}{h}\right) \quad (47)$$

y resulta que [Definición 3.3.2.1](#) es una caso especial de [Definición 3.4.3.1](#).

3.4.5 Densidad de volumen en la esfera

Una variedad plana tiene *curvatura*³⁴ nula en todo punto. De entre las variedades curvas, las d – esferas son de las más sencillas, y tienen curvatura *positiva y constante*.

Esta estructura vuelven *razonable* el cómputo de $\theta_p(q)$ en S^d .

En *Kernel Density Estimation on Riemannian Manifolds: Asymptotic Results* (Henry y Rodriguez, 2009), Guillermo Henry y Daniela Rodriguez estudian algunas propiedades asintótica de este estimador, y las ejemplifican con datos de sitios volcánicos en la superficie terrestre. Para ello, calculan $\theta_p(q)$ y llegan a que

$$\theta_p(q) = \begin{cases} R \frac{|\sin(d_g(p,q)/R)|}{d_g(p,q)} & \text{si } q \neq p, -p \\ 1 & \text{si } q = p \end{cases} \quad (48)$$

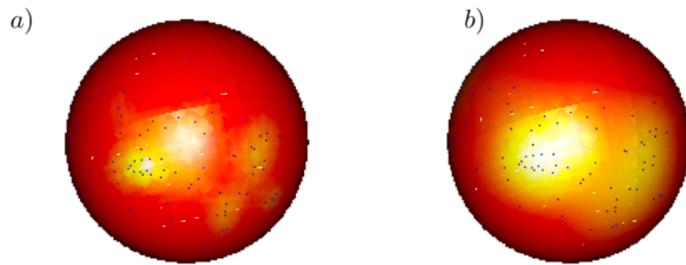


Fig. 1 The nonparametric density estimator using different bandwidth, **a** $h = 1500$ and **b** $h = 3000$

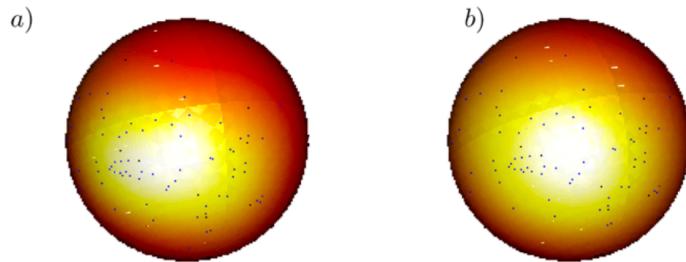


Fig. 2 The nonparametric density estimator using different bandwidth, **a** $h = 5000$ and **b** $h = 7000$

Figura 6: KDE en S^2 para $X = \text{sth}$ los flujos de lava de Fisher TODO mejorar imagen

³⁴la *curvatura* de un espacio es una de las propiedades fundamentales que estudia la geometría riemanniana; en este contexto, basta con la comprensión intuitiva de que una variedad no-plana tiene *cierta* curvatura

³⁵Recordemos que la antípoda de $p, -p$ cae justo fuera de $\text{iny}_p S^d$

3.5 Clasificación en variedades

Un desarrollo directo del estimador de [Definición 3.4.3.1](#) consta en *A kernel based classifier on a Riemannian manifold* (Loubes y Pelletier, 2008), donde construyen un clasificador para un objetivo de dos clases $\mathcal{G} \in \{0, 1\}$ con inputs X soportadas sobre una variedad de Riemann. A tal fin, minimizan la pérdida 0 – 1 y siguen la regla de Bayes, de manera que su clasificador *duro* resulta:

$$\hat{G}(X) = \begin{cases} 1 & \text{si } \widehat{\Pr}(G = 1|X) > \widehat{\Pr}(G = 0|X) \\ 0 & \text{si no} \end{cases} \quad (49)$$

que está de acuerdo con el estimador del clasificador de Bayes basado en densidad por núcleos para K clases propuesto [Definición 3.2.1.1](#).

Una notación simplificada surge de estudiar la expresión que el clasificador intenta maximizar. Para todo $i \in [K]$,

$$\widehat{\Pr}(G = i|X) = \frac{\hat{f}_i(x) \times \hat{\pi}_i}{\underbrace{\left(\sum_{i \in [K]} \hat{f}_i(x) \times \hat{\pi}_i \right)}_c} = c^{-1} \times \hat{f}_i(x) \times \hat{\pi}_i \quad (50)$$

de modo que la tarea es equivalente a maximizar $\hat{f}_i(x) \times \hat{\pi}_i$ sobre $i \in [K]$. Es fácil ver que podemos escribir el estimador de densidad de la clase k como:

$$\begin{aligned} \hat{f}_k(x) &= N_k^{-1} \sum_{i=1}^N K_h(x, X_i) \\ &= \frac{\sum_{i=1}^N \mathbb{1}(G_i = k) K_h(x, X_i)}{\sum_{i=1}^N \mathbb{1}(G_i = k)} \end{aligned} \quad (51)$$

como además $\hat{\pi}_k = N_k/N = N^{-1} \sum_{i=1}^N \mathbb{1}(G_i = k)$, resulta que

$$\begin{aligned} \hat{f}_i(x) \times \hat{\pi}_i &= \frac{\sum_{i=1}^N \mathbb{1}(G_i = k) K_h(x, X_i)}{\sum_{i=1}^N \mathbb{1}(G_i = k)} \times \frac{\sum_{i=1}^N \mathbb{1}(G_i = k)}{N} \\ &= N^{-1} \sum_{i=1}^N \mathbb{1}(G_i = k) K_h(x, X_i) \end{aligned} \quad (52)$$

Y suprimiendo la constante N concluimos que la regla de clasificación resulta equivalente a:

$$\hat{G}(p) = \arg \max_{k \in [K]} \sum_{i=1}^N \mathbb{1}(G_i = k) K_h(p, X_i) \quad (53)$$

para todo $p \in \mathcal{M}$ con K_{h_n} un núcleo isotrópico con sucesión de ventanas h_n (Loubes y Pelletier, 2008, Ecuación 3.1).

La belleza de esta regla, es que combina «sin costuras» el peso de los *priors* $\hat{\pi}_i$ - a través de los elementos no nulos de la suma cuando $\mathbb{1}(G_i =$

$k) = 1$) - con el peso de la «evidencia» - vía su cercanía «suavizada» al punto de interés $K_h(p, X_i)$.

Los autores toman de (Devroye, Györfi y Lugosi, 1996) el siguiente concepto de *consistencia fuerte*:

Definición 3.5.1 (consistencia de un clasificador (Devroye, Györfi y Lugosi, 1996, §6.1)): Sea $\hat{G}_1, \dots, \hat{G}_n$ una secuencia de clasificadores³⁶ de modo que el i -ésimo clasificador está construido con las primeras i observaciones de la muestra \mathbf{X}, \mathbf{g} . Sea L_n la pérdida 0 – 1 que alcanza el n -ésimo clasificador de la regla, y L^* la pérdida que alcanza el clasificador de Bayes de Ecuación 4.

Diremos que la regla \hat{G}_n es (débilmente) consistente - o asintóticamente eficiente en el sentido del riesgo de Bayes - para cierta distribución (X, G) si cuando $n \rightarrow \infty$

$$\mathbb{E}L_n = \Pr(\hat{G}_n(X) \neq G) \rightarrow L^* \quad (54)$$

y fuertemente consistente si

$$\lim_{n \rightarrow \infty} L_n = L^* \text{ con probabilidad 1} \quad (55)$$

En el trabajo, se prueba que el clasificador propuesto es fuertemente consistente para $K = 2$.

3.6 Aprendizaje de distancias

La hipótesis de la variedad nos ofrece un marco teórico en el que abordar la clasificación en alta dimensión, y encontramos en la literatura que la estimación de densidad por núcleos en variedades está estudiada y tiene buenas garantías de convergencia. Por alentador que resulte, nos resta un problema fundamental: no solemos conocer la variedad que soporta las X . Salvo que los datasets estén generados sintéticamente o el dominio de estudio tenga historia de trabajar con ciertas variedades bien definidas, tendremos problemas tanto para definir adecuadamente la distancia d_g como en el cálculo de la densidad de volumen $\theta_p(q)$ de Definición 3.4.3.1.

³⁶A veces también llama una *regla* de clasificación

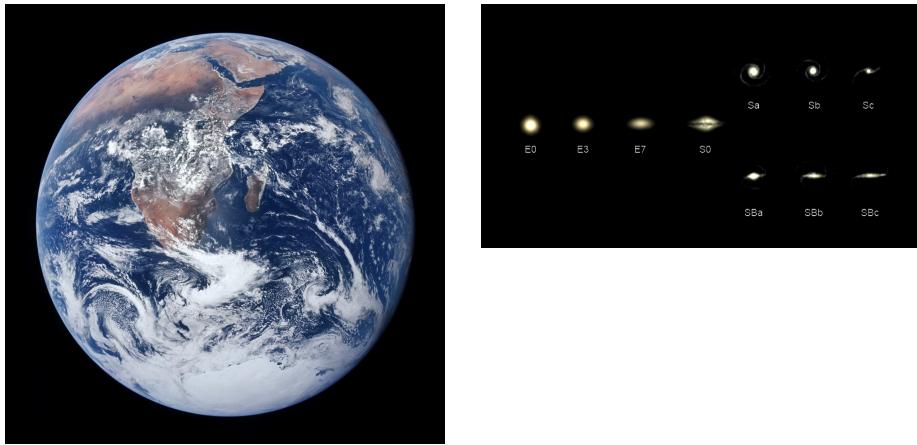


Figura 7: Data espacial con dimensiones bien definidas. Los datos geoespaciales están sobre la corteza terrestre, que es aproximadamente la 2 – esfera $S^2 \in \mathbb{R}^3$ que representa la frontera de nuestra «canica azul» (izq.), una 3 – bola. La clasificación clásica de Hubble distingue literalmente *variedades* «elípticas», «espirales» e «irregulares» de galaxias (der.).³⁷ Considere, por caso, el diagrama de Figura 8 una 1 – variedad - una curva - $\mathcal{U} \subset \mathbb{R}^2$. El espacio ambiente (\mathbb{R}^3) es también su propio espacio tangente, y las geodésicas que irradian desde el punto verde alcanzan antes al rojo que al amarillo. Sobre la variedad \mathcal{U} , el punto amarillo está aproximadamente en la dirección del espacio tangente al punto verde, mientras que el rojo está en dirección perpendicular al mismo.

³⁷Se me perdonará la simplificación; es bien sabido que en realidad la [topología del espacio-tiempo](#) es un tópico de estudio clave en la relatividad general.

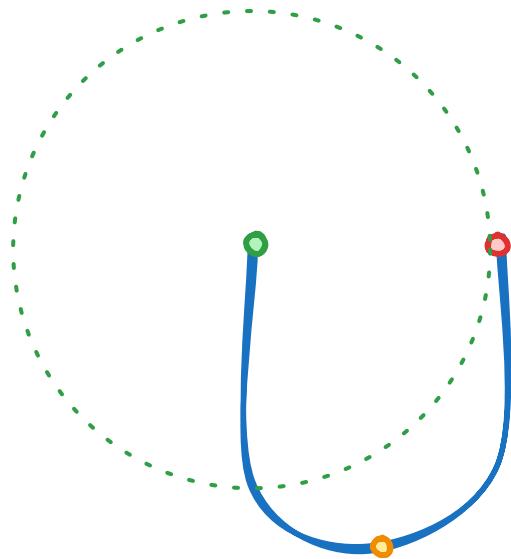


Figura 8: La variedad \mathcal{U} con $\dim(\mathcal{U}) = 1$ embebida en \mathbb{R}^2 . Nótese que en el espacio ambiente, el punto rojo está más cerca del verde, mientras que a través de \mathcal{U} , el punto amarillo está más próximo que el rojo. A los fines de estimar la densidad de X entonces, lo que nos importa es contar con una noción de *distancia* apropiada en \mathcal{M} . La distancia entre p y q es la longitud de la curva geodésica que los une; la longitud de una curva se obtiene integrándola en toda su extensión; integrarla implica conocer el espacio tangente y la métrica g en toda su extensión. Por ende, «conocer la variedad» $(\mathcal{M}, g) = \text{sop } X$ y «computar la distancia d_g inducida por su métrica g » son esencialmente la misma tarea.

En este ejemplo con tan solo $n = 3$ observaciones, es casi imposible distinguir \mathcal{U} , pero con una muestra \mathbf{X} «suficientemente grande», es de esperar que los propios datos revelen la forma de la variedad, y por eso hablamos de «aprendizaje de distancias» a partir de la propia muestra.

La distancia nos da entonces una *representación* útil de la similitud entre puntos: a mayor similitud, menor distancia. Y el *aprendizaje de representaciones*, es exactamente otro de los nombres que se le da a la estimación de variedades. En un extenso censo del campo de aprendizaje de representaciones, (Bengio, Courville y Vincent, 2014) así lo explican:

(...) [L]a principal tarea del aprendizaje no-supervisado se considera entonces como el modelado de la estructura de la variedad que sustenta los datos. La representación asociada que se aprende puede asociarse con un sistema de coordenadas intrínseco en la variedad embebida.

— (Bengio, Courville y Vincent, 2014, §8)

3.6.1 El ejemplo canónica: Análisis de Componentes Principales (PCA)

El término «hipótesis de la variedad es bastante moderno», pero el concepto está presente hace más de un siglo en la teoría estadística³⁸.

El algoritmo arquetípico de modelado de variedades es, como era de esperar, también el algoritmo arquetípico de aprendizaje de representaciones de baja dimensión: el Análisis de Componentes Principales, PCA (Pearson, 1901), que dada $\mathbf{X} \in \mathbb{R}^p$, devuelve en orden decreciente las «direcciones de mayor variabilidad» en los datos, $\mathbf{U}_p = (u_1, u_2, \dots, u_p)$. Proyectar \mathbf{X} sobre las primeras $k \leq p$ direcciones,

$$\hat{\mathbf{X}} = \mathbf{X}\mathbf{U}_k \in \mathbb{R}^{n \times k}, \hat{X}_i = (\hat{X}_{i1}, \dots, \hat{X}_{ik})^T \quad (56)$$

nos devuelve la «mejor»³⁹ representación lineal de dimensión k .

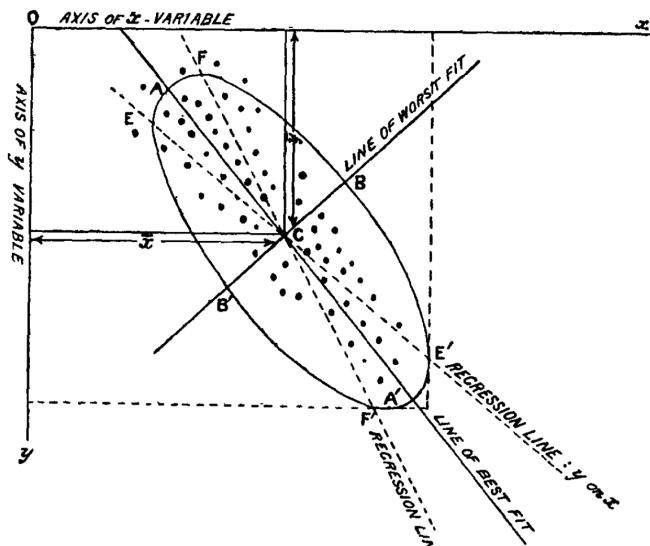


Figura 9: Ilustración de \mathbf{X} y sus componentes principales en «*LIII. On lines and planes of closest fit to systems of points in space.*» (Pearson, 1901)

Hemos hecho ya hincapié en que las variedades que buscamos seguramente sea fuertemente no-lineales; sin embargo, todavía hay lugar para PCA en esta aventura: cuando el dataset tiene dimensión verdaderamente muy alta, un proceso razonable consistirá en primero disminuir la dimensión a un subespacio lineal casi idéntico al original con PCA, y recién en este subespacio aplicar técnicas más complejas de aprendizaje de distancias. Aprovechando que al menos las observaciones de entrena-

³⁸estas referencias vienen del mismo Bengio [comentando en Reddit sobre el origen del término](#)

³⁹cuya definición precisa obviamos.

miento son puntos conocidos de la variedad⁴⁰, y que en la variedad el espacio es *localmente euclídeo* (Vincent y Bengio, 2002) parten del estimador de [Definición 3.3.2.1](#) pero en lugar de utilizar un núcleo $K_{\mathbf{H}}$ fijo en cada observación x_i , se proponen primero hacer análisis de componentes principales de la matriz de covarianza *pesada* estimada en cada punto,

$$\hat{\Sigma}_{\mathcal{K}_i} = \hat{\Sigma}_{\mathcal{K}}(x_i) = \frac{\sum_{j \in [N]-i} \mathcal{K}(x_i, x_j)(x_j - x_i)(x_j - x_i)^T}{\sum_{j \in [N]-i} \mathcal{K}(x_i, x_j)} \quad (57)$$

donde \mathcal{K} es alguna medida de cercanía en el espacio ambiente (e.g. la densidad normal multivariada Φ ya mencionada), con lo cual la estimación de densidad resulta:

$$\hat{f}(x) = N^{-1} \sum_{i=1}^N |\det \hat{\Sigma}_i|^{-\frac{1}{2}} K\left(\hat{\Sigma}_i^{-\frac{1}{2}} t\right) \quad (58)$$

Ahora bien, computar una $\hat{\Sigma}_{\mathcal{K}_i} \forall i \in [N]$ y su inversa es sumamente costoso, por lo que los autores agregan un refinamiento: si la variedad en cuestión es d -dimensional, es de esperar que las direcciones principales a partir de la $d+1$ -ésima sean «negligibles»⁴¹ en lugar computar las componentes principales de $\hat{\Sigma}_{\mathcal{K}_i}$, simplemente fijan de antemano la dimensión d esperada para la variedad, se quedan con las d direcciones principales⁴², «ponen en cero» el resto y «completan» la aproximación con un poco de «ruido» $\sigma^2 \mathbf{I}$. La aproximación resultante $\hat{\Sigma}_i = f(\hat{\Sigma}_{\mathcal{K}_i}) + \sigma^2 \mathbf{I}$ es mucho menos costosa de invertir, y tiene una interpretación geométrica bastante intuitiva en cada punto. Usando el mismo clasificador basado en la regla de Bayes Ecuación 4 que ya mencionamos, obtienen así resultados superadores a los de [Definición 3.3.2.1](#) con $\mathbf{H} = h^2 \mathbf{I}$. Hemos de notar, sin embargo, dos dificultades:

- todavía no está nada claro cuál debería ser la dimensión intrínseca d cuando la variedad es desconocida, y
- no es suficiente para computar KDE en variedades según [Definición 3.4.3.1](#), pues $\hat{\Sigma}_i$ sólo aproxima el tensor métrico en cada x_i , y para computar $\theta_p(q)$ necesitamos conocer g en todo punto.⁴³

⁴⁰módulo el error de medición y/o el efecto de covariables no medidas

⁴¹la sugerente metáfora que usan en el trabajo, es que en lugar de ubicar una «bola» de densidad alrededor de cada observación x_i , quieren ubicar un «panqueque» tangente a la variedad

⁴²en la práctica, las obtienen usando SVD - descomposición en valores singulares, TODO at wikipedia (Hastie, Tibshirani y Friedman, 2009, pág. 64)

⁴³El grupo de investigación de Bengio, Vincent, Rifai et al. continuó trabajando estos estimadores, con especial énfasis en la necesidad de aprender una geometría *global* de la variedad para evitar el crecimiento exponencial de tamaño muestral que exigen los métodos locales como KDE en alta dimensión o variedades muy «rugosas», pero aquí se separan nuestros caminos. Una brevísima reseña: en (Bengio, Larochelle y Vincent, 2005) agregan restricciones globales a las estimaciones de los núcleos punto a punto que computan simultáneamente con redes neuronales, y en (Rifai et al., 2011)

En un trabajo contemporáneo a (Vincent y Bengio, 2002), «Charting a Manifold» (Brand, 2002), los autores intentan encarar frontalmente las limitaciones recién mencionadas, en tres etapas:

1. estimar la dimensión intrínseca de la variedad $d_{\mathcal{M}}$; luego
2. definir un conjunto de cartas centradas en cada observación $x_i \in \mathcal{M}$ que minimicen una *divergencia* global, y finalmente
3. «coser» las cartas a través de una *conexión* global sobre la variedad.

El procedimiento para estimar $d_{\mathcal{M}}$ es ingenioso, pero costoso. Sean $\mathbf{X} = (x_1^T, \dots, x_N^T)$ observaciones p -dimensionales, que han sido muestreados de una distribución en (\mathcal{M}, g) , $\dim \mathcal{M} = d < p$ con algo de ruido *isotrópico*⁴⁴ p -dimensional. Consideremos una bola $B_r(0)$ centrada en un punto cualquiera de \mathcal{M} , y consideremos la tasa $t(r)$ a la que incorpora observaciones vecinas. Cuando r está en la escala del ruido, la bola incorpora puntos «rápidamente», pues hay dispersión en todas las direcciones. A medida que r llega a la escala en la que el espacio es localmente análogo a \mathbb{R}^d , la incorporación de nuevos puntos disminuye, pues sólo habrá nuevas observaciones en las d direcciones tangentes. Si r sigue creciendo la bola $B_r(0)$ eventualmente alcanzará la escala de la *curvatura* de la variedad, momento en el que comenzará a acelerarse nuevamente la incorporación de puntos. Analizando $\arg \max_r t(r)$ podemos identificar la dimensión intrínseca de la variedad.⁴⁵

aprenden explícitamente un atlas que luego usan para clasificación con TangentProp (Simard *et al.*, 1991), una modificación del algoritmo de *backpropagation* que se usa en redes neuronales, que busca conservar «las direcciones tangentes» a las observaciones en la representación aprendida.

⁴⁴Del griego *iso-*, «igual» y *-tropos*, «dirección»; «igual en todas las direcciones»

⁴⁵Más precisamente, el *paper* utiliza otra función de r , $c(r)$ que se *maximiza* cuando $r \approx \frac{1}{d}$, y considera las dificultades entre estimar d punto a punto o globalmente.

Scale behavior of a 1D manifold in 2-space

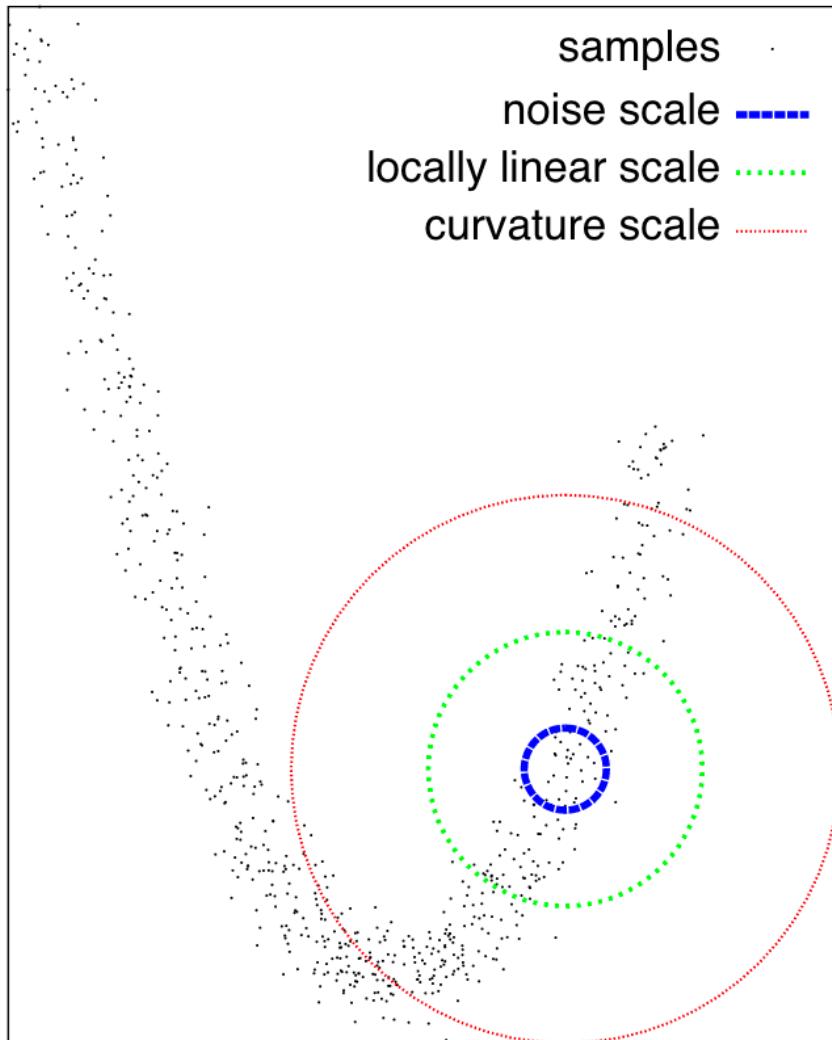


Figura 10: Una bola de radio r creciente centrada en un punto de una 1-variedad muestreada con ruido en \mathbb{R}^2 minimiza la tasa a la que incorpora observaciones cuando r está en la escala «localmente lineal» de la variedad.

Definido d , los pasos siguientes no son menos complejos. Por un lado, plantean un sistema ecuaciones para obtener *al mismo tiempo* todos los entornos coordinados (que no son otra cosa más que un GMM - gaussian mixture modelling⁴⁶ - centrado en cada observación (o sea que $\mu_j = x_j$, y resuelve simultáneamente $\Sigma_j \forall j \in [N]$) minimizando la *divergencia*

⁴⁶modelo de mezcla de (distribuciones) gaussianas

entre Σ_j vecinos⁴⁷. Finalmente, han de encontrar una *conexión* entre los entornos coordenados de cada observación, de manera que se puedan definir coordenadas para *cualquier* punto de la variedad y con ellas formar un atlas diferenciable.

Una *conexión* es otro - y van... - término de significado muy preciso en geometría riemanniana que aquí usamos coloquialmente. Es un *objeto geométrico* que *conecta* espacios tangentes cercanos, describiendo precisamente cómo éstos varían a medida que uno se desplaza sobre la variedad, y permite entonces *diferenciarlos* para computar g_p y la métrica inducida en cualquier punto. Desde ya que con tal estructura es posible calcular $\theta_p(q) \forall p, q \in \mathcal{M}$, pero a esta altura, hemos reemplazado el problema difícil original - encontrar una buena representación de baja dimensión de una muestra \mathbf{X} para clasificarla en clases - por uno *muy difícil* sustituto: encontrar la dimensión intrínseca, un atlas diferenciable y su conexión global para una variedad desconocida. El proceso es sumamente interesante, pero complejiza en lugar de simplificar nuestro desafío inicial.

3.6.2 El algoritmo más *cool*: Isomap

Recordemos que toda esta aventura comenzó cuando identificamos que

1. en alta dimensión, la *distancia euclídea* «explotaba», y rápidamente dejaba de proveer información útil sobre la similitud entre observaciones de \mathbf{X} y además
2. de haber una estructura de menor dimensión que represente mejor las observaciones, habría de ser fuertemente no-lineal.

En rigor, *no es necesario conocer \mathcal{M}* , bastaría con conocer una aproximación a la distancia geodésica en \mathcal{M} que sirva de sustituto a la distancia euclídea en el espacio ambiente. Probablemente el algoritmo más conocido que realiza tal tarea, sea Isomap - por «mapeo isométrico de *features*».

Desarrollado a caballo del cambio de siglo por Joshua Tenenbaum et ales (Tenenbaum, 1997; Tenenbaum, Silva y Langford, 2000), el algoritmo consta de tres pasos:

⁴⁷Aquí «divergencia» tiene un significado preciso que obviamos, pero intuitivamente, representa el «costo» - la variación - que uno encuentra cuando quiere representar un punto a en el vecindario U de x_i , en las coordenadas cptes. a un vecindario V de x_j . Se puede mostrar que el cociente entre las densidad de a en ambos sistemas coordinados - la entropía cruzada entre $\mathcal{N}(x_i, \Sigma_i)$ y $\mathcal{N}(x_j, \Sigma_j)$ - es la divergencia que se busca minizar.

Definición 3.6.2.1 (algoritmo Isomap): Sean $\mathbf{X} = (x_1, \dots, x_N)$, $x_i \in \mathbb{R}^p$ N observaciones p -dimensionales. El mapeo isométrico de *features* es el resultado de:

1. Construir el grafo de vecinos más cercanos $\mathbf{NN} = (\mathbf{X}, E)$, donde cada observación x_i es un vértice y la arista⁴⁸ $e(a, b)$ que une a con b está presente si y sólo si
 - (ε -Isomap): la distancia entre a, b en el espacio ambiente es menor o igual a ε , $d_{\mathbb{R}^p}(a, b) \leq \varepsilon$.
 - (k -Isomap): b es uno de los k vecinos más cercanos de a ⁴⁹
2. Computar la distancia geodésica - el «costo» de los caminos mínimos - entre todo par de observaciones, $d_{\mathbf{NN}}(a, b) \forall a, b \in \mathbf{X}$ ⁵⁰.
3. Construir la representación - d -dimensional utilizando MDS⁵¹ en el espacio euclídeo \mathbb{R}^d que minimice una métrica de discrepancia denominada «estrés», entre las distancias $d_{\mathbf{NN}}$ de (2) y sus equivalentes en la representación, $d_{\mathbb{R}^d}$. Para elegir el valor óptimo de d - la dimensión intrínseca de los datos-, búsquese el «codo» en el gráfico de estrés en función de la dimensión de MDS.

⁴⁸edge en inglés

⁴⁹O viceversa, pues en un grafo no-dirigido la relación de vecinos más cercanos es mutua

⁵⁰A tal fin, se puede utilizar segón convenga el algoritmo de [Floyd-Warshall](#) o [Dijkstra](#)

⁵¹«Multi Dimensional Scaling», o [escalamiento multidimensional](#), un algoritmo de reducción de dimensionalidad

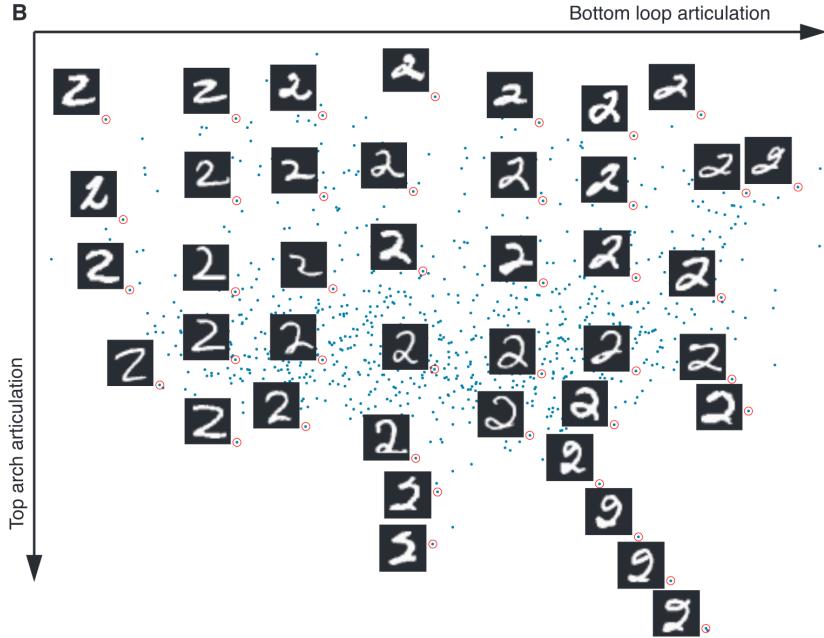


Figura 11: Isomap aplicado a 1.000 dígitos «2» manuscritos del dataset *MNIST* con $d = 2$ (Tenenbaum, Silva y Langford, 2000). Nótese que las dos direcciones se corresponden fuertemente con características de los dígitos: el rulo inferior en el eje X , y el arco superior en el eje Y .

La pieza clave del algoritmo, es la estimación de la distancia geodésica en \mathcal{M} a través de la distancia en el grafo de vecinos más cercanos. Si la muestra disponible es «suficientemente grande», es razonable esperar que en un entorno de x_0 , las distancias euclídeas aproximen bien las distancias geodésicas, y por ende un «paseo» por el grafo **NN** debería describir una curva prácticamente contenida en \mathcal{M} . Isomap resultó ser un algoritmo sumamente efectivo que avivó el interés por el aprendizaje de distancias, per todavía cuenta con un talón de Aquiles: la elección del parámetro de cercanía, ε ó k :

- valores demasiado pequeños pueden partir **NN** en más de una componente conexa, otorgando distancia «infinita» a puntos en componentes disjuntas, mientras que
- valores demasiado grandes pueden «cortocircuitar» la representación - en particular en variedades con muchos pliegues -, uniendo secciones de la variedad subyacente a través del espacio ambiente.

3.6.3 Distancias basadas en densidad

Algoritmos como isomap aprenden la *geometría* de los datos, reemplazando la distancia euclídea ambiente por la distancia euclídea en el grafo NN_k , que con $n \rightarrow \infty$ converge a la distancia d_g en \mathcal{M} . La distancia

de Mahalanobis TODO at dist mahalonobis, por su parte, aprende la *densidad* de los datos.

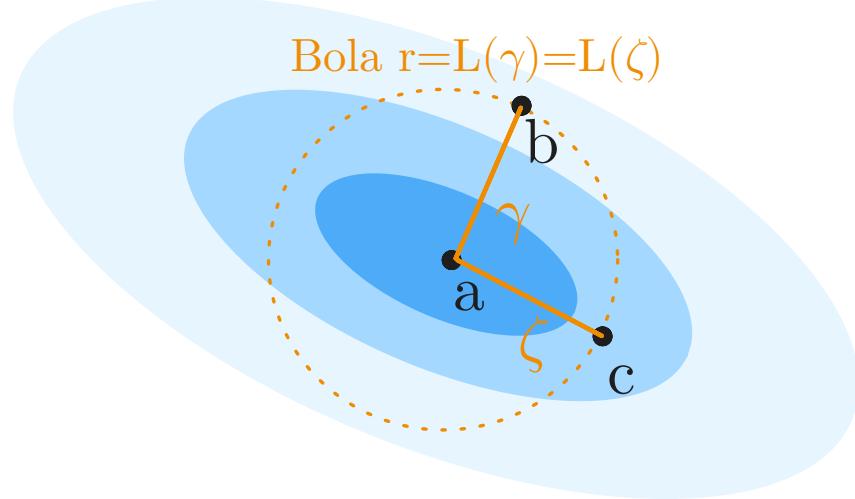


Figura 12: Cuando por ejemplo $\mathcal{M} = (\mathbb{R}^2, g = \mathbf{I})$, $X \sim \mathcal{N}_d(a, \Sigma)$, tenemos que $d_g(a, b) = L(\gamma) = r = L(\zeta) = d_g(a, c)$, mientras que $d_{\Sigma}(a, b) < d_{\Sigma}(a, c)$: la normal multivariada tiene distintas tasas de cambio en distintas direcciones, y medir distancia ignorando este hecho puede llevar a conclusiones erróneas.

Combinando estas dos nociones, podemos considerar la categoría de *distancias basadas en densidad* - DBDs -, donde curvas γ que atravesen regiones de *baja* densidad f_X en \mathcal{M} sean más «costosas» de transitar que otras de igual longitud pero por regiones de mayor densidad. Esta área del aprendizaje de distancias vio considerables avances durante el siglo XXI, a continuación del éxitop empírico de Isomap, y pavimentó el camino para técnicas de reducción de dimensionalidad basales en el «aprendizaje profundo»⁵² como los «autocodificadores»⁵³.

Aprender una DBD nos permite saltarnos el problema ya harto descrito de aprender la variedad desconocida \mathcal{M} , e ir directamente a lo único que necesitamos extraer de la variedad para tener un algoritmo de clasificación funcional: una noción de distancia adecuada.

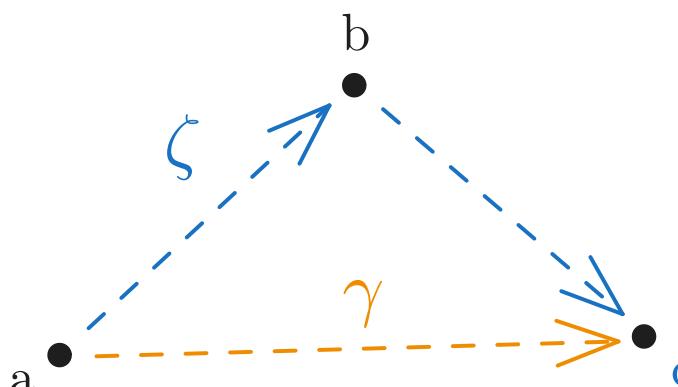
⁵²O «deep learning» en inglés. Llamamos genéricamente de tal modo a la pléthora de arquitecturas de redes neuronales con múltiples capas que dominan hoy el procesamiento de información de alta dimensión. TODO at wikipedia

⁵³«autoencoders» en inglés, algoritmo que dada \mathbf{X} , aprende un codificador $c(x) : \mathbb{R}^D \rightarrow \mathbb{R}^d$, $d \ll D$ y un decodificador $d(-1)(x) : \mathbb{R}^d \rightarrow \mathbb{R}^D$ tal que $d(c(x)) \approx x$. De hecho, uno de los «padres de la IA», Yoshua Bengio, cuyo trabajo ya mencionamos en este área, menciona en [Reddit](#) TODO at Reddit (!) cómo su grupo de investigación en la U. de Montréal trabajando en estas ideas: aprendizaje de variedades primero, y autocodificadores posteriormente.

(Vincent y Bengio, 2003) proveen una de las primeras heurísticas para una DBD: al igual que Isomap, toma las distancias de caminos mínimos pesados en un grafo con vértices \mathbf{X} , pero

- considera el grafo completo \mathbf{C} en lugar del de k -vecinos \mathbf{NN}_k y
- pesa las aristas del grafo por la distancia euclídea en el espacio ambiente entre sus extremos *al cuadrado*.

Esta noción de distancia «arista-cuadrada»⁵⁴ tiene el efecto de «desalentar grandes saltos» entre observaciones lejanas, que es otra manera de decir «asignar un costo alto a trayectos por regiones de baja densidad», por lo cual ya califica - tal vez rudimentariamente - como una DBD.



$$\|a-b\| = \|b-c\| = 2 \quad \|a-c\| = 3$$

Figura 13: En el grafo completo de 3 vértices, hay sólo dos caminos entre a y c : $\zeta = a \rightarrow b \rightarrow c$, y $\gamma = a \rightarrow c$

. Bajo la norma euclídea, $L(\gamma) = 3 < 4 = 2 + 2 = L(\zeta)$ de modo que $d(a, c) = 3$ con geodésica γ . Con la distancia de arista cuadrada, $L(\zeta) = 2^2 + 2^2 = 8 < 3^2 = L(\gamma)$, y por lo tanto $d(a, c) = 8$ con geodésica ζ . La distancia de arista cuadrada cambia las geodésicas, y también cambia la escala en que se miden las distancias.

Hay numerosos algoritmos y estudios comparativos de los mismos en esta era, así que sólo nos detendremos arbitrariamente en algunos. (Cayton, 2005) provee un resumen temprano de algunos de los algoritmos de aprendizaje de variedades más relevantes hasta entonces, y comenta además sobre el torrente aparentemente inacabable de algoritmos sugeridos: es tan amplio el espectro de variedades subyacentes y de representaciones «útiles» que se pueden concebir, que (a) en el plano

⁵⁴«edge-squared distance» en inglés

teórico resulta muy difícil de obtener garantías «amplias» de eficiencia y performance, y (b) en el plano experimental, quedamos reducidos a «elegir un conjunto representativo de variedades» y observar si los resultados obtenidos son «intuitivamente agradables». Veinte años más tarde, esto mismo seguiremos haciendo en una sección posterior.

(Bijral, Ratliff y Srebro, 2012) ofrece - a nuestro entender - una de las primera formalizaciones «amplias» de qué constituye una DBD. Para abordarla, revisaremos una definición previa. En Ecuación 26 mencionamos sin precisiones que dada una variedad de Riemann compacta y sin frontera (\mathcal{M}, g) , la longitud de una *curva rectificable* $\gamma \subset \mathcal{M}$ parametrizada en $[0, 1]$ es

$$L(\gamma) = \int_0^1 \|\gamma'(t)\| dt = \int_0^1 \sqrt{g_{\gamma(t)}(\gamma'(t), \gamma'(t))} dt \quad (59)$$

Definición 3.6.3.1 (curva rectificable): Una *curva rectificable* es una curva que tiene longitud finita. Más formalmente, sea $\gamma : [a, b] \rightarrow \mathcal{M}$ una curva parametrizada. La curva es rectificable si su longitud de arco es finita:

$$L(\gamma) = \sup \sum_{i=1}^n |\gamma(t_i) - \gamma(t_{i-1})| < \infty \quad (60)$$

donde el supremo se toma sobre todas las particiones posibles $a = t_0 < t_1 < \dots < t_n = b$ del intervalo $[a, b]$.

Equivalentemente, si γ es diferenciable por tramos, entonces es rectificable si y solo si:

$$L(\gamma) = \int_a^b |\gamma'(t)| dt < \infty \quad (61)$$

Las curvas rectificables son importantes porque permiten definir conceptos como la longitud de arco y la parametrización por longitud de arco, que son fundamentales en geometría diferencial y análisis. En particular, sea $\gamma : [a, b] \rightarrow \mathbb{R}^n$ una curva rectificable parametrizada y diferenciable por tramos y $f : \mathbb{R}^n \rightarrow \mathbb{R}$ una función diferenciable. La integral de línea de f sobre γ se define como:

$$\int_{\gamma} f ds = \int_a^b f(\gamma(t)) |\gamma'(t)| dt \quad (62)$$

donde ds representa el elemento de longitud de arco.

Si γ tiene longitud finita y f es continua – como en nuestro caso de uso –, el resultado de la integral existe y es independiente de la parametrización.

Sea entonces $X \sim f$, $f : \mathcal{M} \rightarrow \mathbb{R}_+$ un elemento aleatorio distribuido según f sobre una variedad de Riemann compacta y sin frontera –

potencialmente desconocida – \mathcal{M} . Sea además $g(t) : \mathbb{R}_+ \rightarrow \mathbb{R}$ una función *monótonicamente decreciente* en su parámetro. Consideraremos el *costo* J_f de un camino $\gamma : [0, 1] \rightarrow \mathcal{M}$, $\gamma(0) = p, \gamma(1) = q$ entre p, q como la integral de $g \circ f$ a lo largo de γ :

$$J_{g \circ f}(\gamma) = \int_0^1 g \text{LR}((f(\gamma(t)))) \|\gamma'(t)\|_p dt \quad (63)$$

Y la distancia basada en la densidad f pesada por g entre dos puntos cualesquiera $p, q \in \mathcal{M}$ como

$$D_{g \circ f}(p, q) = \inf_{\gamma} J_{g \circ f}(\gamma) \quad (64)$$

, donde la minimización es con respecto a todos los senderos rectificables con extremos en p, q , y $\|\cdot\|_p$ es la p –norma o distancia de Minkowski con parámetro p .

Observación: La longitud de Ecuación 26 es equivalente a tomar una función constante $g(t) = 1$ y $p = 2$

Definición 3.6.3.2 (norma p): Sea $p \geq 1$. Para $x, y \in \mathbb{R}^d$, la norma ℓ_p ⁵⁵ se define como:

$$\|x\|_p = \left(\sum_{i=1}^d |x_i|^p \right)^{\frac{1}{p}} \quad (65)$$

Observación: Cada p –norma induce su propia distancia d_p . Algunas son muy conocidas:

- $p = 1$ da la distancia «taxi» o «de Manhattan»⁵⁶:

$$d_1(x, y) = \|x - y\|_1 = \sum_{i=1}^d |x_i - y_i| \quad (66)$$

,

- $p = 2$ da la distancia euclídea que ya hemos usado, omitiendo el subíndice 2:

$$d_2(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^d (x_i - y_i)^2} \quad (67)$$

,

- $p \rightarrow \infty$ da la distancia de Chebyshev,

$$\|x\|_{p \rightarrow \infty} = \max_{1 \leq i \leq d} |x_i - y_i| \quad (68)$$

¿Es posible estimar $D_{g \circ f}$ de manera consistente? Intuitivamente, consideremos dos puntos $a, b \in U \subset \mathcal{M}$, $\dim \mathcal{M} = d$ en un vecindario U de a lo

⁵⁵También conocida como « p –norma» o «distancia de Minkowski»

⁵⁶Llamada así porque representa la distancia que recorrería un taxi en una grilla urbana. Una traducción razonable sería *distancia de San Telmo*

«suficientemente pequeño» como para que f sea esencialmente uniforme en él, y en particular en el segmento $\gamma_{ab} = \overline{ab}$ y tomemos $g = 1/f^r$:

$$\begin{aligned} J_r(\gamma_{ab}) &= D_r(a, b) \approx g(\text{alrededor de } a \text{ y } b) \|b - a\|_p \\ &\propto g(\|b - a\|_p^{-d}) \|b - a\|_p \\ &= \|b - a\|_p^{rd+1} = \|b - a\|_p^q \end{aligned} \quad (69)$$

,

donde $q = r \times d + 1$. Nótese que como ya mencionamos, tomar $q = 1$ (o $r = 0$) devuelve la distancia de Minkowski.

Luego, el costo de un paseo de k pasos por el grafo completo de \mathbf{X} , $\gamma = (\pi_0, \pi_1, \dots, \pi_{i_k}), \pi_j^T \in \mathbf{X} \forall j \in [k]$ por el grafo completo de \mathbf{X} se puede computar con una simple suma:

$$J_r(\gamma) = \sum_{j=1}^k D_r(\pi_{j-1}, \pi_j) \approx \sum_{j=1}^k \|\pi_j - \pi_{j-1}\|_p^q \quad (70)$$

se puede computar similarmente,

que a su vez nos permite estimar las distancias geodésicas D_r como los «caminos mínimos» en el grafo completo de \mathbf{X} con aristas pesadas por $\|b - a\|_p^q$, $a^T, b^T \in \mathbf{X}$.

Esta estimación es particularmente atractiva, en tanto no depende para nada de la dimensión ambiente D , y sólo depende de la dimensión intrínseca d de \mathcal{M} a través de $q = rd + 1$. De hecho, los autores mencionan que «casi cualquier par de valores (p, q) funciona», y en particular encuentran que en sus experimentos, $p = 2, q = 8$ «anda bien en general» (Bijral, Ratliff y Srebro, 2012, 5.1)⁵⁷.

Queda de manifiesto que hay una estrecha relación entre las distancias de caminos mínimos con aristas pesadas por una potencia $q = rd + 1$ - que sólo está definida entre observaciones de \mathbf{X} , con la distancia $D_r = \inf_\gamma \left(\int_\gamma \frac{1}{f^r} ds \right)$, que a priori está definida globalmente en \mathcal{M} .

Un resultado interesante por lo exacto, aparece en (Chu, Miller y Sheehy, 2019). Dado un conjunto de puntos $P = \{p_1, \dots, p_N\}, p_i \in \mathcal{M} \forall i \in [N]$, Considérese la «métrica de vecino más cercano»

$$r_{P(q)} = 4 \min_{p \in P} \|q - p\| \quad (71)$$

,

que da lugar a la función de costo

$$J_{r_P}(\gamma) = \int_0^1 r_P(\gamma(t)) \|\gamma'(t)\| dt \quad (72)$$

, que a su vez define la distancia

$$D_{r_P} = \inf_\gamma J_{r_P}(\gamma) \quad (73)$$

⁵⁷tendremos más para decir al respecto en la sección de Experimentos TODO link experimentos

que llaman distancia de vecino más cercano, $d_N = D_{r_P}$.

Considérese además la distancia de arista-cuadrada:

$$d_2(a, b) = \inf_{(p_0, \dots, p_k)} \sum_{i=1}^k \|p_i - p_{i-1}\|^2 \quad (74)$$

donde el ínfimo se toma sobre toda posible secuencia de puntos $p_0, \dots, p_k \in P$, $p_0 = a$, $p_k = b$. Resulta entonces que la distancia de vecino más cercano d_N y la métrica de arista cuadrada d_2 son equivalentes para todo conjunto de puntos P en dimensión arbitraria. (Chu, Miller y Sheehy, 2019, Teorema 1.1)⁵⁸.

Probar la equivalencia para el caso trivial con $P = \{a, b\} \subset \mathbb{R}^D$ se convierte en un ejercicio de análisis muy sencillo, que cementa la intuición y explica el factor de 4 original:

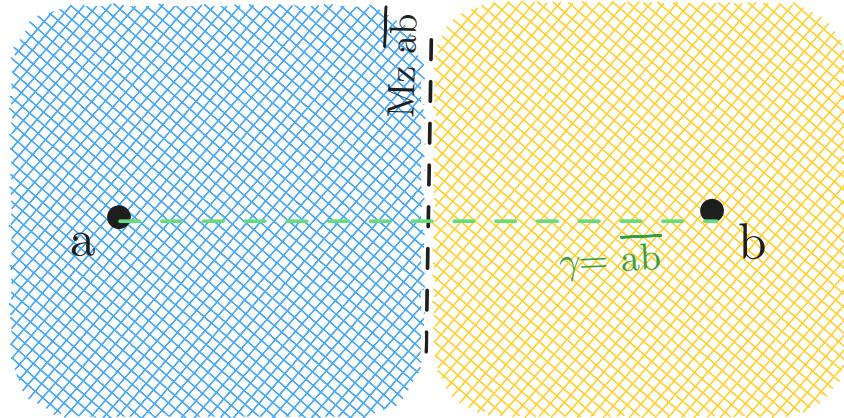


Figura 14: Ejemplo trivial de la equivalencia $d_N \equiv d_2$ para $P = \{a, b\}$. En la mitad del segmento \overline{ab} más cercana a a (región azul), d_N es $\|z - a\|^2$; análogamente, en la región naranja $d_N = \|z - b\|^2$.

$$\gamma(t) : [0, 1] \rightarrow \mathbb{R}^D, \gamma(t) = (1-t)a + tb, \gamma'(t) = b - a \quad (75)$$

⁵⁸De hecho, la prueba que ofrecen es un poco más general: los elementos de P no tienen por qué ser puntos en \mathcal{M} , sino que pueden ser conjuntos compactos, con costo cero al atravesarlos, cf. (Chu, Miller y Sheehy, 2019, Figura 2)

$$\begin{aligned}
d_{\mathbf{N}}(a, b) &= J_{r_P}(\gamma) = \int_0^1 r_{\{a,b\}}(\gamma(t)) \times \|\gamma'(t)\| dt \\
&= \int_0^1 4 \min_{p \in \{a,b\}} \|(a + (b-a)t) - p\| \|b - a\| dt \\
&= 4\|b - a\| \left(\int_0^{\frac{1}{2}} \|a + (b-a)t - a\| dt + \int_{\frac{1}{2}}^1 \|a + (b-a)t - b\| dt \right) \tag{76} \\
&= 4\|b - a\| \left(\int_0^{\frac{1}{2}} \|(b-a)t\| dt + \int_{\frac{1}{2}}^1 \|(a-b)(1-t)\| dt \right) \\
&= 4\|b - a\|^2 \left(\int_0^{\frac{1}{2}} t dt + \int_{\frac{1}{2}}^1 (1-t) dt \right) = 4\|b - a\| \left(\frac{1}{8} + \frac{1}{8} \right) \\
&= \|b - a\|^2 = d_2(a, b)
\end{aligned}$$

El grueso del trabajo de Chu et al consiste en una prueba más general de esta igualdad, que se desarrolla en tres partes:

1. Para toda colección finita de puntos $P = \{p_i : p_i \in \mathbb{R}^D\}$,

- 1.a. $d_{\mathbf{N}} \leq d_2$
- 1.b. $d_{\mathbf{N}} \geq d_2$

2. (1) también es válido para toda colección de compactos P de \mathbb{R}^D .

Una utilidad de este resultado, es que permite calcular con precisión para qué valores de k , estimar $d_{\mathbf{N}}$ sobre el grafo pesado por aristas cuadradas $\mathbf{NN}_k(\mathbf{X})$ es «suficientemente buen sustituto» por el más costoso $\mathbf{C}(\mathbf{X})$. En (Chu, Miller y Sheehy, 2019, Theorema 1.3), observan que basta $k = O(2^d \ln n)$

Lo que Chu et al llaman d_2 y figura en (Vincent y Bengio, 2003; Chu, Miller y Sheehy, 2019) como «distancia de arista-cuadrada», es la misma distancia D_r que (Bijral, Ratliff y Srebro, 2012) considera, con $p = 2$ (norma euclídea) y $r = \frac{1}{d}$ (de modo que $q = rd + 1 = 2$). A nuestro entender, no hay pruebas de tal equivalencia para valores arbitrarios de p, q , pero sí existen resultados asintóticos para casos más generales.

3.6.4 Distancia de Fermat

We tackle the problem of learning a distance between points, able to capture both the geometry of the manifold and the underlying density. We define such a sample distance and prove the convergence, as the sample size goes to infinity, to a macroscopic one that we call Fermat distance as it minimizes a path functional, resembling Fermat principle in optics.

— P. Groisman et al (2019)

El trabajo de (Groisman, Jonckheere y Sapienza, 2019) considera la misma familia de distancias basadas en funciones monótonamente decrecientes de la densidad que (Bijral, Ratliff y Srebro, 2012), $g = \frac{1}{f^r}$, salvo que en (Groisman, Jonckheere y Sapienza, 2019),

$$p = 2; \quad q = \alpha; \quad r = \beta = \frac{\alpha - 1}{d} \quad (77)$$

y no se limita a sugerir que la distancia en el espacio ambiente, $D_r = D_{g \circ f}$ se puede aproximar a través de la distancia basada en el grafo completo de \mathbf{X} con aristas pesadas por $\|\cdot\|_2^\alpha$, sino que precisan en qué sentido la una converge a la otra, y a qué tasa.⁵⁹

Definición 3.6.4.1 (Distancia «macroscópica» de Fermat (Groisman, Jonckheere y Sapienza, 2019, Definición 2.2)):

Sea f una función continua y positiva, $\beta \geq 0$ y $x, y \in S \subseteq \mathbb{R}^D$. Definimos la *Distancia de Fermat* $\mathcal{D}_{f,\beta}(x, y)$ como:

$$\mathcal{T}_{f,\beta}(\gamma) = \int_\gamma f^{-\beta} ds, \quad \mathcal{D}_{f,\beta}(x, y) = \inf_\gamma \mathcal{T}_{f,\beta}(\gamma) \quad (78)$$

... donde el ínfimo se toma sobre el conjunto de todos los «senderos» o curvas rectificables entre x e y contenidos en \bar{S} , la clausura de S , y la integral es entendida con respecto a la longitud de arco ds dada por la distancia euclídea como siempre.

Este objeto «macroscópico» se puede aproximar a partir de una versión «microscópica» del mismo, que en límite converge a $\mathcal{D}_{f,\beta}$:

⁵⁹Con respecto a fijar $p = 2$, en la «Observación 2.6» los autores mencionan que es posible y hasta sería interesante reemplazar la norma euclídea – 2 –norma – por otra distancia – otra p –norma, por ejemplo –, reemplazando las integrales con respecto a la longitud de arco, por integrales con respecto a la distancia involucrada. Entendemos de ello que no es una condición *necesaria* para el desarrollo del trabajo, sino sólo *conveniente*.

⁶⁰Es decir, que para todo compacto $U \subset \mathbb{R}^D$, la cardinalidad de $Q \cap U$ es finita, $|Q \cap U| < \infty$.

Definición 3.6.4.2 (Distancia muestral de Fermat):

Sea Q un conjunto no-vacío, *localmente finito*⁶⁰ de \mathbb{R}^D . Para $\alpha \geq 1$ y $x, y \in \mathbb{R}^d$, la *Distancia Muestral de Fermat* se define como

$$D_{Q,\alpha} = \inf \left\{ \sum_{j=1}^{K-1} \|q_{j+1} - q_j\|^\alpha : (q_1, \dots, q_K) \text{ es un camino de } x \text{ a } y, K \geq 1 \right\} \quad (79)$$

donde los q_j son elementos de Q . Nótese que $D_{Q,\alpha}$ satisface la desigualdad triangular, define una métrica sobre Q y una pseudo-métrica⁶¹ sobre \mathbb{R}^d .

Definición 3.6.4.3 (variedad isométrica): Diremos que \mathcal{M} es una variedad d -dimensional C^1 *isométrica* embebida en \mathbb{R}^D si existe un conjunto abierto y conexo $S \subset \mathbb{R}^D$ y $\varphi : S \rightarrow \mathbb{R}^D$ una transformación isométrica⁶² tal que $\varphi(\overline{S}) = \mathcal{M}$. Como se mencionó con anterioridad, se espera que $d \ll D$, pero no es necesario.

Definición 3.6.4.4 (Convergencia de $D_{Q,\alpha}$, (Groisman, Jonckheere y Sapienza, 2019, Teorema 2.7)):

Asuma que \mathcal{M} es una variedad C^1 d -dimensional isométrica embebida en \mathbb{R}^D y $f : M \rightarrow R_+$ es una función de densidad de probabilidad continua. Sea $Q_n = \{q_1, \dots, q_n\}$ un conjunto de elementos aleatorios independientes con densidad común f . Entonces, para $\alpha > 1$ y $x, y \in M$ tenemos:

$$\lim_{n \rightarrow \infty} n^\beta D_{Q_n,\alpha}(x, y) = \mu D_{f,\beta}(x, y) \text{ casi seguramente.} \quad (80)$$

Aquí,

- $\beta = (\alpha - 1)/d$,
- μ es una constante que depende únicamente de α y d y
- la minimización se realiza sobre todas las curvas rectificables $\gamma \subset M$ que comienzan en x y terminan en y .

Observación: El factor de escala $\beta = \frac{\alpha-1}{d}$ depende de la dimensión intrínseca d de la variedad, y no de la dimensión D del espacio ambiente.

La distancia muestral de Fermat $D_{Q,\alpha}$:

⁶¹una métrica tal que la distancia puede ser nula entre puntos no-idénticos $\exists a \neq b : d(a, b) = 0$

⁶²Que preserva las métricas o distancias; del griego «isos» (igual) y «metron» (medida)

- se puede aproximar a partir de una muestra «lo suficientemente grande»
- sin conocer ni la variedad \mathcal{M} ni su dimensión intrínseca; además
- tiene garantías de convergencia a una distancia basada en densidad (DBD) «macroscópica» (la distancia de Fermat «a secas» $\mathcal{D},(f,\beta)$) y
- por definición, aprende «a la vez» la geometría del dominio y la densidad de la variable aleatoria objetivo sobre éste.

Es decir, que pareceríamos haber conseguido la pieza faltante para nuestro clasificador en variedades *desconocidas* y estaríamos en condiciones de proponer un algoritmo de clasificación que reúna todos los cabos del tejido teórico hasta aquí desplegado.

Nobleza obliga, hemos de mencionar que los trabajos de (McKenzie y Damelin, 2019; Little, McKenzie y Murphy, 2021), contemporáneos a Groisman et al, también consideran lo que ellos llaman «distancias de caminos mínimos pesadas por potencias»⁶³, y las aplican no a problemas de clasificación, sino de *clustering*⁶⁴. Hay algunas diferencias en la minucia del tratamiento⁶⁵, mas no así en la sustancia, por lo cual pasaremos directamente a la próxima sección.

4 Propuesta Original

Al comienzo de este sendero teórico nos preguntamos: ¿es posible mejorar un algoritmo de clasificación reemplazando la distancia euclídea por una aprendida de los datos? Habiendo explorado el área en profundidad, entendemos que sí pareciera ser posible, y en particular la distancia muestral de Fermat es un buen candidato de reemplazo.

Para saldar la cuestión, nos propusimos:

1. Implementar un clasificador basado en estimación de densidad por núcleos como el de [Definición 3.4.3.1](#) (Loubes y Pelletier, 2008), que llamaremos «KDC». Además,
2. Implementar un estimador de densidad por núcleos basado en la distancia de Fermat, a fines de poder comparar la *performance* de KDC con distancia euclídea y de Fermat.

Nótese que el clasificador de k -vecinos más cercanos de [Definición 3.1.1.2](#) (k -NN, Ecuación 9), tiene un pariente cercano, ε -NN

⁶³«power-weighted shortest-path distances» o PWSPDs por sus siglas en inglés

⁶⁴de identificación de grupos en datos no etiquetados

⁶⁵En particular, la distancias microscópica que plantean Little et al no es la suma de las aristas pesadas por $q = \alpha$ como hacen Bijral et al y Groisman et al, sino la raíz α -ésima de tal suma, en una especie de reversión de la distancia de Minkowski. Además, el contexto de *clustering* los lleva a considerar una muestra compuesta de elementos provenientes de variedad disjuntas, una representando a cada *cluster*.

Definición 4.1 (clasificador de ε -vecinos-más-cercanos): Sean $B_{\varepsilon(x)}$ una bola normal de radio ε centrada en x , y $\mathcal{N}_\varepsilon(x) = \mathbf{X} \cap B_{\varepsilon(x)}$ el ε -vencindario de x . El clasificador de ε -vecinos-más-cercanos $\varepsilon-NN$ le asignará a x la clase más frecuente entre la de sus vecinos $y \in \mathcal{N}_\varepsilon(x)$

Ecuación 9 es esencialmente equivalente a KDC con un núcleo «rectangular», $k(t) = \frac{\mathbb{1}(d(x,t) < \varepsilon)}{\varepsilon}$, pero su implementación es considerablemente más sencilla. Para comprender más cabalmente el efecto de la distancia de Fermat en *la tarea de clasificación*, y no solamente en *cierto* algoritmo de clasificación, nos propusimos también

3. Implementar un clasificador cual [Definición 3.1.1.2](#), pero con distancia muestral de Fermat en lugar de euclídea.

4.1 Evaluación

Nos interesa conocer en qué circunstancias, si es que hay alguna, la distancia muestral de Fermat provee ventajas a la hora de clasificar por sobre la distancia euclídea. Además, en caso de existir, quisiéramos en la medida de lo posible comprender por qué (o por qué no) es que tal ventaja existe. A nuestro entender resulta imposible hacer declaraciones demasiado generales al respecto de la capacidad del clasificador: la cantidad de *datasets* posibles, junto con sus *configuraciones de evaluación* es tan densamente infinita como lo permite la imaginación del evaluador. Con un ánimo exploratorio, nos proponemos explorar la *performance* de nuestros clasificadores basados en distancia muestral de Fermat en algunas *tareas* puntuales.

4.1.1 Métricas de *performance*

En tareas de clasificación, la métrica más habitual es la *exactitud*⁶⁶

Definición 4.1.1.1 (exactitud): Sean $(\mathbf{X}, \mathbf{g}) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$ una matriz de n observaciones de p atributos y sus clases asociadas. Sea además $\hat{\mathbf{g}} = \hat{G}(\mathbf{X})$ las predicciones de clase resultado de una regla de clasificación \hat{G} . La *exactitud* (exac) de \hat{G} en \mathbf{X} se define como la proporción de coincidencias con las clases verdaderas \mathbf{g} :

$$\text{exac}(\hat{G} | \mathbf{X}) = n^{-1} \sum_{i=1}^n \mathbb{1}(\hat{g}_i = g_i) \quad (81)$$

La exactitud está bien definida para cualquier clasificador que provea una regla *dura* de clasificación. Ahora bien, cuando un clasificador

⁶⁶Más conocida por su nombre en inglés, *accuracy*.

provee una regla suave, la exactitud como métrica « pierde información »: dos clasificadores binarios que asignen respectivamente 0.51 y 1.0 de probabilidad de pertenecer a la clase correcta a todas las observaciones tendrán la misma exactitud, 100%, aunque el segundo es a las claras mejor. A la inversa, cuando un clasificador erra al asignar la clase: ¿lo hace con absoluta confianza, asignando una alta probabilidad a la clase equivocada, o con cierta incertidumbre, repartiendo la masa de probabilidad entre varias clases que considera factibles?

Una métrica natural para evaluar una regla de clasificación suave, es la *verosimilitud* (y su logaritmo) de las predicciones.

Definición 4.1.1.2 (verosimilitud): Sean $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$ una matriz de n observaciones de p atributos y sus clases asociadas. Sea además $\hat{\mathbf{Y}} = \hat{G}(\mathbf{X}) \in \mathbb{R}^{n \times k}$ la matriz de probabilidades de clase resultado de una regla suave de clasificación \hat{G} . La *verosimilitud* (vero) de \hat{G} en \mathbf{X} se define como la probabilidad conjunta que asigna \hat{G} a las clases verdaderas \mathbf{y} :

$$L(\hat{G}) = \text{vero}(\hat{G} | \mathbf{X}) = \Pr(\hat{\mathbf{y}} = \mathbf{y}) = \prod_{i=1}^n \Pr(\hat{y}_i = y_i) = \prod_{i=1}^n \hat{\mathbf{Y}}_{(i,y_i)}$$

Por conveniencia, se suele considerar la *log-verosimilitud promedio*,

$$\ell(\hat{G}) = n^{-1} \log(L(\hat{G})) = n^{-1} \sum_{i=1}^n \log(\hat{\mathbf{Y}}_{(i,y_i)}) \quad (83)$$

La verosimilitud de una muestra varía en $[0, 1]$ y su log-verosimilitud, en $(-\infty, 0]$, pero como métrica esta sólo se vuelve comprensible *relativa a otros clasificadores*. Una forma de «normalizar» la log-verosimilitud, se debe a (McFadden, 1974).

Definición 4.1.1.3 (R^2 de McFadden): Sea \hat{G}_0 el clasificador «nulo», que asigna a cada observación y posible clase, la frecuencia empírica de clase encontrada en la muestra de entrenamiento $\mathbf{X}_{\text{train}}$. Para todo clasificador suave \hat{G} , definimos el R^2 de McFadden como

$$R^2(\hat{G} | \mathbf{X}) = 1 - \frac{\ell(\hat{G})}{\ell(\hat{G}_0)} \quad (84)$$

Observación: $R^2(\hat{G}_0) = 0$. A su vez, para un clasificador perfecto \hat{G}^* que otorgue toda la masa de probabilidad a la clase correcta, tendrá $L(\hat{G}^*) = 1$ y log-verosimilitud igual a 0, de manera que $R^2(\hat{G}^*) = 1 - 0 = 1$.

Sin embargo, un clasificador *peor* que \hat{G}_0 en tanto asigne bajas probabilidades a las clases correctas, puede tener un R^2 infinitamente negativo.

Visto y considerando que tanto \mathcal{F} -KDC como \mathcal{F} -kNN son clasificadores suaves, evaluaremos su comportamiento en comparación con ambas métricas, la exactitud y el R^2 de McFadden⁶⁷

4.1.2 Algoritmos de referencia

Además de medir qué (des)ventajas otorga el uso de una distancia aprendida de los datos en la tarea de clasificación, quisiéramos entender (a) por qué sucede, y (b) si tal (des)ventaja es significativa en el amplio abanico de algoritmos disponibles. Pírrica victoria sería mejorar con la distancia de Fermat la *performance* de cierto algoritmo, para encontrar que aún con la mejora, el algoritmo no es competitivo en la tarea de referencia.

Consideraremos a modo de referencia los siguientes algoritmos:

- Naive Bayes Gaussiano (`GNB`),
- Regresión Logística (`LR`) y
- Clasificador de Soporte Vectorial (`svc`)

Esta elección no pretende ser exhaustiva, sino que responde a un «capricho informado» del investigador. `GNB` es una elección natural, ya que es la simplificación que surge de asumir independencia en las dimensiones de X para KDE multivariado ([Definición 3.3.2.1](#)), y se puede computar para grandes conjuntos de datos en muy poco tiempo. `LR` es «el» método para clasificación binaria, y su extensión a múltiples clases no es particularmente compleja: para que sea mínimamente valioso un nuevo algoritmo, necesita ser al menos tan bueno como `LR`, que tiene ya más de 65 años en el campo (TODO REF bliss1935, cox1958). Por último, fue nuestro deseo incorporar algún método más cercano al estado del arte. A tal fin, consideramos incorporar alguna red neuronal (TODO REF), un método de *boosting* (TODO REF) y el antedicho clasificador de soporte vectorial, `svc`. Finalmente, por la sencillez de su implementación dentro del marco elegido⁶⁸ y por la calidad de los resultados obtenidos, decidimos dejar fuera las redes neuronales, pero introdujimos `svc`, en dos variantes: con núcleos (*kernels*) lineales y RBF; y `GBT`.

4.1.3 Metodología

La unidad de evaluación de los algoritmos a considerar es una `Tarea`, que se compone de:

- un *diccionario de algoritmos* a evaluar en condiciones idénticas, definidas por
- un *dataset* con el conjunto de N observaciones en D dimensiones repartidas en K clases, (\mathbf{X}, \mathbf{g}) ,

⁶⁷de aquí en más, R^2 para abbreviar

⁶⁸Utilizamos *scikit-learn*, un poderoso y extensible paquete para tareas de aprendizaje automático en Python

- un *split de evaluación* $r \in (0, 1)$, que determina las proporciones de los datos a usar durante el entrenamiento ($1 - r$) y la evaluación (r), junto con
- una *semilla* $s \in [2^{32}]$ que alimenta el generador de números aleatorios y define determinísticamente cómo realizar la división antedicha.

4.1.4 Entrenamiento de los algoritmos

La especificación completa de un clasificador, requiere, además de la elección del algoritmo, la especificación de sus *hiperparámetros*, de manera tal de optimizar su rendimiento bajo ciertas condiciones de evaluación. Para ello, se definió de antemano para cada clasificador una *grilla* de hiperparámetros: durante el proceso de entrenamiento, la elección de los «mejores» hiperparámetros se efectuó maximizando la log-verosimilitud [Definición 4.1.1.2](#) para los clasificadores suaves, y la exactitud [Definición 4.1.1.1](#) para los duros⁶⁹ con una búsqueda exhaustiva por convalidación cruzada de 5 pliegos⁷⁰ sobre la grilla entera.

4.1.5 Estimación de la variabilidad en la *performance* reportada

En última instancia, cualquier métrica evaluada, no es otra cosa que un *estadístico* que representa la «calidad» del clasificador en la Tarea a mano. A fines de conocer no sólo su estimación puntual sino también darnos una idea de la variabilidad de su performance, para cada dataset y colección de algoritmos, se entrenaron y evaluaron 25 tareas idénticas salvo por la semilla s , que luego se usaron para estimar la varianza y el desvío estándar en la exactitud ([Definición 4.1.1.1](#)) y el pseudo- R^2 ([Definición 4.1.1.3](#)).

Cuando el conjunto de datos proviene del mundo real y por lo tanto *preexiste a nuestro trabajo*, las 25 semillas s_1, \dots, s_{25} fueron utilizadas para definir el split de entrenamiento/evaluación. Por el contrario, cuando el conjunto de datos fue generado sintéticamente, las semillas se utilizaron para generar 25 versiones distintas pero perfectamente replicables del dataset, y en todas se utilizó una misma semilla maestra s^* para definir el split de evaluación.

⁶⁹Entre los mencionados, el único clasificador duro es `svc`. Técnicamente es posible entrenar un clasificador suave a partir de uno duro con un *segundo* estimador que toma como *input* el resultado «crudo» del clasificador duro y da como *output* una probabilidad calibrada (cf. [Calibración](#) en la documentación de `scikit-learn` TODO citar `scikit-learn`), pero es un proceso computacionalmente costoso.

⁷⁰Conocida en inglés como *Grid Search 5-fold Cross-Validation*

5 Resultados

5.1 Chequeo de sanidad: blobs

Antes de considerar ningún tipo de sofisticación, comenzamos asegurándonos que en condiciones benignas, nuestros clasificadores funcionan correctamente. La

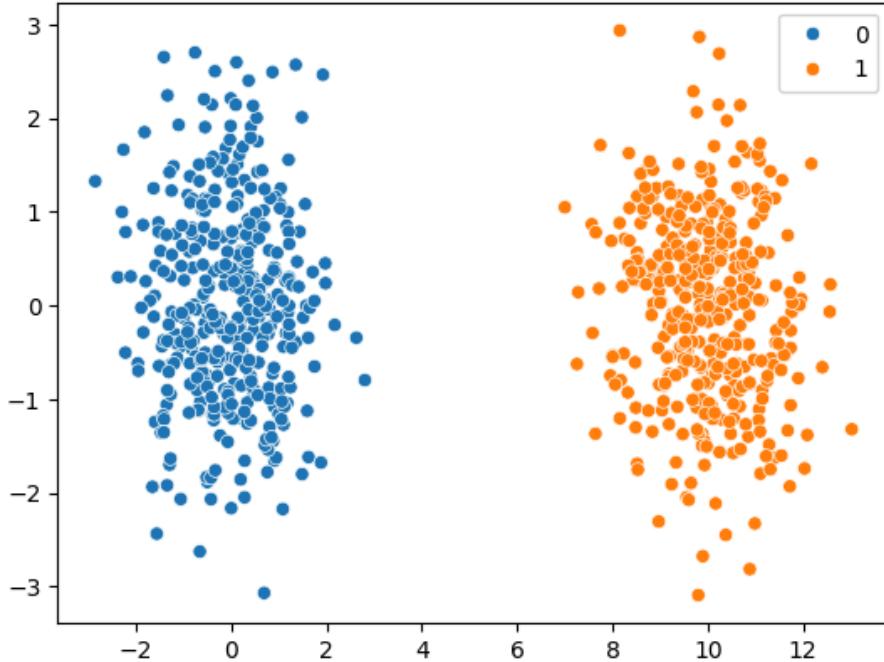


Figura 15: `make_blobs(n_features=2, centers=((0, 0), (10, 0)), random_state=1984)`

En este ejemplo, $d_{\mathcal{M}} = d_x = 2$; $k = 2$; $n_1 = n_2 = 400$ tenemos dos clases perfectamente separables, con lo cual cualquier clasificador razonable debería alcanzar $\text{exac} \approx 1$, $\ell \approx 0$, $R^2 \approx 1$. La evaluación de nuestros clasificadores resulta ser:

	ℓ	R^2	exac
\mathcal{F} -KDC	-0.0	1.0	1.0
KDC	-0.0	1.0	1.0
GNB	-0.0	1.0	1.0
k -NN	-0.0	1.0	1.0
\mathcal{F} -kNN	-0.0	1.0	1.0
LR	-4.5678	0.9589	1.0
SVC			1.0
LSVC			1.0
base	-111.1567	0.0	0.4625

Tabla 1: Resultados de entrenamiento en Figura 15

¡Excelentes noticias! Todos los clasificadores bajo estudio tienen exactitud perfecta, y salvo por una ligeramente negativa ℓ para LR, el resto da exactamente 0. Pasemos entonces a algunos dataset mínimamente más complejos.

5.1.1 Datasets sintéticos baja dimensión

Consideremos ahora algunas curvas unidimensionales embebidas en \mathbb{R}^2 :

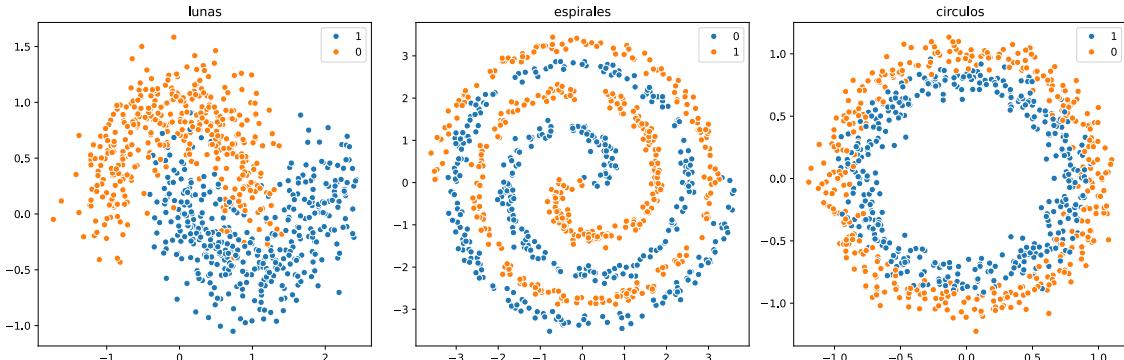


Figura 16: «Lunas», «Círculos» y «Espirales», con $d_x = 2$, $d_{\mathcal{M}} = 1$ y $s = 4107$

Resultará obvio al lector que los conjuntos de datos expuestos en Figura 16 no son exactamente variedades «1D» embebidas en «2D», sino que tienen un poco de «ruido blanco» agregado para incrementar la dificultad de la tarea.

Definición 5.1.1.1 (ruido blanco): Sea $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ una variable aleatoria tal que $E(X_i) = 0$, $\text{Var}(X_i) = \Sigma \forall i \in [d]$. Llamaremos «ruido blanco con escala Σ » a toda realización de X .

Veamos entonces cómo les fue a los contendientes, considerando primero la exactitud. Recordemos que para cada experimento se realizaron 25 repeticiones: en cada celda reportaremos la exactitud *promedio*, y a su lado entre paréntesis el error estándar cpte.:

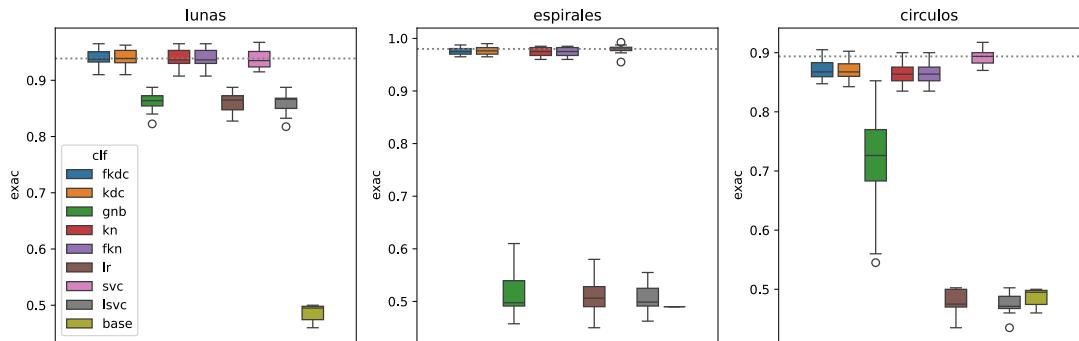


Figura 17: Boxplots con la distribución de dxactitud en las 25 repeticiones de cada experimento de Figura 16

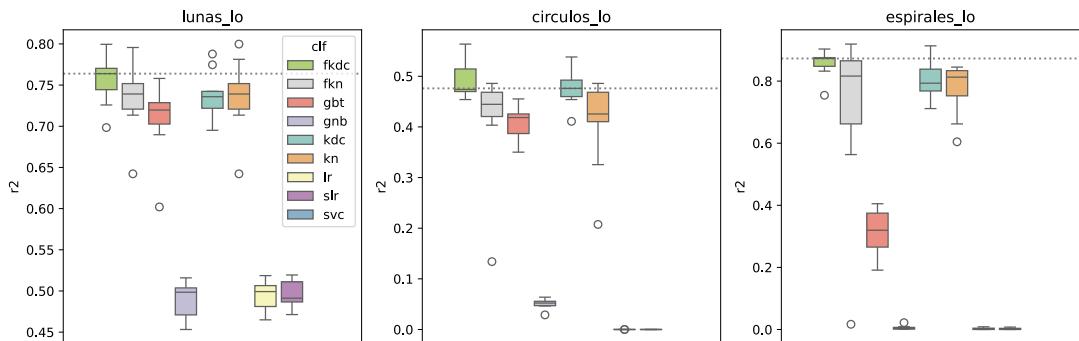


Figura 18: Boxplots con la distribución de dxactitud en las 25 repeticiones de cada experimento de Figura 16

dataset	circulos	circulos	espirales	espirales	lunas	lunas
	mean	std	mean	std	mean	std
clf						
base	48.61	1.42	49.0	0.0	48.61	1.42
fkdc	65.31	2.79	84.88	2.18	81.7	2.22
fkn	64.69	2.46	84.97	2.57	81.86	2.29
gnb	63.98	3.37	51.41	4.0	80.5	2.1
kdc	65.38	2.98	85.09	2.21	81.8	2.44
kn	64.69	2.46	84.97	2.57	81.86	2.29
lr	48.12	1.74	51.88	3.01	80.25	2.17
lsvc	47.86	1.79	50.55	2.68	80.52	1.99
svc	67.8	2.6	86.8	1.82	82.02	2.44

Tabla 2: «mi caption, bo».

KDC (en sus dos variantes), KNN y SVC (con kernel RBF) parecieran ser los métodos más competitivos, con mínimas diferencias de performance entre sí: sólo en «círculos» se observa un ligero ordenamiento de los métodos, svc > KDC > k - NN, aunque la performance mediana de svc está dentro de «los bigotes» de todos los métodos antedichos. La tarea «lunas» pareciera ser la más fácil de todas, en la que hasta una regresión logística sin modelado alguno es adecuada. Para «espirales» y «círculos», GNB, LR y LSVC no logran performar significativamente mejor que el clasificador base.

Definición 5.1.1.2 (clasificador base):

¿Cómo se comparan los métodos en términos de la log-verosimilitud y el R^2 ?

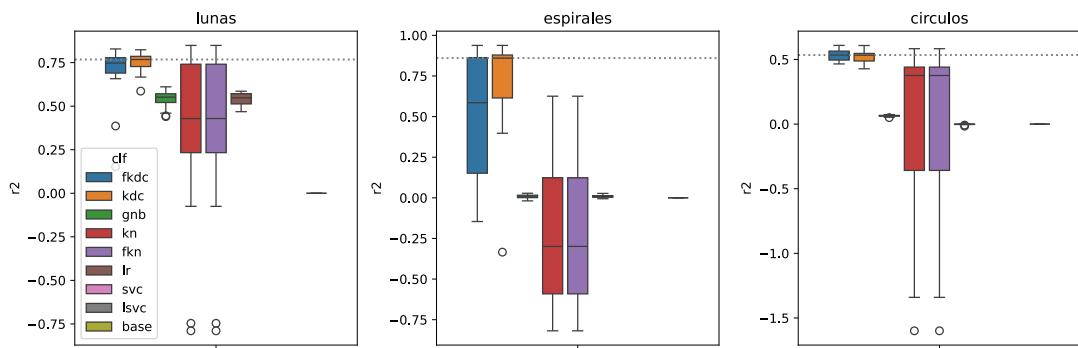


Figura 19: Boxplots con la distribución de R^2 en las 25 repeticiones de cada experimento.

dataset	circulos	circulos	espirales	espirales	lunas	lunas
	mean	std	mean	std	mean	std
clf						
base	0.0	0.0	0.0	0.0	0.0	0.0
fkdc	0.077	0.061	-0.067	1.596	0.347	0.133
fkn	-0.082	0.465	-3.081	3.123	0.299	0.153
gnb	0.05	0.009	0.007	0.012	0.384	0.045
kdc	0.094	0.026	-0.556	2.374	0.374	0.06
kn	-0.082	0.465	-3.081	3.123	0.299	0.153
lr	-0.002	0.004	0.009	0.009	0.364	0.044

Tabla 3: «mi caption, bo-bo».

Como los métodos basados en máquinas de soporte vectorial resultan en clasificadores *duros* (clasificador-duro), no es posible analizar la log-verosimilitud u otras métricas derivadas. De entre los dos métodos con exactitud similar a esos, es notoriamente mejor el R^2 que alcanzan ambos KDC. A primera vista, se ve que la dispersión de la métrica es considerable, pues las «cajas» del rango intercuartil son bastante amplias, y aún así se observan *outliers*. En las tres tareas, los clasificadores de estimación de densidad por núcleos tienen las cajas más angostas y los bigotes más cortos, con KDC mostrando una dispersión menor o igual que \mathcal{F} -KDC. En la Figura 17, observamos que la exactitud de los métodos de k vecinos más cercanos era muy similar a la de KDC y svc, sin embargo en términos de R^2 ,

- en el dataset de «espirales» el R^2 promedio y mediano son *negativos*, y
- en el de «círculos», aunque la locación⁷¹ es positiva, la distribución tiene una pesada cola a izquierda, que entra de lleno en los negativos.

En otras palabras, pareciera ser que aunque la exactitud de los métodos basados en vecinos más cercanos es buena, cuando clasifican *mal*, lo hacen con *alta seguridad*, lo que resulta en un pésimo R^2 .

En esta terna inicial de *datasets*, obtenemos unos resultados aceptables:

- Observamos que el clasificador de (Loubes y Pelletier, 2008) es competitivo (es decir que no sólo Loubes y Pelletier propusieron),
- aunque la distancia de Fermat muestral no parece mejorar significativamente la exactitud de los calsificadores en ella basados.

Que la bondad de los clasificadores *no empeore* con el uso de $D_{Q,\alpha}$ en lugar de $\|\cdot\|_2$ es importante. Por una parte, cuando $\alpha = 1$ y $n \rightarrow \infty$, $D_{Q,\alpha} \rightarrow \mathcal{D}_{f,\beta} = \|\cdot\|_2$, con lo cual \mathcal{F} -KDC debería performar al

⁷¹Entendemos tanto al *promedio* o *media* y la *mediana* como *medidas de locación*

menos tan bien como KDC cuando la grilla de hiperparámetros en la que lo entrenamos incluye a $\alpha = 1$. Sin embargo, el cómputo de $D_{Q,\alpha}$ es numéricamente bastante complejo, y bien podríamos haber encontrado dificultades computacionales⁷².

5.1.1.1 Comparación entre KDC y \mathcal{F} -KDC para ("circulos", 4479)

Concentrémonos en un segundo en una corrida específica de un experimento particular. Por caso, tomemos el dataset «circulos», con la semilla 4479. Los parámetros óptimos de \mathcal{F} -KDC resultaron ser (`alpha: 1.1875, bandwidth: 0.0562`) , mientras que los de KDC fueron (`bandwidth: 0.1202`). Los anchos de banda son diferentes, y el α óptimo encontrado por \mathcal{F} -KDC es distinto de 1. Sin embargo, la exactitud de \mathcal{F} -KDC fue 0.885, y la de KDC, 0.88, prácticamente idénticas⁷³. ¿Por qué? ¿Será que los algoritmos no son demasiado sensibles a los hiperparámetros elegidos?

Recordemos que la elección de hiperparámetros se hizo con una búsqueda exhaustiva por convalidación cruzada de 5 pliegos. Por lo tanto, *durante el entrenamiento* se generaron suficientes datos como para graficar la exactitud promedio en los pliegos, en función de (α, h) . A esta función de los hiperparámetros a una función de pérdida⁷⁴ se la suele denominar *superficie de pérdida*.

⁷²De hecho, hubo montones de ellas, cuya resolución progresiva dio lugar a la pequeña librería que acompaña esta tesis y documentamos en el anexo. A mi entender, ningún error de cálculo persiste en el producto final

⁷³Con 400 observaciones para evaluación, dichos porcentajes representan 352 y 354 observaciones correctamente clasificadas, resp.

⁷⁴En realidad, la exactitud es un «score» o puntaje - mientras más alto mejor-, pero el negativo de cualquier puntaje es una pérdida - mientras más bajo, mejor.

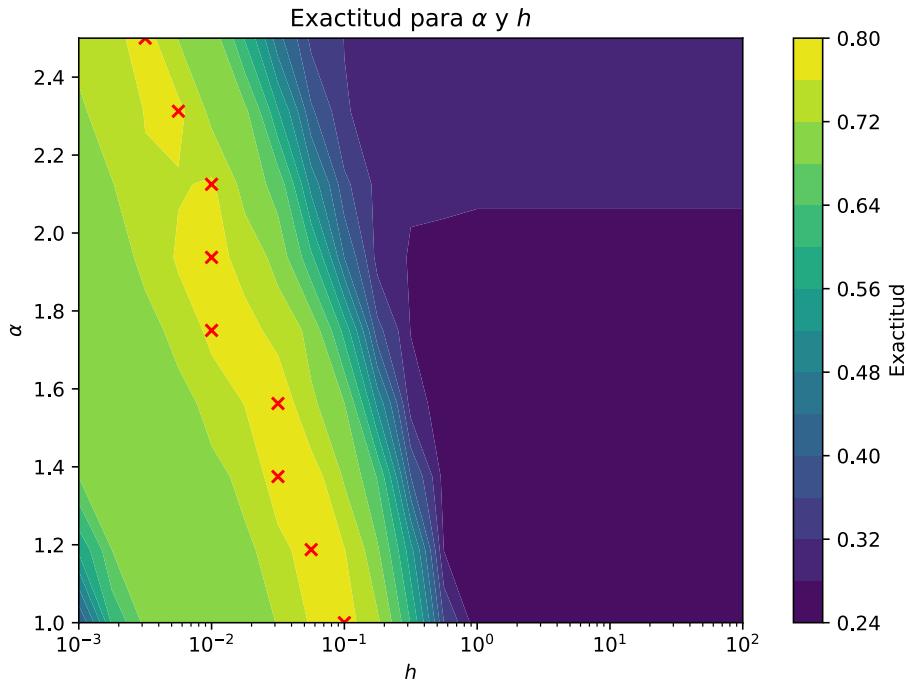


Figura 20: Exactitud promedio en entrenamiento para la corrida ("[circulos](#)", [4479](#)). Las cruces rojas indican la ventana h óptima para cada α .

Nótese que la región amarilla, que representa los máximos puntajes durante el entrenamiento, se extiende diagonalmente a través de todos los valores de α . Es decir, no hay un *par* de hiperparámetros óptimos (α^*, h^*) , sino que fijando α , siempre parecería existir un(os) $h^*(\alpha)$ que alcanza (o aproxima) la máxima exactitud *possible* con el método en el dataset. En este ejemplo en particular, hasta parecería ser que una relación log-lineal captura bastante bien el fenómeno, $\log(h^*) \propto \alpha$. En particular, entonces, $\text{exac}(h^*(1), 1) \approx \text{exac}(h^*, \alpha^*)$, y se entiende que el algoritmo \mathcal{F} -KDC, que agrega el hiperparámetro α a KDC no mejore significativamente su exactitud.

Ahora bien, esto es sólo en *un* dataset, con *una* semilla específica. ¿Se replicará el fenómeno en los otros datasets estudiados? Y si tomásemos datasets con otras características?

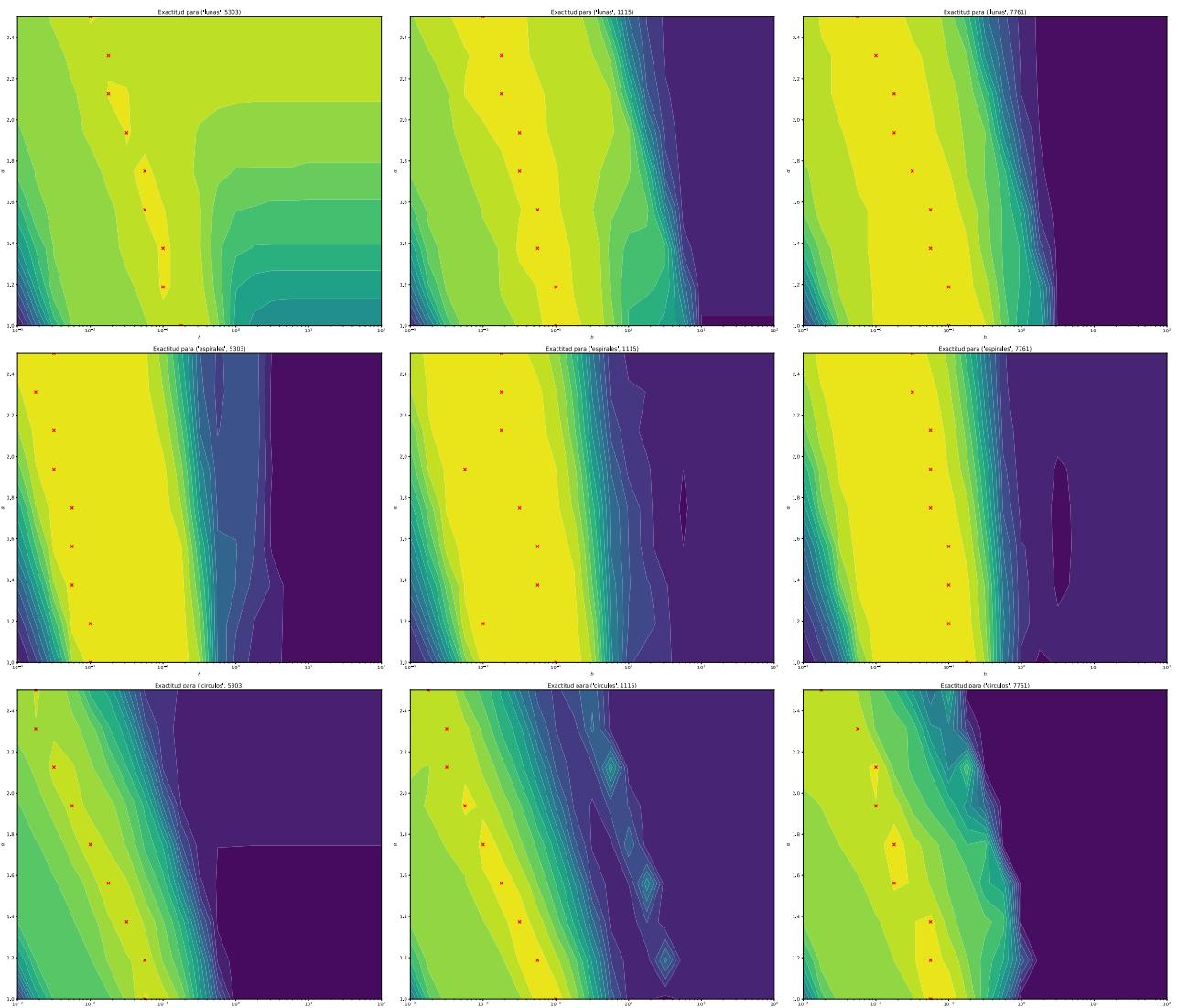
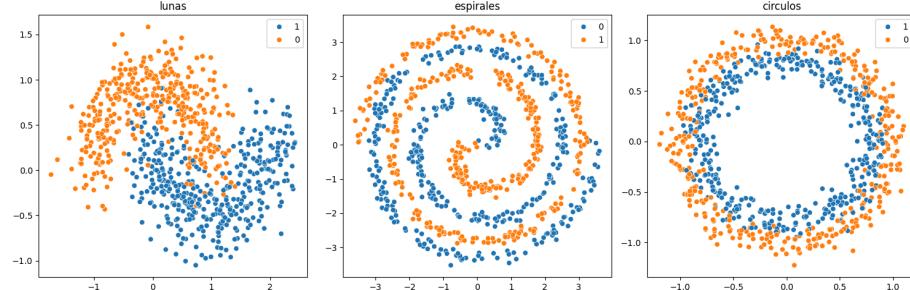


Figura 21: It does replicate

Antes de avanzar hacia el siguiente conjunto de datos, una pregunta más: ¿qué sucede si aumentamos el nivel de ruido? Es decir, mantenemos los dataset hasta aquí considerados, pero subimos Σ de [Definición 5.1.1.1](#)?

5.1.2 2D, 2 clases: excelente R^2 con exactitud competitiva

5.1.3 Con Bajo Ruido



lunas_lo (acc: 93.66%)

	delta_acc	r2
clf		
fkdc	-0.29	75.58
fkn	-0.04	74.43
kn	0.00	74.30
kdc	-0.33	73.42
gbt	-0.67	70.14
lr	-8.71	49.90
slr	-7.93	49.62
gnb	-10.12	48.83
base	-45.12	0.00
svc	-0.98	NaN

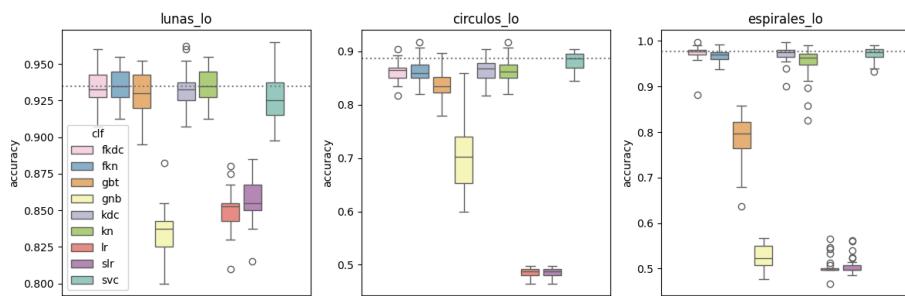
circulos_lo (acc: 88.15%)

	delta_acc	r2
clf		
fkdc	-1.91	49.25
kdc	-1.71	48.58
fkn	-1.82	45.13
kn	-1.82	44.92
gbt	-4.64	43.42
gnb	-17.43	5.13
base	-39.61	0.00
lr	-39.61	-0.00
slr	-39.61	-0.00
svc	0.00	NaN

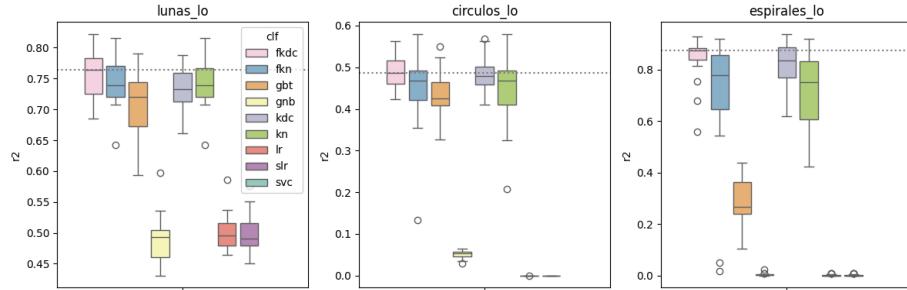
espirales_lo (acc: 97.27%)

	delta_acc	r2
clf		
fkdc	0.00	84.66
kdc	-0.14	82.36
kn	-2.29	71.47
fkn	-0.50	70.64
gbt	-18.68	28.82
gnb	-44.59	0.35
lr	-46.74	0.17
slr	-46.56	0.17
base	-47.52	0.00
svc	-0.18	NaN

5.1.4 Boxplot Accuracy



5.1.5 Boxplot R^2



5.1.6 Superposición de parámetros: α y h

- El uso de la distancia de Fermat muestral no hiere la performance, pero las mejoras son nulas o marginales. ¿Por qué?

Si recordamos $\hat{f}_{K,N}$ según Loubes & Pelletier, al núcleo K se lo evalúa sobre

$$\frac{d(x_0, X_i)}{h}, \quad d = D_{Q_i, \alpha} \quad (85)$$

Lo que α afecta a \hat{f} vía d , también se puede conseguir vía h .

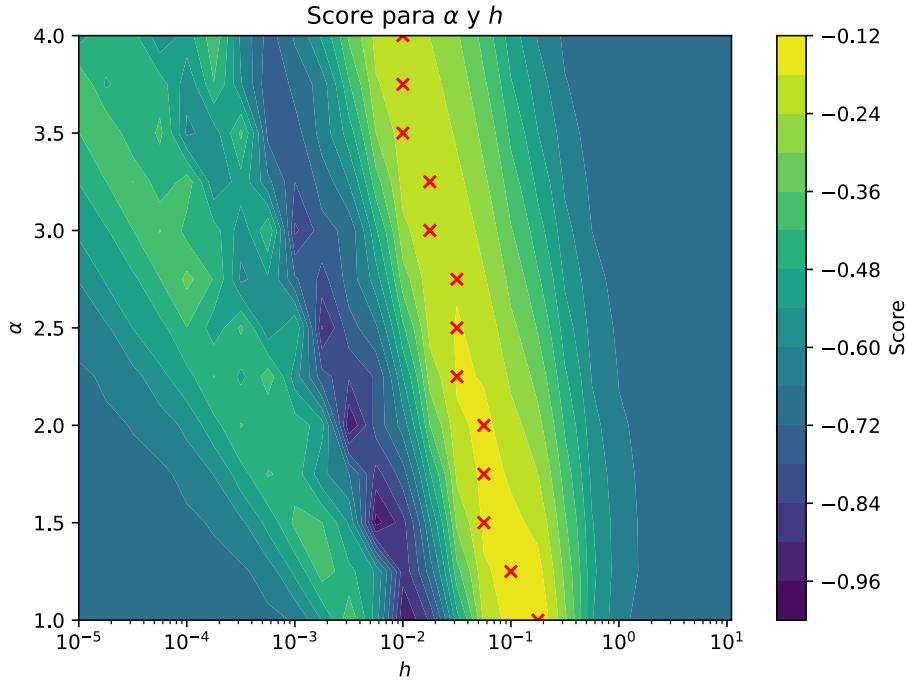
Si $D_{Q_i, \alpha} \propto \|\cdot\|$ (la distancia de fermat es proporcional a la euclídea), los efectos de α y h se «solapan»

... y sabemos que localmente, eso es cierto 😊

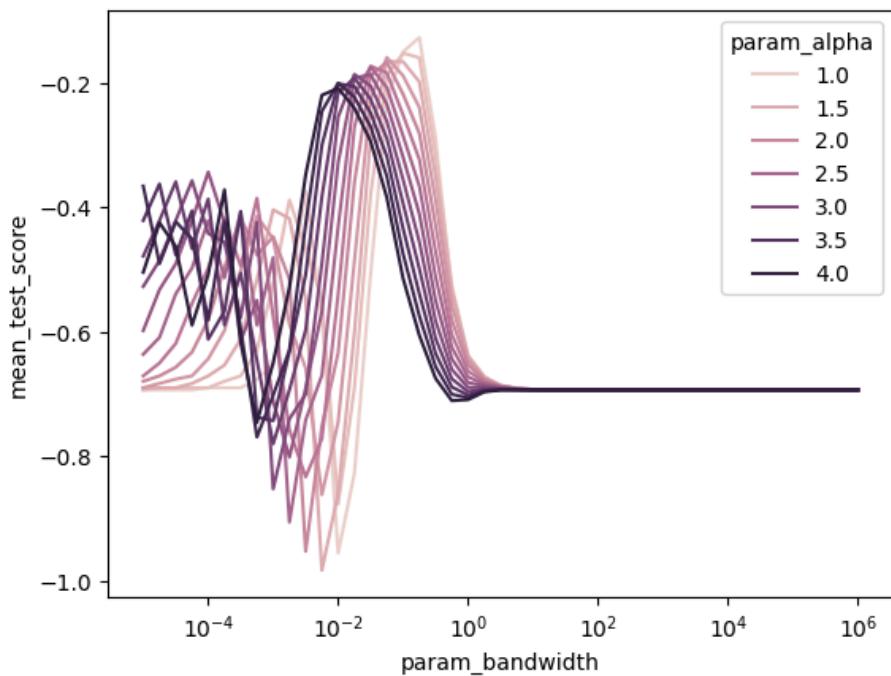
5.1.7 Parámetros óptimos para (F)KDC en espirales_lo

(fkdc, alpha)	(fkdc, bandwidth)	(kdc, bandwidth)	count
1.0	0.1000	0.1431	1
1.0	0.1778	0.1726	8
1.0	0.1778	0.2082	7
1.0	0.1778	0.2512	7
1.0	0.3162	0.3030	1
1.5	0.0032	0.2082	1

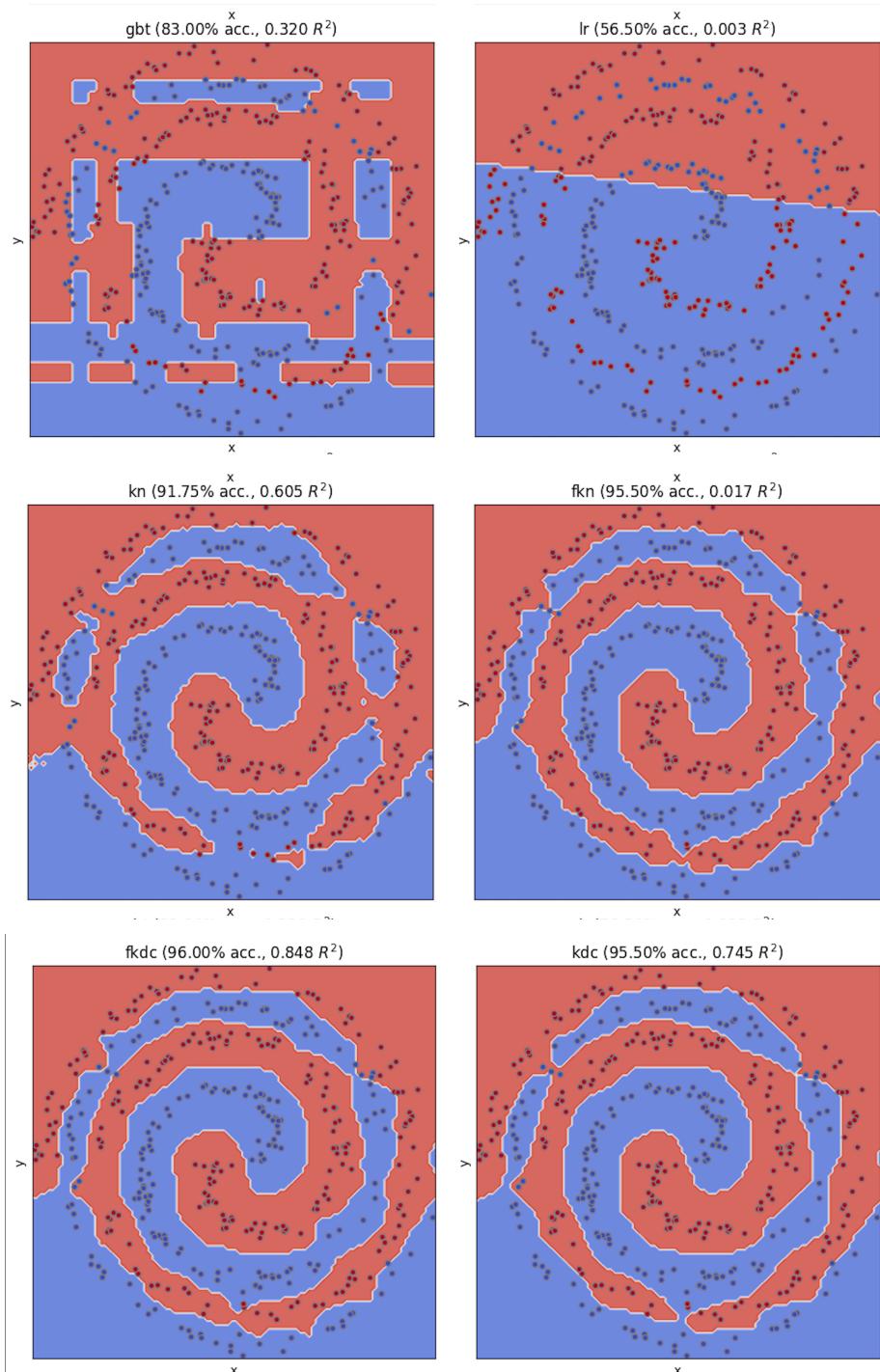
5.1.8 Superficies (o paisajes) de score para (espirales_lo, 1434)

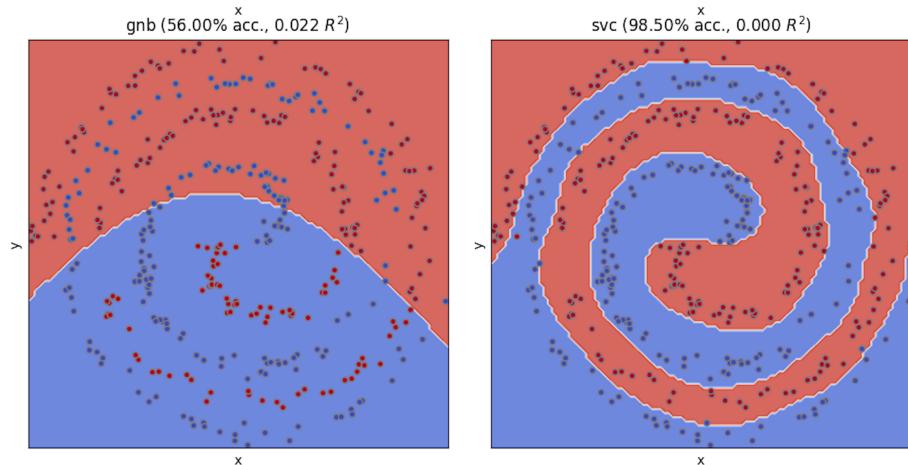


5.1.9 Alt-viz: Perfiles de pérdida para (espirales_lo, 1434)



5.1.10 Fronteras de decisión para `(espirales_lo, 1434)`





5.1.10.1 Efecto del ruido en la performance de clasificación

lunas_hi (acc: 81.44%)

	clf	delta_acc	r2
gbt	-0.21	38.14	
fkdc	-0.33	37.46	
fkn	0.00	36.57	
kn	-0.02	36.50	
kdc	-0.47	35.67	
slr	-1.21	33.75	
lr	-1.73	33.70	
gnb	-1.98	33.34	
base	-32.90	0.00	
svc	-1.18	NaN	

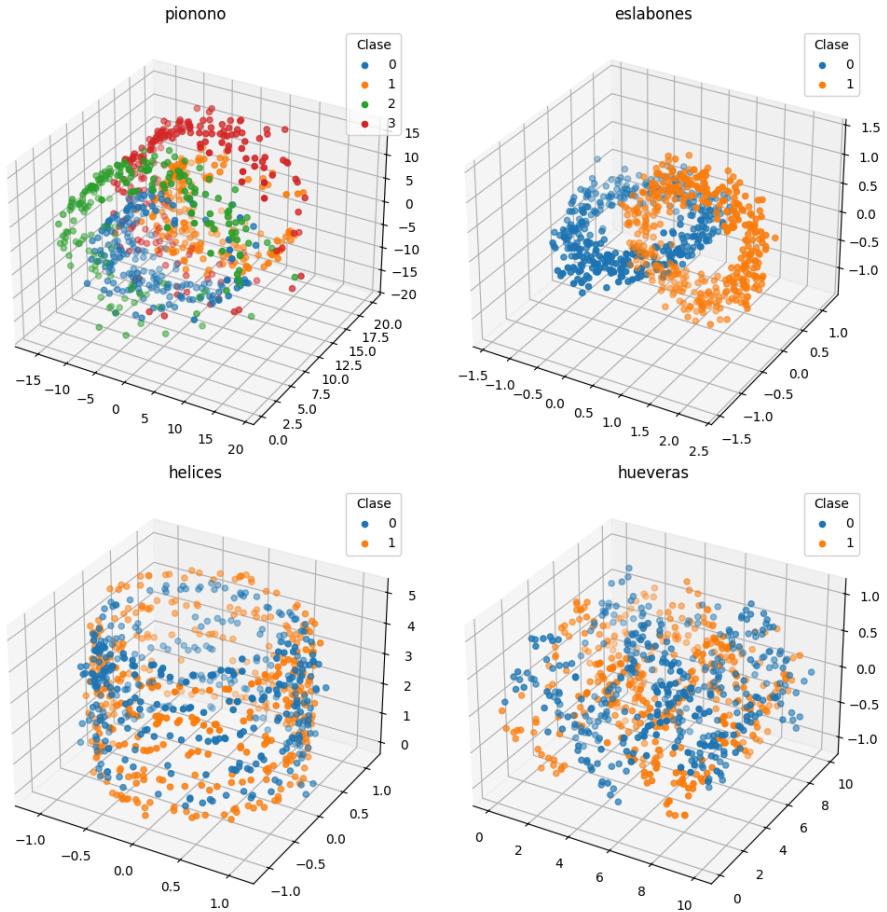
circulos_hi (acc: 65.93%)

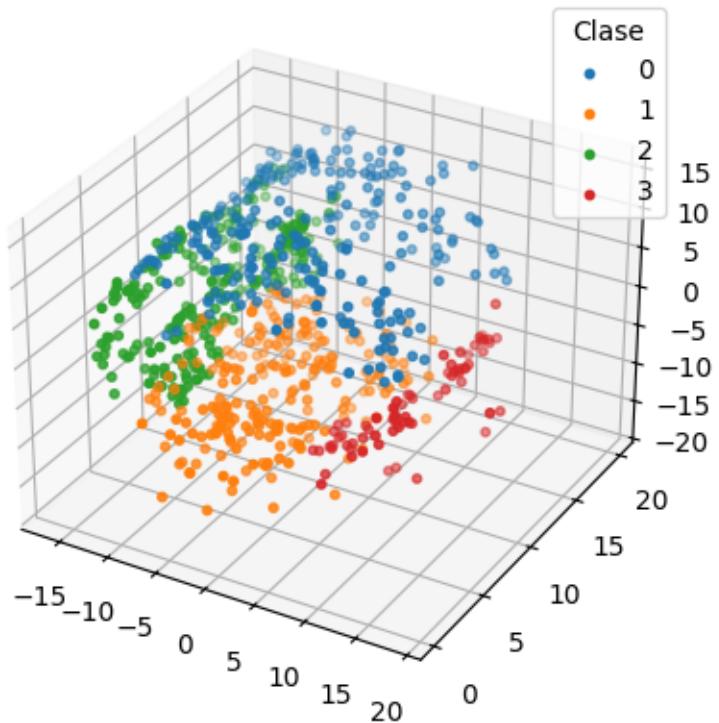
	clf	delta_acc	r2
gbt	-0.23	9.49	
fkdc	-3.82	6.98	
kdc	-4.78	5.54	
kn	-4.87	5.14	
fkn	-5.41	5.12	
gnb	-5.34	3.33	
base	-17.39	0.00	
lr	-17.39	-0.00	
slr	-17.39	-0.00	
svc	0.00	NaN	

espirales_hi (acc: 85.54%)

	clf	delta_acc	r2
fkdc	-1.86	44.41	
kdc	-1.98	43.45	
fkn	-2.57	40.36	
kn	-3.06	39.93	
gbt	-14.68	15.22	
gnb	-32.64	0.33	
lr	-34.95	0.17	
slr	-34.97	0.17	
base	-35.79	0.00	
svc	0.00	NaN	

5.1.11 3D, 2 clases + piononos





pionono_0 (acc: 94.19%)

	clf	delta_acc	r2
fkdc		-1.06	81.04
gbt		-2.37	81.02
kdc		-0.83	79.89
fkn		-2.65	72.55
kn		-3.57	70.62
gnb		-20.40	58.65
sir		-29.69	47.45
lr		-29.72	47.44
base		-71.44	0.00
svc		0.00	NaN

eslabones_0 (acc: 99.9%)

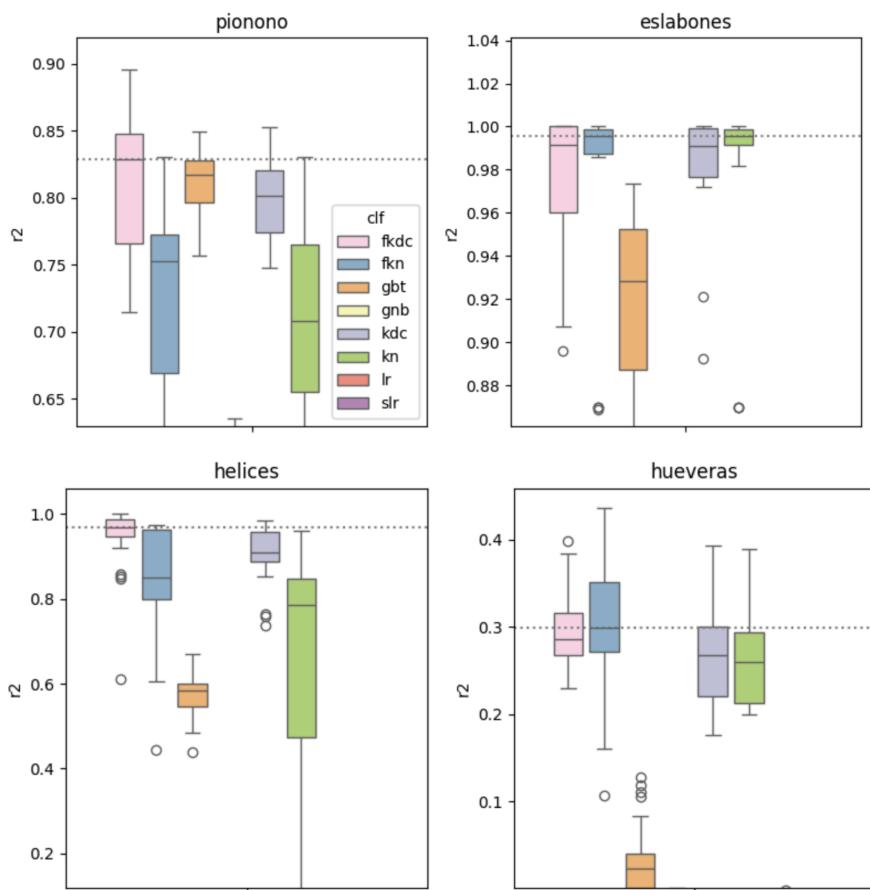
	clf	delta_acc	r2
kdc		-0.10	97.75
kn		-0.02	96.51
fkdc		-0.06	96.44
fkn		-0.02	96.04
gbt		-1.14	91.97
gnb		-10.68	75.16
lr		-33.32	22.04
sir		-33.30	21.87
base		-50.15	0.00
svc		0.00	NaN

helices_0 (acc: 99.16%)

	clf	delta_acc	r2
fkdc		0.00	94.41
kdc		-0.06	86.10
fkn		-0.76	83.45
kn		-4.90	62.81
gbt		-9.48	57.56
gnb		-49.29	0.01
base		-49.41	0.00
lr		-49.41	0.00
sir		-49.41	0.00
svc		-24.03	NaN

hueveras_0 (acc: 77.88%)

	clf	delta_acc	r2
fkn		-1.60	30.20
fkdc		-2.57	29.40
kdc		-3.36	26.62
kn		-2.71	26.43
gbt		-19.90	3.44
gnb		-27.43	0.01
base		-28.13	0.00
sir		-28.13	0.00
lr		-28.15	-0.01
svc		0.00	NaN



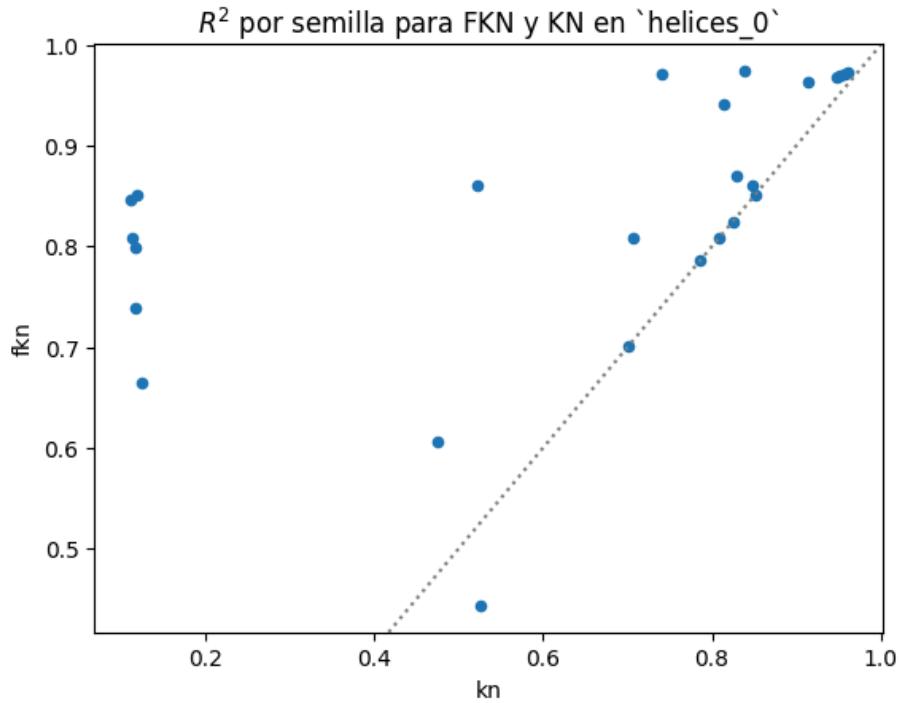
5.1.12 Parámetros óptimos para (F)KDC en helices_0

(fkdc, alpha)	(fkdc, bandwidth)	(kdc, bandwidth)	count
1.00	0.1000	0.1431	6
1.00	0.0100	0.1431	3
1.25	0.0056	0.1726	3
1.00	0.0100	0.1726	2
1.00	0.1000	0.1186	1
1.25	0.0056	0.1431	1
1.25	0.0056	0.2082	1
1.25	0.0100	0.1431	1
1.50	0.0056	0.1726	1
1.75	0.0032	0.1431	1
1.75	0.0032	0.1726	1
2.00	0.0032	0.1431	1
2.00	0.0032	0.1726	1
2.50	0.0010	0.2082	1
2.50	0.0018	0.1726	1

5.1.13 Microindiferencia, macrodiferencia

- En zonas con muchas observaciones (por tener alta f o alto N) sampleadas, la distancia de Fermat y la euclídea coinciden.
- «Localmente», siempre van a coincidir, aunque sea en un vecindario muy pequeño.
- Si el algoritmo de clasificación sólo depende de ese vecindario local para clasificar, no hay ganancia en la distancia de Fermat.
- ¡Pero tampoco hay pérdida si se elige mal `n_neighbors`! 🧐

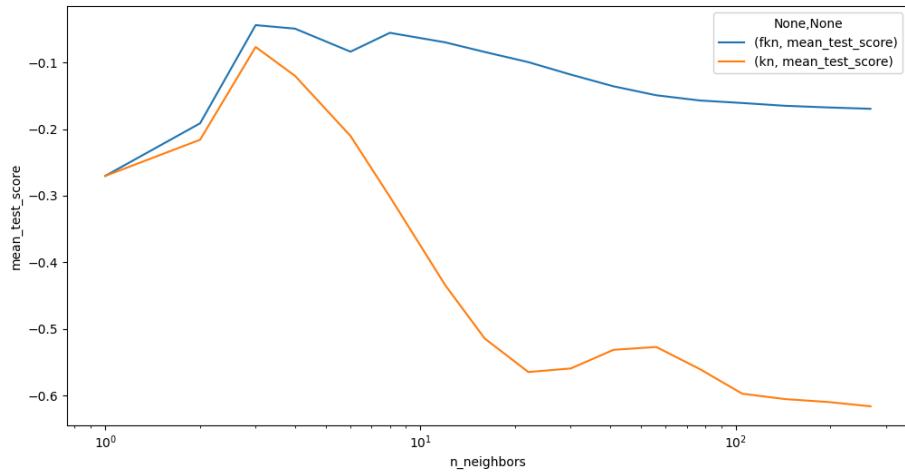
5.1.14 R^2 por semilla para (F)KN en `helices_0`



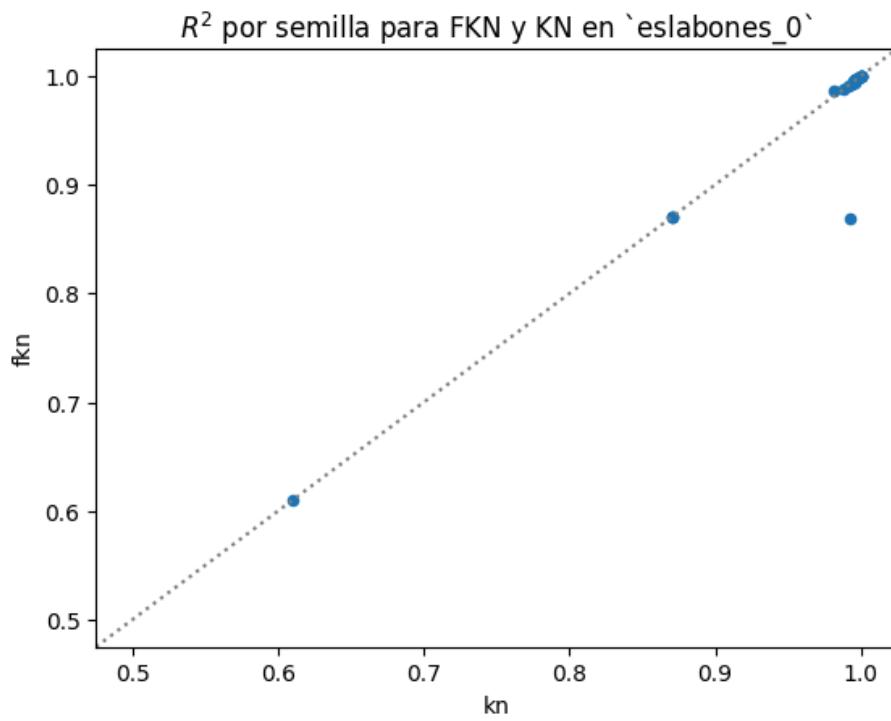
5.1.15 R^2 y α^* para (F)KN en `helices_0`, n_neighbors seleccionados

param_n_neighbors	fkn		kn	
	param_alpha	mean_test_score	mean_test_score	mean_test_score
1	3.50	-0.270327	-0.270327	-0.270327
3	1.75	-0.043833	-0.076631	-0.076631
12	4.00	-0.069771	-0.434716	-0.434716
41	4.00	-0.135744	-0.531532	-0.531532
144	4.00	-0.165027	-0.605601	-0.605601

5.1.16 Mejor R^2 para (F)KN en helices_0, en función de n_neighbors



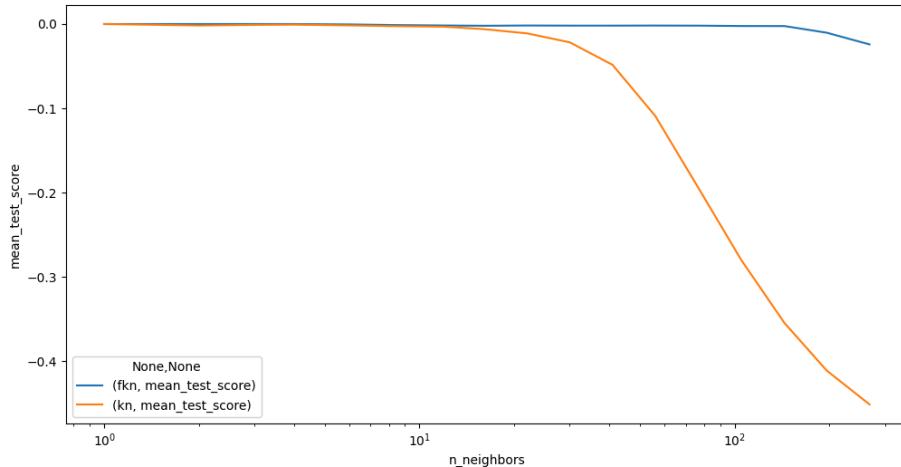
5.1.17 R^2 por semilla para (F)KN en eslabones_0



5.1.18 R^2 y α^* para (F)KN en eslabones_0, n_neighbors seleccionados

	fkn		kn	
	param_alpha	mean_test_score	mean_alpha	mean_test_score
param_n_neighbors				
1	1.00	-0.000	-0.000	-0.000
3	3.75	-0.000	-0.001	-0.001
12	3.75	-0.002	-0.003	-0.003
41	3.00	-0.002	-0.048	-0.048
144	4.00	-0.002	-0.355	-0.355

5.1.19 Mejor R^2 para (F)KN en eslabones_0, en función de n_neighbors



5.1.20 Otros datasets: 15D

pionono_12 (acc: 91.07%)	eslabones_12 (acc: 98.48%)	helices_12 (acc: 53.15%)	hueveras_12 (acc: 53.55%)
clf	clf	clf	clf
delta_acc r2	delta_acc r2	delta_acc r2	delta_acc r2
gbt 0.00 78.96	gbt 0.00 91.66	fkn -0.99 0.20	kn -0.87 0.20
gnb -20.78 54.30	gnb -8.67 75.24	kn -0.99 0.20	fkn -0.99 0.17
slr -28.42 42.42	fkdc -22.66 24.58	gnb -1.01 0.13	gnb -1.30 0.14
lr -28.56 42.22	fkn -21.98 24.43	base -3.40 0.00	lr -3.81 0.02
fkdc -46.30 10.18	kdc -22.40 24.32	lr -3.40 0.00	base -3.80 0.00
kdc -41.83 10.05	kn -22.52 24.11	slr -3.40 0.00	slr -3.80 0.00
fkn -44.47 9.99	lr -31.16 21.01	gbt 0.00 -0.37	gbt 0.00 -0.62
kn -44.15 9.95	slr -30.98 20.69	fkdc -3.41 -2.69	fkdc -3.74 -10.75
base -68.32 0.00	base -48.73 0.00	kdc -3.45 -6.01	kdc -3.97 -15.08
svc -29.07 NaN	svc -19.20 NaN	svc -2.92 NaN	svc -3.69 NaN

5.1.21 Otros datasets: multiclas



anteojos (acc: 97.68%)

clf	delta_acc	r2
fkdc	-0.08	92.89
kdc	0.00	91.84
kn	-0.16	89.32
fkn	-0.17	87.08
gbt	-1.36	86.12
gnb	-0.62	85.17
lr	-55.51	27.80
slr	-56.00	27.60
base	-48.34	0.00
svc	-0.26	NaN

5.1.22 Otros datasets: digitos y mnist

digitos (acc: 98.41%)

mnist (acc: 87.07%)

clf	delta_acc	r2
fkdc	0.00	97.67
kdc	-0.10	96.80
gbt	-2.43	94.17
lr	-2.24	93.38
slr	-2.33	93.10
fkn	-1.98	92.22
kn	-2.91	90.62
gnb	-8.42	85.67
base	-89.43	0.00
svc	-0.16	NaN

clf	delta_acc	r2
kdc	-4.04	76.38
lr	-4.32	76.10
fkdc	-4.38	73.38
gbt	-10.11	66.57
slr	-12.33	61.43
gnb	-19.10	56.91
fkn	-16.28	54.02
kn	-19.10	50.40
base	-76.57	0.00
svc	0.00	NaN

5.1.23 Conclusiones

5.1.23.1 Para Patu

Lo junan a (Carpio *et al.*, 2019)? «Fingerprints of cancer by persistent homology»

We have carried out a topological data analysis of gene expressions for different databases based on the Fermat distance between the z scores of different tissue samples. There is a critical value of the fil-

tration parameter at which all clusters collapse in a single one. This critical value for healthy samples is gapless and smaller than that for cancerous ones. After collapse in a single cluster, topological holes persist for larger filtration parameter values in cancerous samples. Barcodes, persistence diagrams and Betti numbers as functions of the filtration parameter are different for different types of cancer and constitute fingerprints thereof.

6 Glosario

clausura	???
Riemanniana, métrica	sdfsdf
Lebesgue, medida de	???
densidad, estimación de	cf. (Berenfeld y Hoffmann, 2021)
ventana	parámetro escalar que determina la «unidad» de distancia
núcleo, función	K

7 Listados

Listado de Figuras

Figura 1	Dos círculos concéntricos y sus KDE marginales por clase: a pesar de que la frontera entre ambos grupos de puntos es muy clara, es casi imposible distinguirlas a partir de sus densidades marginales.	9
Figura 2	Ejemplos de variedades en el mundo físico: tanto la hoja de un árbol como una bandera flameando al viento tienen dimensión intrínseca $d_{\mathcal{M}} = 2$, están embedidas en \mathbb{R}^3 , y son definitivamente no-lineales.	13
Figura 3	Espacio tangente $T_p \mathcal{M}$ a una esfera $\mathcal{M} = S^2$ por p . Nótese que el espacio tangente varía con p , pero siempre mantiene la misma dimensión ($d = 2$) que \mathcal{M}	17
Figura 4	Espacio tangente y mapa exponencial para $p_N \in S^1$. Nótese que $\text{iny } S^1 = \pi$. Prolongando una geodésica $\gamma(t)$ más allá de $t = \pi$, ya no se obtiene un camino mínimo, pues hubiese sido más corto llegar por $-\gamma(s)$, $s = t \bmod \pi$	20
Figura 5	Pretendido «error» - diferencia módulo 1 - de los pesos atómicos medidos para ciertos elementos, sobre S^1 . Nótese como la mayoría de las mediciones se agrupan en torno al 0.0.	22

Figura 6	KDE en S^2 para $X = \text{sth}$ los flujos de lava de Fisher TODO mejorar imagen	28
Figura 7	Data espacial con dimensiones bien definidas. Los datos geoespaciales están sobre la corteza terrestre, que es aproximadamente la 2 –esfera $S^2 \in \mathbb{R}^3$ que representa la frontera de nuestra «canica azul» (izq.), una 3 –bola. La clasificación clásica de Hubble distingue literalmente <i>variedades</i> «elípticas»,«espirales» e «irregulares» de galaxias (der.). ⁷⁵	31
Figura 8	La variedad \mathcal{U} con $\dim(\mathcal{U}) = 1$ embebida en \mathbb{R}^2 . Nótese que en el espacio ambiente, el punto rojo está más cerca del verde, mientras que a través de \mathcal{U} , el punto amarillo está más próximo que el rojo	32
Figura 9	Ilustración de \mathbf{X} y sus componentes principales en « <i>LIII.</i> <i>On lines and planes of closest fit to systems of points in space.</i> » (Pearson, 1901)	33
Figura 10	Una bola de radio r creciente centrada en un punto de una 1 –variedad muestreada con ruido en \mathbb{R}^2 <i>minimiza</i> la tasa a la que incorpora observaciones cuando r está en la escala «localmente lineal» de la variedad.	36
Figura 11	Isomap aplicado a 1.000 dígitos «2» manuscritos del dataset <i>MNIST</i> con $d = 2$ (Tenenbaum, Silva y Langford, 2000). Nótese que las dos direcciones se corresponden fuertemente con características de los dígitos: el rulo inferior en el eje X , y el arco superior en el eje Y	39
Figura 12	Cuando por ejemplo $\mathcal{M} = (\mathbb{R}^2, g = \mathbf{I})$, $X \sim \mathcal{N}_d(a, \Sigma)$, tenemos que $d_g(a, b) = L(\gamma) = r = L(\zeta) = d_g(a, c)$, mientras que $d_\Sigma(a, b) < d_\Sigma(a, c)$: la normal multivariada tiene distintas tasas de cambio en distintas direcciones, y medir distancia ignorando este hecho puede llevar a conclusiones erróneas.	40
Figura 13	En el grafo completo de 3 vértices, hay sólo dos caminos entre a y c : $\zeta = a \rightarrow b \rightarrow c$, y $\gamma = a \rightarrow c$	41
Figura 14	Ejemplo trivial de la equivalencia $d_{\mathbf{N}} \equiv d_2$ para $P =$ $\{a, b\}$	45
Figura 15	<code>make_blobs(n_features=2, centers=(0, 0), (10, 0)), random_state=1984)</code>	54
Figura 16	«Lunas», «Círculos» y «Espirales», con $d_x = 2$, $d_{\mathcal{M}} = 1$ y $s = 4107$	55
Figura 17	Boxplots con la distribución de dxactitud en las 25 repeticiones de cada experimento de Figura 16	56

⁷⁵Se me perdonará la simplificación; es bien sabido que en realidad la [topología del espacio-tiempo](#) es un tópico de estudio clave en la relatividad general.

Figura 18	Boxplots con la distribución de dxactitud en las 25 repeticiones de cada experimento de Figura 16	56
Figura 19	Boxplots con la distribución de R^2 en las 25 repeticiones de cada experimento.	57
Figura 20	Exactitud promedio en entrenamiento para la corrida ("circulos", 4479). Las cruces rojas indican la ventana h óptima para cada α	60
Figura 21	It does replicate	61

8 Tablas

Listado de Tablas

Tabla 1	Resultados de entrenamiento en Figura 15	55
Tabla 2	«mi caption, bo».....	57
Tabla 3	«mi caption, bo-bo».....	58

9 Código

Listado de código

Bibliografía

- Bengio, Y. (2019) «The Consciousness Prior». arXiv.
- Bengio, Y., Courville, A. y Vincent, P. (2014) «Representation Learning: A Review and New Perspectives». arXiv.
- Bengio, Y., Larochelle, H. y Vincent, P. (2005) «Non-Local Manifold Parzen Windows», en *Advances in Neural Information Processing Systems*. MIT Press.
- Berenfeld, C. y Hoffmann, M. (2021) «Density Estimation on an Unknown Submanifold», *Electronic Journal of Statistics*, 15(1), pp. 2179-2223. Disponible en: <https://doi.org/10.1214/21-EJS1826>.
- Bijral, A.S., Ratliff, N. y Srebro, N. (2012) «Semi-Supervised Learning with Density Based Distances». arXiv. Disponible en: <https://doi.org/10.48550/arXiv.1202.3702>.
- Brand, M. (2002) «Charting a Manifold», en *Advances in Neural Information Processing Systems*. MIT Press.
- Carpio, A. et al. (2019) *Fingerprints of Cancer by Persistent Homology*. Disponible en: <https://doi.org/10.1101/777169>.
- Cayton, L. (2005) «Algorithms for Manifold Learning».
- Chu, T., Miller, G.L. y Sheehy, D.R. (2019) «Exact Computation of a Manifold Metric, via Lipschitz Embeddings and Shortest Paths on a

- Graph», *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics (Proceedings). Disponible en: <https://doi.org/10.1137/1.9781611975994.25>.
- Devroye, L., Györfi, L. y Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*. Repr. New York Berlin Heidelberg: Springer (Applications of Mathematics).
- Fisher, R.A. (1957) «Dispersion on a Sphere», *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130), pp. 295-305. Disponible en: <https://doi.org/10.1098/rspa.1953.0064>.
- Gallese, V. (2003) «The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity», *Psychopathology*, 36(4), pp. 171-180. Disponible en: <https://doi.org/10.1159/000072786>.
- Groisman, P., Jonckheere, M. y Sapienza, F. (2019) «Nonhomogeneous Euclidean First-Passage Percolation and Distance Learning». arXiv.
- Hall, P. y Kang, K.-H. (2005) «Bandwidth Choice for Nonparametric Classification», *The Annals of Statistics*, 33(1). Disponible en: <https://doi.org/10.1214/009053604000000959>.
- Hastie, T., Tibshirani, R. y Friedman, J. (2009) *Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer London, Limited.
- Henry, G. y Rodriguez, D. (2009) «Kernel Density Estimation on Riemannian Manifolds: Asymptotic Results», *Journal of Mathematical Imaging and Vision*, 34(3), pp. 235-239. Disponible en: <https://doi.org/10.1007/s10851-009-0145-2>.
- Jupp, P.E. y Mardia, K.V. (1989) «A Unified View of the Theory of Directional Statistics, 1975-1988», *International Statistical Review / Revue Internationale de Statistique*, 57(3), pp. 261-294. Disponible en: <https://doi.org/10.2307/1403799>.
- Lee, J.M. (2018) *Introduction to Riemannian Manifolds*. Cham: Springer International Publishing (Graduate Texts in Mathematics). Disponible en: <https://doi.org/10.1007/978-3-319-91755-9>.
- Little, A., McKenzie, D. y Murphy, J. (2021) «Balancing Geometry and Density: Path Distances on High-Dimensional Data». arXiv.
- Loubes, J.-M. y Pelletier, B. (2008) «A Kernel-Based Classifier on a Riemannian Manifold», *Statistics & Decisions*, 26(1), pp. 35-51. Disponible en: <https://doi.org/10.1524/stnd.2008.0911>.
- Mardia, K.V. (1975) «Distribution Theory for the Von Mises-Fisher Distribution and Its Application», en G.P. Patil, S. Kotz, y J.K. Ord (eds.) *A Modern Course on Statistical Distributions in Scientific Work*. Dordrecht: Springer Netherlands (NATO Advanced Study Institutes Series), pp. 113-130. Disponible en: https://doi.org/10.1007/978-94-010-1842-5_10.

- McFadden, D. (1974) «Conditional Logit Analysis of Qualitative Choice Behavior», *Frontiers in econometrics* [Preprint].
- Mckenzie, D. y Damelin, S. (2019) «Power Weighted Shortest Paths for Clustering Euclidean Data». arXiv.
- von Mises, R. (1918) «Über Die "Ganzzahligkeit" Der Atomgewicht Und Verwandte Fragen», *Physikal. Z.*, 19, pp. 490-500.
- Muñoz, A.L. (2011) *Estimación no paramétrica de la densidad en variedades Riemannianas*.
- Parzen, E. (1962) «On Estimation of a Probability Density Function and Mode», *The annals of mathematical statistics*, 33(3), pp. 1065-1076.
- Pearson, K. (1901) «LIII. On Lines and Planes of Closest Fit to Systems of Points in Space». Disponible en: <https://doi.org/10.1080/14786440109462720>.
- Pelletier, B. (2005) «Kernel Density Estimation on Riemannian Manifolds», *Statistics & Probability Letters*, 73(3), pp. 297-304. Disponible en: <https://doi.org/10.1016/j.spl.2005.04.004>.
- Rifai, S. et al. (2011) «The Manifold Tangent Classifier», en *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Rosenblatt, M. (1956) «Remarks on Some Nonparametric Estimates of a Density Function», *The Annals of Mathematical Statistics*, 27(3), pp. 832-837.
- Simard, P. et al. (1991) «Tangent Prop - A Formalism for Specifying Selected Invariances in an Adaptive Network», en *Advances in Neural Information Processing Systems*. Morgan-Kaufmann.
- Tenenbaum, J. (1997) «Mapping a Manifold of Perceptual Observations», en *Advances in Neural Information Processing Systems*. MIT Press.
- Tenenbaum, J.B., Silva, V. de y Langford, J.C. (2000) «A Global Geometric Framework for Nonlinear Dimensionality Reduction», *Science*, 290(5500), pp. 2319-2323. Disponible en: <https://doi.org/10.1126/science.290.5500.2319>.
- Vincent, P. y Bengio, Y. (2003) «Density Sensitive Metrics and Kernels», en *Proceedings of the Snowbird Workshop*.
- Vincent, P. y Bengio, Y. (2002) «Manifold Parzen Windows», en *Advances in Neural Information Processing Systems*. MIT Press.
- Wand, M.P. y Jones, M.C. (1993) «Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation», *Journal of the American Statistical Association*, 88(422), pp. 520-528. Disponible en: <https://doi.org/10.1080/01621459.1993.10476303>.
- Wand, M.P. y Jones, M.C. (1995) *Kernel Smoothing*. Boston, MA: Springer US. Disponible en: <https://doi.org/10.1007/978-1-4899-4493-1>.

9.1 Slides sobre diseño experimental

9.1.1 KDC con Distancia de Fermat Muestral

9.1.2 f-KNN

9.1.3 Algunas dudas

- Entrenar el clasificador por validación cruzada está OK: como $\mathbf{X}_{\text{train}} \subseteq \mathbf{X}$ y $\mathbf{X}_{\text{test}} \subseteq \mathbf{X}$, se sigue que $\forall (a, b) \in \{\mathbf{X}_{\text{train}} \times \mathbf{X}_{\text{test}}\} \subseteq \{\mathbf{X} \times \mathbf{X}\}$ y $D_{\mathbf{X}, \alpha}(a, b)$ está bien definida. ¿Cómo sé la distancia *muestral* de una *nueva* observación x_0 , a los elementos de cada clase?

Para cada una de las $g_i \in \mathcal{G}$ clases, definimos el conjunto

$$Q_i = \{x_0\} \cup \{x_j : x_j \in \mathbf{X}, g_j = g_i, j \in \{1, \dots, N\}\} \quad (86)$$

y calculamos $D_{Q_i, \alpha}(x_0, \cdot)$

9.1.4 Algunas dudas

- El clasificador de Loubes & Pelletier asume que todas las clases están soportadas en la misma variedad \mathcal{M} . ¿Quién dice que ello vale para las diferentes clases?

¡Nadie! Pero

1. No hace falta dicho supuesto, y en el peor de los casos, podemos asumir que la unión de las clases está soportada en *cierta* variedad de Riemman, que resulta de (¿la clausura de?) la unión de sus soportes individuales.
2. Sí es cierto que si las variedades (y las densidades que soportan) difieren, tanto el α_i^* como el h_i * «óptimos» para los estimadores de densidad individuales no tienen por qué coincidir.
3. Aunque las densidades individuales f_i estén bien estimadas, el clasificador resultante puede ser mal(ard)o si no diferencia bien «en las fronteras». Por simplicidad, además, decidimos parametrizar el clasificador con dos únicos hiperparámetros globales: α, h .

(Hall y Kang, 2005) h óptimo para clasificación con KDE

9.1.5 Diseño experimental

1. Desarrollamos un clasificador compatible con el *framework* de [scikit-learn](#) según los lineamientos de Loubes & Pelleteir, que apodamos **KDC**.
2. Implementamos el estimador de la distancia muestral de Fermat, y combinándolo con KDC, obtenemos la titular «Clasificación por KDE con Distancia de Fermat», **FKDC**.
3. Evaluamos el *pseudo-R²* y la *exactitud* («accuracy») de los clasificadores propuestos en diferentes *datasets*, relativa a técnicas bien establecidas:
 - regresión logística (LR)
 -

- clasificador de soporte vectorial (svc)⁷⁶
 - k-vecinos-más-cercanos (KN)
 - Naive Bayes Gaussiano (GNB)
- gradient boosting trees (GBT)
- La implementación de KNeighbors de referencia acepta distancias pre-computadas, así que incluimos una versión con distancia de Fermat, que apodamos F(ermat)KN.
- Para ser «justos», se reservó una porción de los datos para la evaluación comparada, y del resto, cada algoritmo fue entrenado repetidas veces por validación cruzada de 5 pliegos, en una extensísima grilla de hiperparametrizaciones. Este procedimiento se repitió 25 veces en cada dataset.
- La función de score elegida fue neg_log_loss (= ℓ) para los clasificadores suaves, y accuracy para los duros.
- Para tener una idea «sistémica» de la performance de los algoritmos, evaluaremos su performance con datasets que varíen en el tamaño muestral N , la dimensión p de las X_i , el nro. de clases k y su origen («real» o «sintético»).
- Cuando creamos datos sintéticos en variedades con dimensión intrínseca menor a la ambiente, (casi) cualquier clasificador competente alcanza exactitud perfecta; para complejizar la tarea, agregamos un poco de «ruido» a las observaciones, y también analizamos sus efectos.

9.1.6 Regla de Parsimonia

- ¿Qué parametrización elegir cuando «en test da todo igual»?
 - de Occam: la más «sencilla» (TBD)
- ¿Qué parametrización elegir cuando «en test da casi todo igual»?
 - Regla de 1Σ : De las que estén a 1Σ de la mejor, la más sencilla.
 - ¿Sabemos cuánto vale Σ ?

9.1.7 R^2 de McFadden

Sea \mathcal{C}_0 el clasificador «base», que asigna a cada observación y posible clase, la frecuencia empírica de clase encontrada en la muestra \mathbf{X} . Para todo clasificador suave \mathcal{C} , definimos el R^2 de McFadden como

$$R^2(\mathcal{C} | \mathbf{X}) = 1 - \frac{\ell(\mathcal{C})}{\ell(\mathcal{C}_0)} \quad (87)$$

donde $\ell(\cdot)$ es la log-verosimilitud clásica. Nótese que $R^2(\mathcal{C}_0) = 0$. A su vez, para un clasificador perfecto \mathcal{C}^* que otorgue toda la masa de probabilidad a la clase correcta, tendrá $L(\mathcal{C}^*) = 1$ y log-verosimilitud igual a 0, de manera que $R^2(\mathcal{C}^*) = 1 - 0 = 1$.

⁷⁶sólo se consideró su exactitud, ya que no es un clasificador suave

Sin embargo, un clasificador *peor* que \mathcal{C}_0 en tanto asigne bajas probabilidades (≈ 0) a las clases correctas, puede tener un R^2 infinitamente negativo.