Tecnológico de Monterrey

Análisis Topológico de Datos

Daniel Otero Fadul

Departamento de Ciencias
Escuela de Ingeniería y Ciencias

Análisis Topológico de Datos

El **Análisis Topológico de Datos** (TDA) es un método de análisis de datos que aprovecha los conceptos de la topología, una rama de las matemáticas que se ocupa de las propiedades espaciales que se preservan bajo transformaciones continuas. El TDA pretende extraer y comprender la forma, la estructura y los patrones subyacentes en conjuntos de datos complejos centrándose en las características topológicas en lugar de medidas geométricas específicas.

Análisis Topológico de Datos

El TDA es especialmente útil para analizar conjuntos de datos de gran dimensión y con ruido en los que los métodos tradicionales, como la agrupación o la reducción de la dimensionalidad, pueden resultar menos eficaces. Proporciona un potente marco para descubrir estructuras ocultas, detectar anomalías y comprender las propiedades globales de espacios de datos complejos.

A continuación se presentan dos herramientas muy básicas sobre las que se construye el análisis de datos topológicos: los **complejos simpliciales** y los **grupos homología**.

Complejo Simplicial

Para $k \geq 0$, un k-**símplex** σ en un espacio Euclidiano $\mathbb{R}^m$ es el casco convexo de un conjunto P de $k+1$ puntos *afínmente independientes* en $\mathbb{R}^m$. En particular, un 0-símplex es un vértice, un 1-símplex es un borde, un 2-símplex es un triángulo, y un 3-símplex es un tetrahedro. Un k-símplex tiene dimensión k.

Complejo Simplicial

Para $0 \leq k' \leq k$, una k'-**cara**, o simplemente una cara, de σ es un k'-símplex que es un casco convexo de un subconjunto no vacío de P. Las dimensionalidad de las caras de σ van desde cero (los vértices de σ) hasta k; y σ es una cara de σ. Una **cara propia** de σ es un símplex que es un casco convexo de un subconjunto propio de P; en otras palabras, cualquier cara excepto σ. Las $(k-1)$-caras de σ se les llama **facetas** de σ.

Complejo Simplicial

Un **complejo geométrico simplicial** K, también conocido como triangulación, es un conjunto que contiene un número finito de símplices y que satisface las dos restricciones siguientes:

- K contiene todas las caras de cada símplex de K.
- Para dos símplices cualesquiera $\sigma, \tau \in K$, su intersección $\sigma \cap \tau$ es vacía o es una cara de σ y τ.

Los elementos de $V(K)$ son los vértices de K. Se dice que cada k-símplex en K tiene dimensión k. También decimos que K es un k-complejo simplicial si la mayor dimensión de cualquier símplex en K es k.

Complejo Simplicial

Una colección K de subconjuntos no vacíos de un conjunto dado $V(K)$ es un **complejo simplicial abstracto** si cada elemento $\sigma \in K$ tiene todos sus subconjuntos no vacíos $\sigma' \subseteq \sigma$ también en K. Cada elemento σ con $|\sigma| = k + 1$ se llama un k-símplex (o simplemente un símplex).

Complejo Simplicial

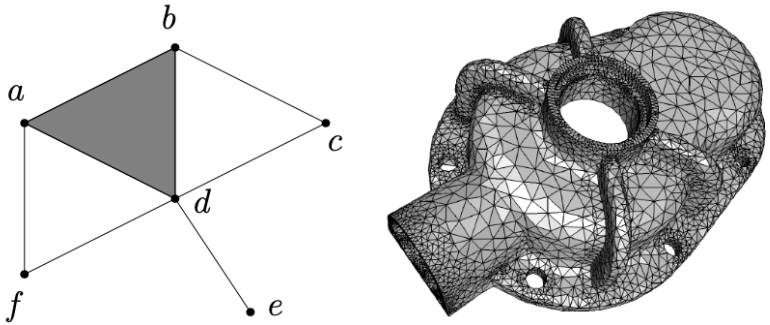

Figura: A la izquierda, un complejo simplicial con seis vértices, ocho aristas y un triángulo; a la derecha, complejo simplicial que triangula una superficie en $\mathbb{R}^3$. Imagen tomada de [1].

Complejo Simplicial

Dada una colección finita de conjuntos $U = \{U_\alpha\}_{\alpha \in A}$, definimos el **nervio** del conjunto U como el complejo simplicial $N(U)$ cuyo conjunto de vértices es el conjunto índice A, y donde un subconjunto $\{\alpha_0, \alpha_1, \ldots, \alpha_k\} \subseteq A$ es un k-símplex en $N(U)$ si y solo si $U_{\alpha_0} \cap U_{\alpha_1} \cap \ldots \cap U_{\alpha_k} \neq \emptyset$.

En este contexto, U es una **cobertura** de un espacio métrico o topológico, el cual denotaremos como M.

Complejo Simplicial

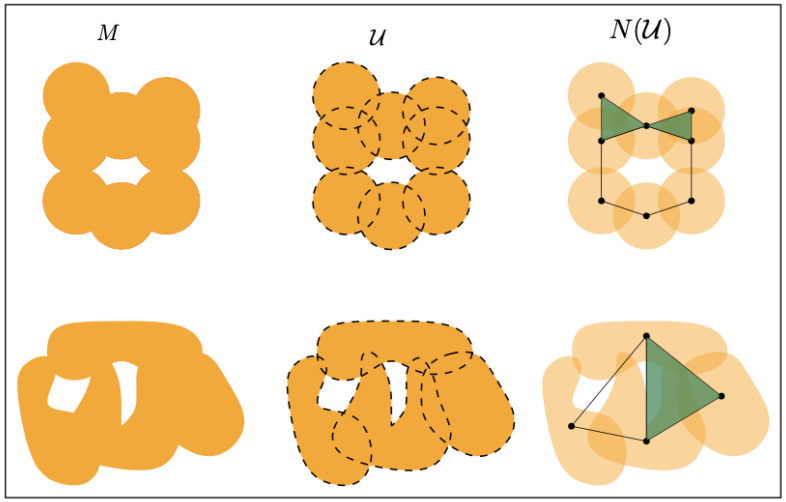

Figura: Ejemplos de dos espacios (izquierda), cubiertas abiertas de ellos (centro) y sus nervios (derecha). Imagen tomada de [1].

Complejo Simplicial

Sea (M, d) un espacio métrico y P un subconjunto finito de M. Dado $r \in \mathbb{R}$, $r > 0$, se define el **complejo de Čech** $C_r(P)$ como el nervio del conjunto $\{B(p_i, r)\}$, donde $B(p_i, r)$ es la bola cerrada de radio r centrada en p_i:

$$B(p_i, r) = \{x \in M \mid d(p_i, x) \leq r\}.$$

Complejo Simplicial

Sea (P, d) un espacio métrico finito. Dado un real $r > 0$, el **complejo de Vietoris-Rips** (abreviado Rips) es el complejo simplicial abstracto $VR_r(P)$ donde un símplex $\sigma \in VR_r(P)$ si y solo si $d(p, q) \leq 2r$ para cada par de vértices de σ.

Complejo Simplicial

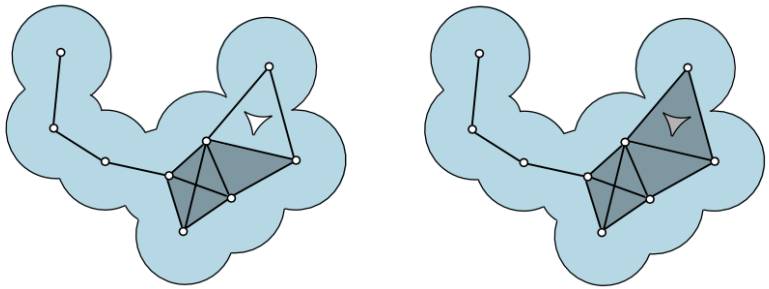

Figura: De izquierda a derecha, el complejo simplicial de Čech y el complejo de Vietori-Rips. El complejo de Čech es más utilizado en la teoría, mientras que el complejo de Vietori-Rips es más utilizado en la práctica ya que es más fácil de calcular. Para el mismo valor r, $C_r(P) \subset VR_r(P)$. Imagen tomada de [1].

Cadenas

Sea K un k-complejo simplicial con m_p p-símplices, $k \geq p \geq 0$. Una p-**cadena** c en K es una suma de p-símplices multiplicados por unos coeficientes, es decir,

$$c = \sum_{i=1}^{m_p} \alpha_i \sigma_i,$$

donde σ_i son los símplices y α_i son los coeficientes. En general, los coeficientes α_i son elementos de un **anillo**.

Cadenas

Dos p-cadenas se pueden sumar para obtener otra cadena:

$$\begin{aligned} c + c' &= \sum_{i=1}^{m_p} \alpha_i \sigma_i + \sum_{i=1}^{m_p} \alpha'_i \sigma_i \\ &= \sum_{i=1}^{m_p} (\alpha_i + \alpha'_i) \sigma_i. \end{aligned}$$

A menos que se diga lo contrario, asumiremos que los α_i son elementos del campo escalar $\mathbb{Z}_2$, el cual es un **campo escalar** finito, que es un tipo especial de anillo.

Cadenas

Las p-cadenas, cuyos coeficientes son elementos de $\mathbb{Z}_2$, forman un grupo. La identidad de este grupo es la cadena

$$0 = \sum_{i=1}^{m_p} 0\sigma_i.$$

Nótese que una cadena c es su propia inversa ya que $c + c = 0$. Este grupo, llamado el grupo **p-ésima cadena**, se denota como $C_p(K)$.

Fronteras

Los grupos de cadenas en diferentes dimensiones están relacionados por un **operador de frontera**. Dado un p-símplex $\sigma = \{v_0, \ldots, v_p\}$, la acción del operador de frontera sobre σ se define como

$$\partial_p \sigma = \sum_{i=0}^{p} \{v_0, \ldots, \hat{v}_i, \ldots, v_p\}$$

donde $\hat{v}_i$ indica que se omite el vértice v_i. Informalmente, podemos ver ∂_p como un mapa que envía un p-símplex σ a la $(p-1)$-cadena que tiene coeficientes distintos de cero solo en $(p-1)$-caras de σ, o en otras palabras, la frontera de σ. En este punto, es instructivo observar que la frontera de un vértice es un conjunto vacío, es decir, $\partial_0 \sigma = \emptyset$.

Fronteras

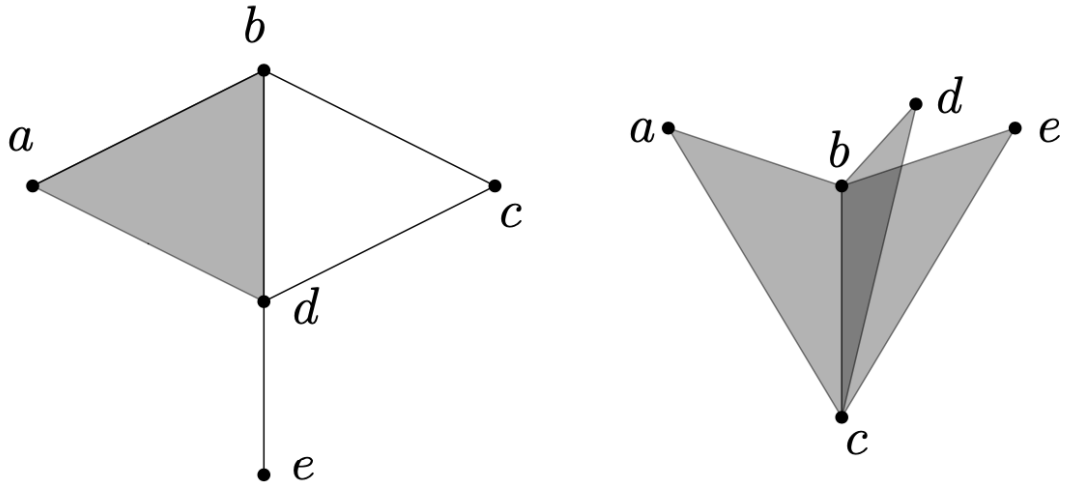

Figura: Dos complejos simpliciales. Imagen tomada de [1].

Consideremos la 2-cadena $[abc] + [bcd]$ del complejo simplicial ubicado a la derecha de la figura. Entonces, se tiene que

$$\begin{aligned} \partial_2([abc] + [bcd]) &= ([ab] + [bc] + [ca]) + ([bc] + [cd] + [db]) \\ &= [ab] + [ca] + [cd] + [db]. \end{aligned}$$

Fronteras

Extendiendo ∂_p a una p-cadena, obtenemos un homomorfismo $\partial_p : C_p \to C_{p-1}$ llamado operador límite que produce una $(p-1)$-cadena cuando se aplica a una p-cadena:

$$\partial_p c = \sum_{i=1}^{m_p} \alpha_i \partial_p \sigma_i,$$

donde $c = \sum_{i=1}^{m_p} \alpha_i \sigma_i$.

Fronteras

De nuevo, observamos el caso especial de $p = 0$ cuando obtenemos $\partial_0 c = \emptyset$. El grupo de cadenas C_{-1} solo tiene un único elemento que es su identidad 0. En el otro extremo, también suponemos que si K es un complejo k, entonces C_p es 0 para $p > k$.

Por otro lado, extendiendo el operador límite a los grupos de cadenas, obtenemos la siguiente secuencia de homomorfismos para un complejo simplicial de dimensión k:

$$0 = C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} C_{k-2} \cdots C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} C_{-1} = 0.$$

Esta secuencia también se le conoce como **complejo de cadena**.

Ciclos

Ciclo y grupo de ciclos. Una p-cadena c es un p-ciclo si $\partial_p c = 0$. En otras palabras, una cadena que tiene un borde vacío es un ciclo. Todos los p-ciclos juntos forman el p-ésimo grupo de ciclos Z_p bajo la adición que se usa para definir los grupos de cadenas. En términos del operador borde, Z_p es el subgrupo de C_p que se envía al cero de C_{p-1}, es decir, $\ker(\partial_p) = Z_p$.

Ciclos

Por ejemplo, en la figura de la diapositiva 18 (derecha), la 1-cadena $ab + bc + ca$ es un 1-ciclo ya que

$$\partial_1([ab] + [bc] + [ca]) = ([a] + [b]) + ([b] + [c]) + ([c] + [a]) = 0.$$

Además, nótese que la 1-cadena anterior es el borde del triángulo abc. No es casualidad que el borde de un símplex sea un ciclo: el borde de una p-cadena es un $(p-1)$-ciclo. Este es un hecho fundamental en la teoría de la homología.

Bordes

El conjunto de p-cadenas que se pueden obtener aplicando el operador de frontera ∂_{p+1} sobre $p+1$-cadenas forma un subgrupo de p-cadenas, el cual se le conoce como **p-ésimo grupo de bordes** $B_p = \partial_{p+1}(C_{p+1})$; en otras palabras, la imagen del homomorfismo de borde es el grupo de bordes $B_p = \text{Im}(\partial_{p+1})$. Tenemos que $\partial_p B_p = 0$ para $p > 0$, por lo tanto, $B_p \subseteq Z_p$.

Bordes

Para un k-complejo simplicial,

1. $C_0 = Z_0$ y $B_k = 0$.
2. Para $p \geq 0$, $B_p \subseteq Z_p \subseteq C_p$.
3. Por cierto, C_p, B_p y Z_p son espacios vectoriales.

Vale la pena decir que C_p, B_p y Z_p al ser espacios vectoriales pueden tener una **base**.

Grupos de Homología

Los **grupos de homología** clasifican los ciclos de un grupo de ciclos juntando los ciclos de la misma clase que se diferencian por un borde. Desde el punto de vista de la teoría de grupos, esto se hace tomando el cociente de los grupos de ciclos con los grupos de bordes, lo que está permitido ya que el grupo de bordes es un subgrupo del grupo de ciclos.

Grupos de Homología

Para $p \geq 0$, el p-**ésimo grupo de homología** es el grupo cociente $H_p = Z_p/B_p$. Dado que usamos un campo, concretamente $\mathbb{Z}_2$, para los coeficientes, H_p es un espacio vectorial y su dimensión se llama el p-**ésimo número de Betti**, denotado por β_p:

$$\beta_p := \dim H_p.$$

Grupos de Homología

Un círculo y la frontera de un triángulo sólido son homeomórficos: un triángulo hueco es un 1-símplice que se puede considerar un complejo simplicial que es una triangulación del círculo. En este caso, las bases de las cadenas son las siguientes:

$$\begin{aligned} C_n &: \{0\},\ n \geq 2, \\ C_1 &: \{[a,b],[b,c],[c,a]\} \\ C_0 &: \{a,b,c\}. \end{aligned}$$

Grupos de Homología

Los elementos de las cadenas son combinaciones lineales de las bases. Por ejemplo, los elementos de C_1 son de la forma

$$m[a,b] + n[b,c] + p[c,a],$$

donde $m, n, p \in \mathbb{Z}_2$.

Nótese que C_1 es isomórfico a $\mathbb{Z}_2^3$: $C_1 \approx \mathbb{Z}_2^3$. En otras palabras $C_1 \approx \mathbb{Z}_2 \bigoplus \mathbb{Z}_2 \bigoplus \mathbb{Z}_2$, donde $\bigoplus$ es una operación entre grupos conocida como **suma directa**.

Grupos de Homología

Ya teniendo caracterizadas las k-cadenas podemos calcular los grupos de homología del círculo, empezando con H_0. Tenemos que $H_0 = Z_0/B_0$, donde $Z_0 = \ker(\partial_0)$ y $B_0 = \text{Im}(\partial_1)$. En este caso, sabemos que

$$\partial_0(C_0) = 0,$$

así que $\ker(\partial_0) = C_0 = \{a, b, c\}$. Es más, $\ker(\partial_0) \approx \mathbb{Z}_2 \bigoplus \mathbb{Z}_2 \bigoplus \mathbb{Z}_2$.

En cuanto a $\text{Im}(\partial_1)$, se tiene que

$$\begin{aligned} \partial_1([a,b]) &= b+a \\ \partial_1([b,c]) &= c+b \\ \partial_1([c,a]) &= a+c. \end{aligned}$$

Nótese que $\partial_1([c,a]) = \partial_1([b,c]) + \partial_1([a,b])$, por lo cual los elementos de $\text{Im}(\partial_1)$ son de la forma

$$m(b+a) + n(c+b),$$

donde $m, n \in \mathbb{Z}_2$. En otras palabras, $\text{Im}(\partial_1)$ es generado por $\langle b+a, c+b \rangle$. Además, $\text{Im}(\partial_1) \approx \mathbb{Z}_2 \bigoplus \mathbb{Z}_2$.

Grupos de Homología

Para encontrar el grupo cociente Z_0/B_0, igualamos $\langle b+a, c+b\rangle$ a $\langle 0,0\rangle$. Esto implica que

$$\begin{aligned} b+a &= 0 \\ c+b &= 0, \end{aligned}$$

por lo tanto $a=b=c$. Así que $H_0 = Z_0/B_0 = \langle a\rangle$. Nótese que $H_0 \approx \mathbb{Z}_2$.

Grupos de Homología

Para encontrar H_1 seguimos un procedimiento similar. En este caso empezamos con $\ker(\partial_1)$:

$$\begin{aligned} \partial_1(m[a,b]+n[b,c]+p[c,a]) &= m(b+a)+n(c+b)+p(a+c) \\ &= (p+m)a+(m+n)b+(n+p)c. \end{aligned}$$

Para que la expresión anterior sea igual a cero debe cumplirse que $m=n=p$, por lo cual

$$\ker(\partial_1)=\langle a+b+c\rangle \approx \mathbb{Z}_2.$$

En cuanto a $\text{Im}(\partial_2)$, ya que no hay 2-símplices, $C_2 = 0$, así que $B_1 = \text{Im}(\partial_2) = 0$. Por lo tanto,

$$\begin{aligned} H_1 &= Z_1/B_1 \\ &= \langle a+b+c \rangle / 0 \\ &= \langle a+b+c \rangle \approx \mathbb{Z}_2. \end{aligned}$$

Grupos de Homología

Ya que $C_2 = 0$, se tiene que

$$\begin{aligned} \ker(\partial_2) &= 0 \\ \mathrm{Im}(\partial_3) &= 0, \end{aligned}$$

así que

$$\begin{aligned} H_2 &= Z_2/B_2 \\ &= 0/0 \\ &= 0. \end{aligned}$$

Es más, $H_n = 0$ para $n \geq 2$.

Grupos de Homología

Ya con los grupos de homología podemos calcular los **números de Betti**. Para el círculo obtuvimos que

$$\begin{aligned} H_0 &\approx \mathbb{Z}_2 \\ H_1 &\approx \mathbb{Z}_2 \\ H_n &\approx 0, \ n \geq 2. \end{aligned}$$

Por consiguiente,

$$\begin{aligned} \beta_0 &= 1 \\ \beta_1 &= 1 \\ \beta_n &= 0, \ n \geq 2. \end{aligned}$$

Grupos de Homología

Intuitivamente, el número de Betti β_k hace referencia al número de agujeros de k dimensiones en una superficie topológica. Un *"hueco k-dimensional"* es una k-cadena que no es una frontera de un objeto de $k+1$ dimensiones.

Volviendo al ejemplo del círculo, obtuvimos un agujero de dimensión cero ($\beta_0 = 1$) —lo cual se puede interpretar como un único objeto totalmente conectado—, un agujero de una dimensión ($\beta_1 = 1$), y cero cavidades de dimensión dos o mayor ($\beta_n = 0$, $n \geq 2$).

Grupos de Homología

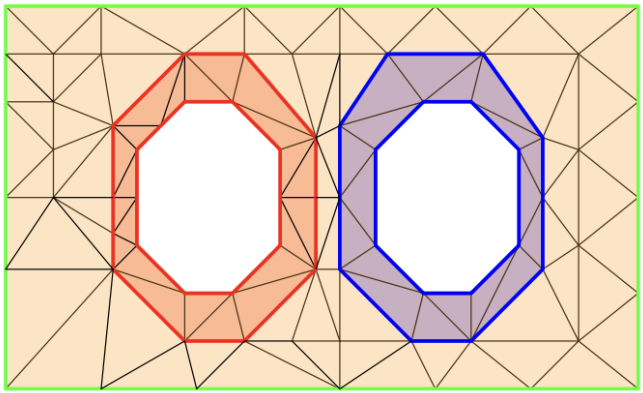

Figura: Cada uno de los ciclos rojo, azul y verde no son un borde porque no enlazan ninguna 2-cadena. Sin embargo, la suma de los dos ciclos rojos y la suma de los dos ciclos azules sí forman un borde ya que limitan con 2-cadenas formadas por triángulos rojizos y azulados, respectivamente. En otras palabras, los ciclos rojos y los ciclos azules pertenecen a clases de equivalencia distintas en H_1 pues son ciclos de distintos huecos de dimensión uno. Imagen tomada de [1].

Filtración

Consideremos un complejo simplicial K y una función $f : K \to \mathbb{R}$. Requerimos que f sea monótona, lo que significa que no decrece a lo largo de cadenas crecientes de caras, es decir, $f(\sigma) \leq f(\tau)$ siempre que σ sea una cara de τ. La monotonía implica que el conjunto de subniveles, $K(a) = f^{-1}(-\infty, a]$, es un subcomplejo de K para cada $a \in \mathbb{R}$.

Filtración

Sea m el número de símplices en K. Entonces, podemos tener $n \leq m$ subcomplejos diferentes, los cuales ordenamos como una secuencia creciente:

$$\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K.$$

En otras palabras, si $a_1 < a_2 < \ldots < a_n$ son los valores de la función de los símplices en K y $a_0 = -\infty$, entonces $K_i = K(a_i)$ para cada i. Llamamos a esta secuencia de complejos la **filtración** de f y pensamos en ella como una construcción en la que se van agregando bloques de símplices a medida que aumenta a.

Filtración

Más que en la secuencia de complejos, estamos interesados en la evolución topológica, tal como se expresa en la secuencia correspondiente de grupos de homología. Para cada $i \leq j$ tenemos un mapa de inclusión desde el espacio subyacente de K_i al de K_j y por lo tanto un homomorfismo inducido,

$$h_p^{i,j} : H_p(K_i) \to H_p(K_j),$$

para cada dimensión p. Así, la filtración corresponde a una secuencia de grupos de homología conectados por homomorfismos:

$$0 = H_p(K_0) \to H_p(K_1) \to \ldots \to H_p(K_n) = H_p(K).$$

Para cada dimensión p, a medida que avanzamos de K_{i-1} a K_i, ganamos nuevas clases de homología y perdemos algunas cuando se vuelven triviales o se fusionan entre sí. Recogemos las clases que nacen en o antes de un umbral dado y mueren después de otro umbral en grupos.

Grupos de Homología Persistente

Los **grupos de homología persistente de orden** p son las imágenes de los homomorfismos inducidos por la inclusión

$$H_p^{i,j} = \operatorname{Im} h_p^{i,j},$$

para $0 \leq i \leq j \leq n$. Nótese que $H_p^{i,i} = H_p(K_i)$. Los **números de Betti persistentes** correspondientes de orden p son los rangos de estos grupos:

$$\beta_p^{i,j} = \operatorname{rank} H_p^{i,j}.$$

En otras palabras, los grupos de homología persistente consisten en las clases de equivalencia de $H_p^{i,j}$ que sobreviven la transición de K_i a K_j o, más formalmente,

$$H_p^{i,j} = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i)).$$

Grupos de Homología Persistente

El grupo $H_p^{i,j}$ existe para cada dimensión p y cada par de índices $i \leq j$. Sea γ una clase de equivalencia en $H_p(K_i)$. Decimos que γ nace en K_i si

$$\gamma \notin H_p^{i-1,i}.$$

Además, si γ nace en K_i, se dice que muere entrando en K_j si se fusiona con una clase más antigua al pasar de K_{j-1} a K_j, es decir,

$$h_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1},$$

pero

$$h_p^{i,j}(\gamma) \in H_p^{i-1,j}.$$

Grupos de Homología Persistente

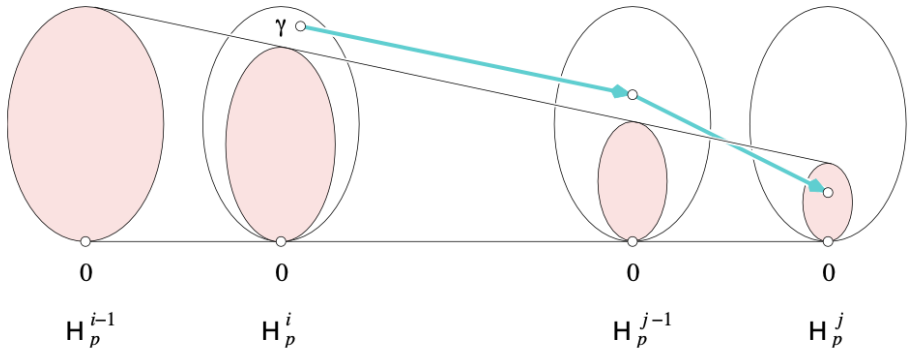

Figura: La clase γ nace en K_i ya que no se encuentra en la imagen (sombreada) de H_{i-1}. Además, γ muere al entrar en K_j ya que esta es la primera vez que su imagen se fusiona con la imagen de H_{i-1}. Imagen tomada de [3].

Grupos de Homología Persistente

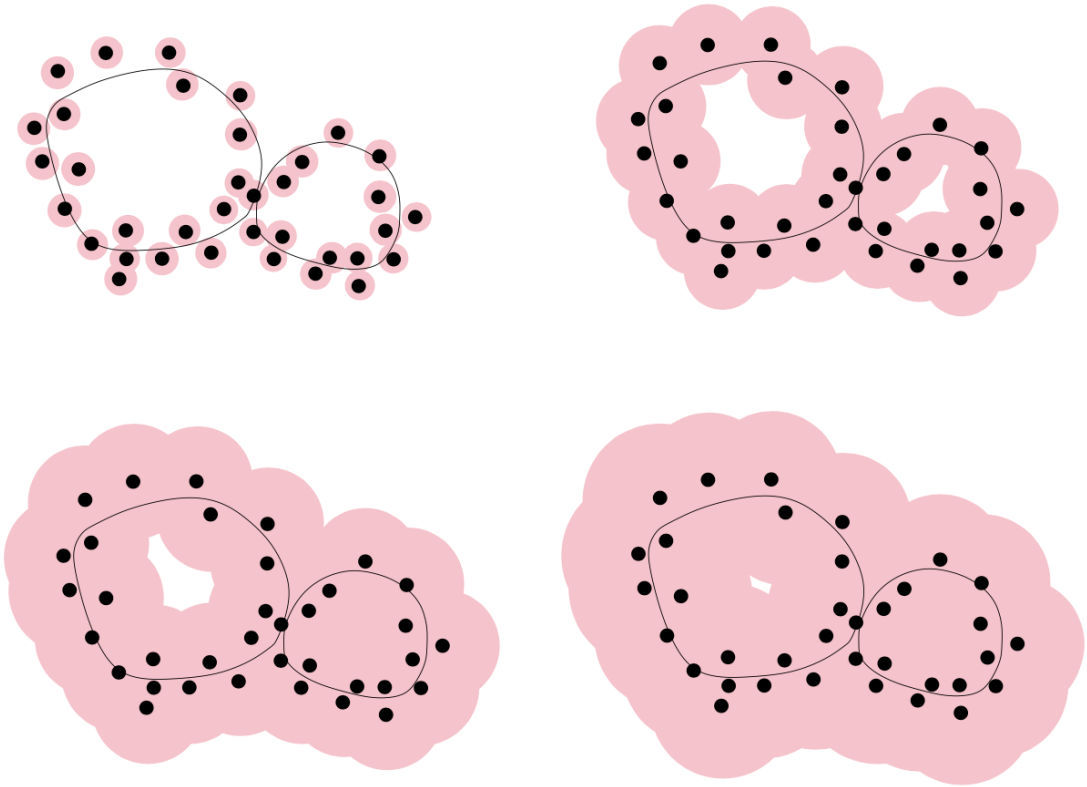

Figura: Muestra ruidosa de una curva con dos bucles y los conjuntos de subnivel crecientes de la función de distancia a los puntos de la muestra: El bucle más grande persiste más tiempo que el bucle más pequeño, mientras que otros agujeros espurios persisten aún menos tiempo. Imagen tomada de [1].

Grupos de Homología Persistente

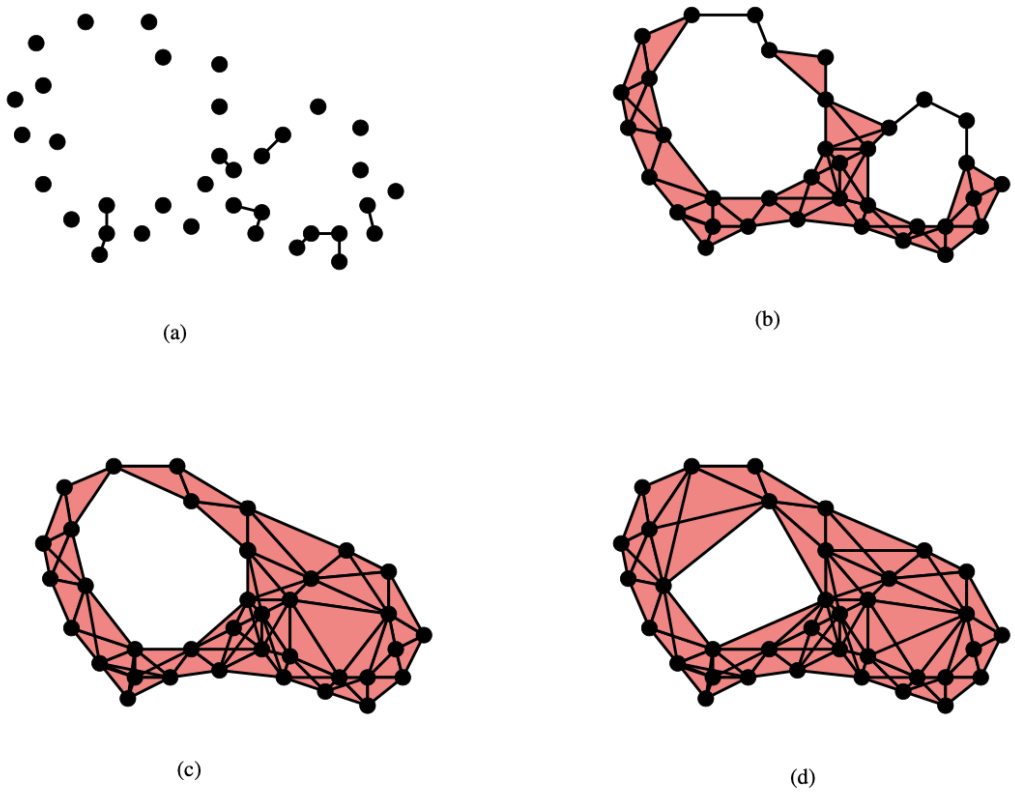

Figura: Complejo de Čech de la unión de bolas consideradas anteriormente. Las clases de homología en H_1 nacen y mueren a medida que la unión crece. Los dos agujeros más prominentes aparecen como las dos clases de homología más persistentes en H_1. Otras clases aparecen y desaparecen rápidamente con una persistencia relativamente mucho más corta. Imagen tomada de [1].

Persistencia

Si γ nace en K_i y muere al entrar en K_j, llamamos a la diferencia en el valor de la función la **persistencia**, $\text{pers}(\gamma) = a_j - a_i$.

A veces preferimos ignorar los valores reales de la función y considerar la diferencia en el índice, $j - i$, lo cual llamamos la **persistencia del índice** de la clase.

Si γ nace en K_i pero nunca muere, entonces establecemos su persistencia, así como su persistencia de índice, como infinito.

Diagramas de Persistencia

Visualizamos la colección de números de Betti persistentes dibujando puntos en el plano real extendido, $\overline{\mathbb{R}}^2 := (\mathbb{R} \cup \{\pm\infty\})^2$. Sea $\mu_p^{i,j}$ el número de clases de dimensión p que nacen en K_i y mueren entrando en K_j, tenemos

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j}),$$

para todo $i < j$ y todo p. La primera diferencia en el lado derecho cuenta las clases que nacen en K_i o antes y mueren entrando en K_j, mientras que la segunda diferencia cuenta las clases que nacen en K_{i-1} o antes y mueren entrando en K_j.

Diagramas de Persistencia

El **diagrama de persistencia** $\mathrm{Dgm}_p(F_f)$ (también escrito $\mathrm{Dgm}_p f$) de una filtración F_f inducida por una función f se obtiene dibujando un punto (a_i, a_j) con multiplicidad no nula $\mu_{i,j}$, $i < j$, en el plano extendido

$$\overline{\mathbb{R}}^2 := (\mathbb{R} \cup \{\pm\infty\})^2,$$

donde los puntos con multiplicidad infinita se añaden en la diagonal $\Delta : \{(x, x)\}$.

Diagramas de Persistencia

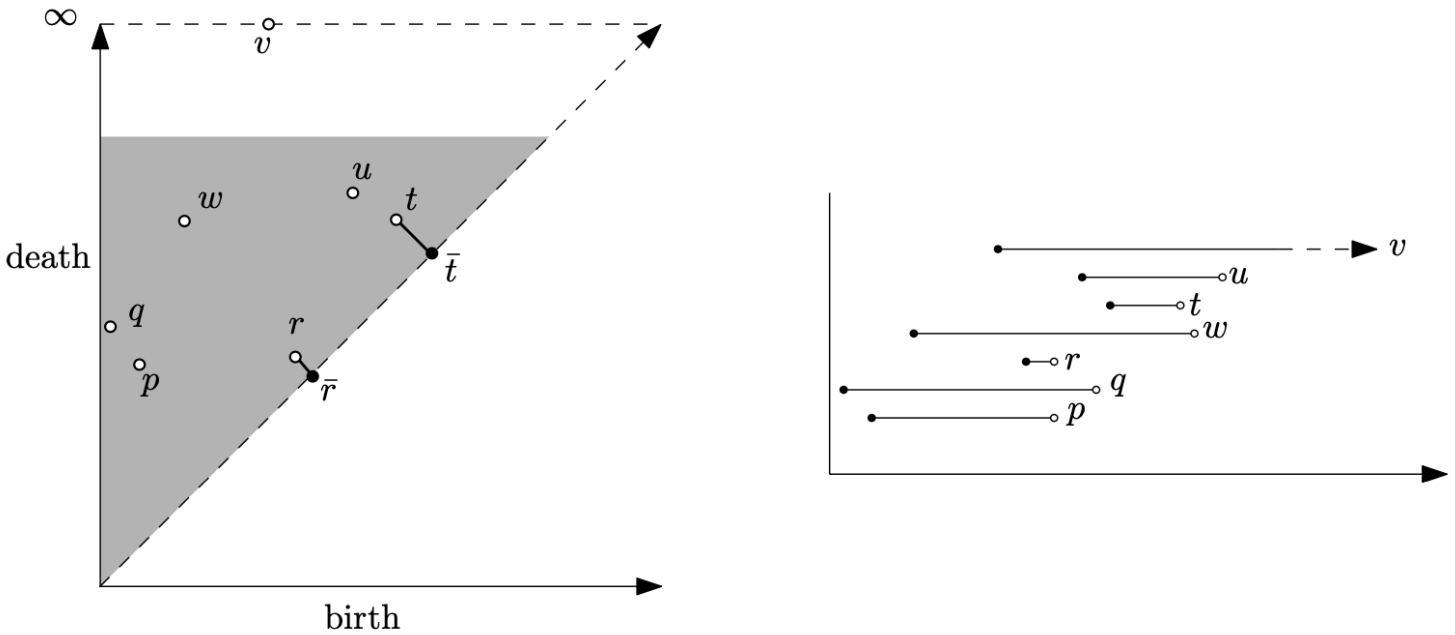

Figura: A la izquierda, diagrama de persistencia con puntos no diagonales sólo en el cuadrante positivo; a la derecha, el código de barras correspondiente. Imagen tomada de [1].

Diagramas de Persistencia

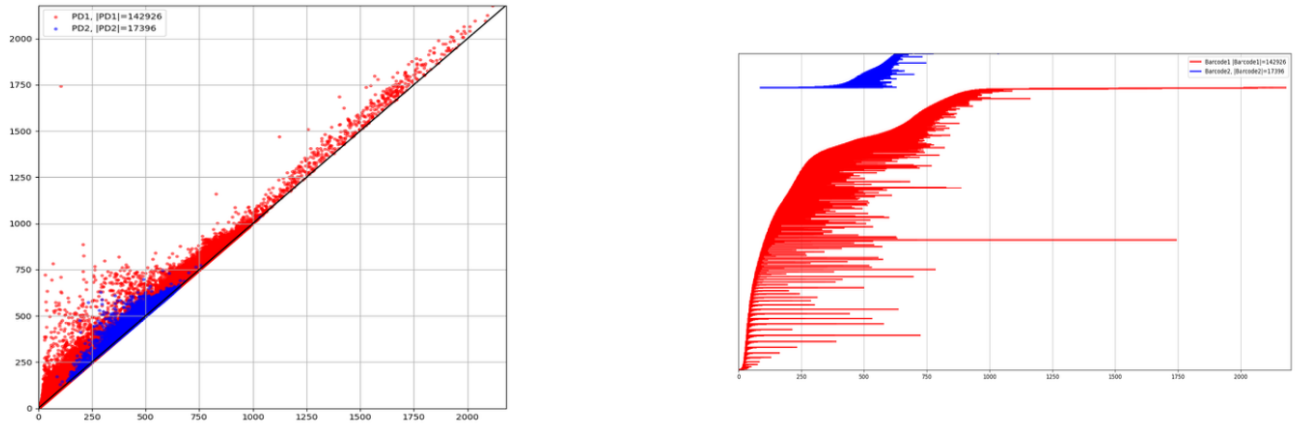

Figura: Diagramas de persistencia típicos y los códigos de barras correspondientes; el rojo y el azul corresponden a los diagramas de persistencia 0-ésima y 1-ésima, respectivamente. Las barras están ordenadas de abajo a arriba por orden creciente según su "hora" de nacimiento. Imagen tomada de [1].

Diagramas de Persistencia

Sea $\Pi = \{\pi : \mathrm{Dgm}_p(F_f) \to \mathrm{Dgm}_p(F_g)\}$ el conjunto de todas las biyecciones que van de $\mathrm{Dgm}_p(F_f)$ a $\mathrm{Dgm}_p(F_g)$. La **distancia del cuello de botella** entre los dos diagramas es:

$$d_B(\mathrm{Dgm}_p(F_f), \mathrm{Dgm}_p(F_g)) = \inf_{\pi \in \Pi} \sup_{x \in \mathrm{Dgm}_p(F_f)} \|x - \pi(x)\|_\infty$$

Diagramas de Persistencia

Para cualquier $p \geq 0$, $q \geq 1$, la distancia q-Wasserstein se define como

$$d_{W,q}(\mathrm{Dgm}_p(F_f), \mathrm{Dgm}_p(F_g)) = \inf_{\pi \in \Pi} \left(\sum_{x \in \mathrm{Dgm}_p(F_f)} \|x - \pi(x)\|^q \right)^{1/q}.$$

La distancia $d_{W,q}$ también es una métrica en el espacio de los diagramas de persistencia, al igual que la distancia del cuello de botella. Nótese que Π es el mismo conjunto de biyecciones que se definió en la diapositiva anterior.

Algoritmo Mapper

El **algoritmo mapper** es una técnica utilizada en el análisis topológico de datos (TDA) para visualizar la estructura de datos de alta dimensión. Combina conceptos de la topología algebraica y la teoría de grafos para simplificar y representar conjuntos de datos complejos.

Algoritmo Mapper

Este algoritmo toma un conjunto de datos de alta dimensión y lo proyecta en un espacio de menor dimensión utilizando una **función de proyección**. Luego, utilizando una **cubierta**, se cubre este espacio proyectado con una colección de conjuntos que se superponen y agrupa los puntos de datos dentro de cada conjunto utilizando algún **algoritmo de agrupamiento**. Finalmente, construye un **grafo** donde los nodos representan los grupos de datos y las aristas indican solapamiento de puntos entre grupos.

Algoritmo Mapper

1. **Proyección**: Aplicar una función de proyección para reducir la dimensionalidad.
2. **Cubierta**: Dividir el espacio proyectado en conjuntos solapados.
3. **Agrupamiento**: Agrupar los puntos de datos en cada conjunto.
4. **Construcción del Grafo**: Crear un nodo para cada grupo y aristas para solapamientos.

Algoritmo Mapper

Algunas de las aplicaciones del algoritmo mapper son las siguientes:

- Análisis de datos biomédicos.
- Segmentación de imágenes.
- Exploración de datos de redes sociales.
- Visualización de datos financieros.

Algoritmo Mapper

Sean X y Z espacios topológicos y sea $f : X \to Z$ un mapa continuo y bien comportado. Sea $U = \{U_\alpha\}_{\alpha \in A}$ una cubierta abierta finita de Z. El "mapper" que surge de estos datos se define como el nervio de la retracción $f^*(U)$ que cubre a X; es decir,

$$M(U, f) := N(f^*(U)).$$

Algoritmo Mapper

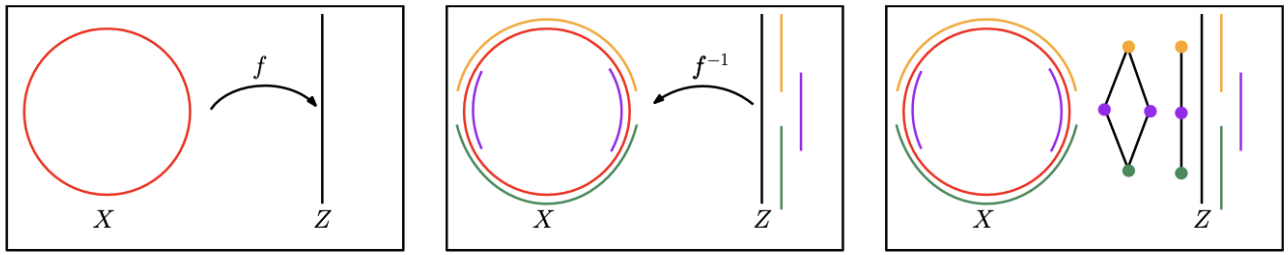

Figura: A la izquierda, un mapa $f : X \to Z$ de un círculo a un subconjunto $Z \subseteq \mathbb{R}$; en el centro, se muestra como el mapa inverso f^{-1} induce una cobertura del círculo a partir de una cobertura U de Z; a la derecha, los nervios de las dos coberturas de X y Z: el nervio a la izquierda (con forma de cuadrilátero) es el "mapper" inducido por f y U. Imagen tomada de [1].

Algoritmo Mapper

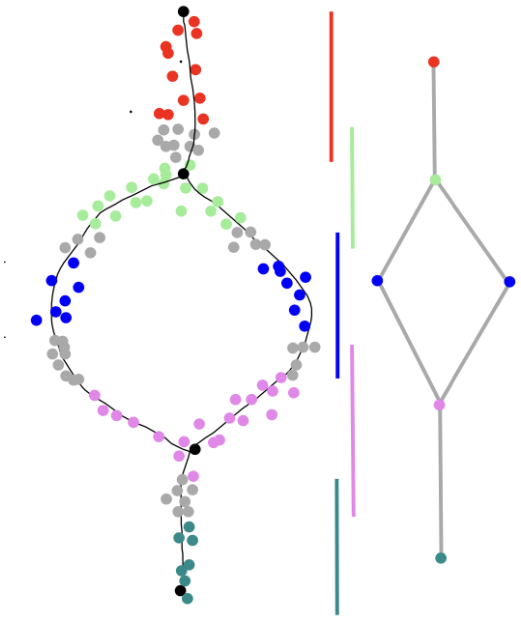

Figura: Un mapa $f : P \to Z$ de una nube de puntos P a un subconjunto $Z \subseteq \mathbb{R}$. Las coberturas de Z son intervalos de distintos colores; los puntos de P están coloreados con los colores de los intervalos, sin embargo, los puntos grises tienen valores en dos intervalos que se sobrelapan. Imagen tomada de [1].

Algunos sitios de interés

1. https://gjkoplik.github.io/pers-hom-examples/
2. https://giotto-ai.github.io/gtda-docs/latest/notebooks/vietoris_rips_quickstart.html
3. https://giotto-ai.github.io/gtda-docs/0.4.0/notebooks/mapper_quickstart.html

Figura: La topología es chida. Úsala en tu trabajo y te convertirás en el científico de datos más "cool" de la comarca. Imagen tomada de `http://www.math.lviv.ua/hearts/`.

BIBLIOGRAFÍA

1 Dey, Tamal Krishna, and Yusu Wang, *"Computational topology for data analysis,"* Cambridge University Press, 2022.

2 V. Robins, *"Computational topology at multiple resolutions: foundations and applications to fractals and dynamics,"* University of Colorado at Boulder, 2000.

3 Edelsbrunner, Herbert, and John L. Harer, *"Computational topology: an introduction,"* American Mathematical Society, 2022.

4 Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson, *"Topological methods for the analysis of high dimensional data sets and 3D object recognition*, In Proc. Eurographics Sympos. Point-Based Graphics (2007), pages 91?100, 2007.