NOTICIERO

ISSN 1514-9595 (web)

Miradas Matemáticas

Un procedimiento de cluster para conjuntos de funciones

Marcela Svarc

Universidad de San Andrés - CONICET

Introducción

El aprendizaje no supervisado ha recibido mucha atención en los últimos años. El aumento de la capacidad de almacenamiento de datos no solo ha incrementado la cantidad de información recopilada, sino que también ha modificado su naturaleza. Hace algunas décadas los datos a analizar eran típicamente observaciones reales o categóricas multivariadas, pero en la actualidad es común que los registros sean curvas o superficies. Ante este contexto es necesario volver a pensar los problemas estadísticos considerando la naturaleza de los datos de forma tal de poder extraer la información que tienen los valores observados de la manera más adecuada posible.

Uno de los grandes problemas que se plantean en el aprendizaje no supervisado es el *análisis de cluster*. En muchas ocasiones, para realizar un análisis de datos adecuado, es conveniente organizarlos en grupos homogéneos que se denominan **clusters.** Este tema ha sido extensamente estudiado en las últimas décadas y hay diversos enfoques para encarar el problema. Uno de los más extendidos y versátiles consiste en considerar una distancia entre los objetos a agrupar y, a partir de ella, construir los grupos. Esta construcción se realiza típicamente mediante un procedimiento jerárquico que comienza considerando que cada observación conforma un cluster, y en cada paso fusiona los dos grupos más próximos. Hay diversos criterios, llamados *enlaces*, para fusionar los dos grupos más próximos, basados en distancias entre grupos. Dentro de los más usuales se encuentra el *enlace completo*. Dados dos grupos C y C' el enlace completo está dado por

$$D(C, C') = \sup_{u \in C, v \in C'} \{d(u, v)\},$$

donde $d(u, v)$ es la distancia entre dos observaciones u y v. El procedimiento termina cuando todas las observaciones son fusionadas en un único cluster. El *dendrograma* ilustra gráficamente este procedimiento, ver Figura 1. En un dendrograma, cada hoja en la parte inferior representa una observación individual. A medida que se desplaza hacia arriba en el diagrama, las ramas se fusionan, representando la formación de clusters. La altura en la que dos ramas se fusionan indica la distancia entre los clusters o puntos que

se están fusionando. Cuanto más alta sea la fusión en el dendrograma, más disímiles son los clusters. Al cortar el dendrograma a una altura particular, queda determinada la conformación de los clusters dada por las componentes conexas del dendrograma.

Uno de los mayores desafíos que presenta el problema de cluster es determinar el *número de grupos*. Un criterio heurístico para determinar el número de grupos a partir de un dendrograma es encontrar una altura donde la distancia vertical sea *significativa*. También hay procedimientos analíticos para determinar el número de grupos y en problemas aplicados el conocimiento del experto es fundamental para tomar estas decisiones.

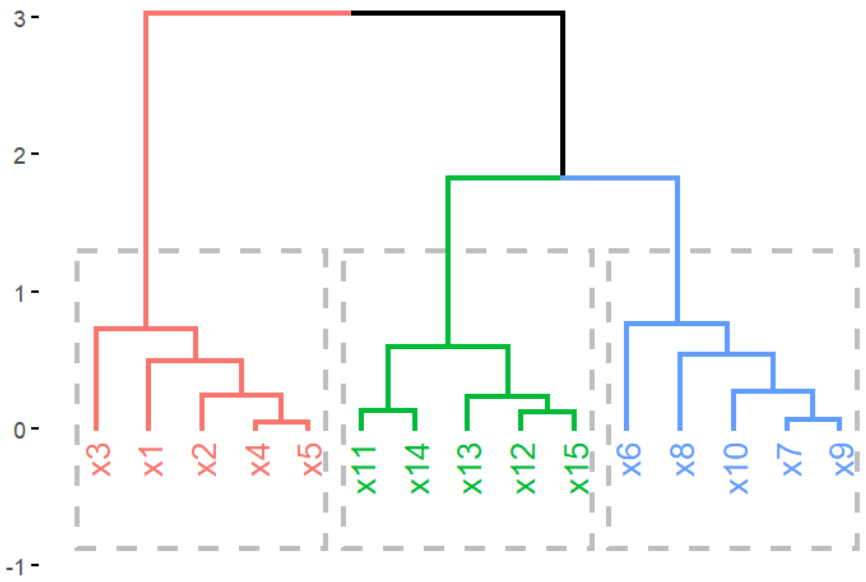

Figura 1: Dendrograma para un conjunto de 15 observaciones que conforman tres clusters, cada cluster se indica en un color diferente.

1. Descripción del problema de neurociencias

El problema de cluster tiene aplicaciones en muchos campos disciplinares de las ciencias, en la segmentación de mercados y de poblaciones. El ejemplo concreto que vamos a estudiar proviene del campo de las neurociencias. Una conjetura clásica es que el cerebro está constantemente estimando regularidades de secuencias de eventos. Uno de los interrogantes que surge en este contexto es si la mente es capaz de reconstruir las estructuras que generaron determinada información. Para tratar de resolver este interrogante se suelen llevar a cabo experimentos donde se expone a individuos a una sucesión de estímulos generados en forma estocástica, que pueden ser por ejemplo visuales [3] o auditivos [4]. Supongamos que se trabaja con una sucesión de estímulos, $\mathcal{U} = \{u_1, u_2, \dots, u_n\}$, pertenecientes a una familia finita de I elementos dividida en k grupos. Simultáneamente, se toman mediciones de electroencefalogramas (EEG). El objetivo es determinar si a partir de estas mediciones se puede recuperar los k grupos de $\mathcal{U}$.

Las señales del electroencefalograma (EEG) se pueden pensar como una función estocástica. A continuación esta señal es segmentada en intervalos de igual longitud, pudiéndose establecer una correspondencia entre cada uno de los estímulos de la sucesión $u_1, u_2, \dots, u_n$ y el correspondiente segmento de señal EEG $Y_1, Y_2, \dots, Y_n$ con $Y_i \in L^2[0, T]$, funciones medibles de cuadrado integrable. Luego, para cada estímulo $u \in \mathcal{U}$ se tiene un

conjunto de segmentos de señales de EEG, $\mathcal{Y}_N^u = \{Y_1^u, \dots, Y_N^u\}$. Consideremos el siguiente ejemplo sintético a modo ilustrativo, con $l = 5$ y $k = 2$. Sean $\mathcal{U} = \{u_1, \dots, u_5\}$ estímulos que ocurren de manera aleatoria, donde $\{u_1, u_2, u_3\}$ y $\{u_4, u_5\}$ conforman la partición que queremos detectar. En la parte superior de la Figura 2 **A** se encuentra la sucesión de eventos generados en forma aleatoria, debajo graficamos la señal de EEG, la segmentación está indicada mediante líneas punteadas. A cada estímulo le asignamos un color para simplificar la visualización. A partir de esta segmentación se obtienen los conjuntos de funciones, $\mathcal{Y}^{u_1}, \dots, \mathcal{Y}^{u_5}$, que corresponden a cada uno de los estímulos, como muestra la Figura 2 **B**.

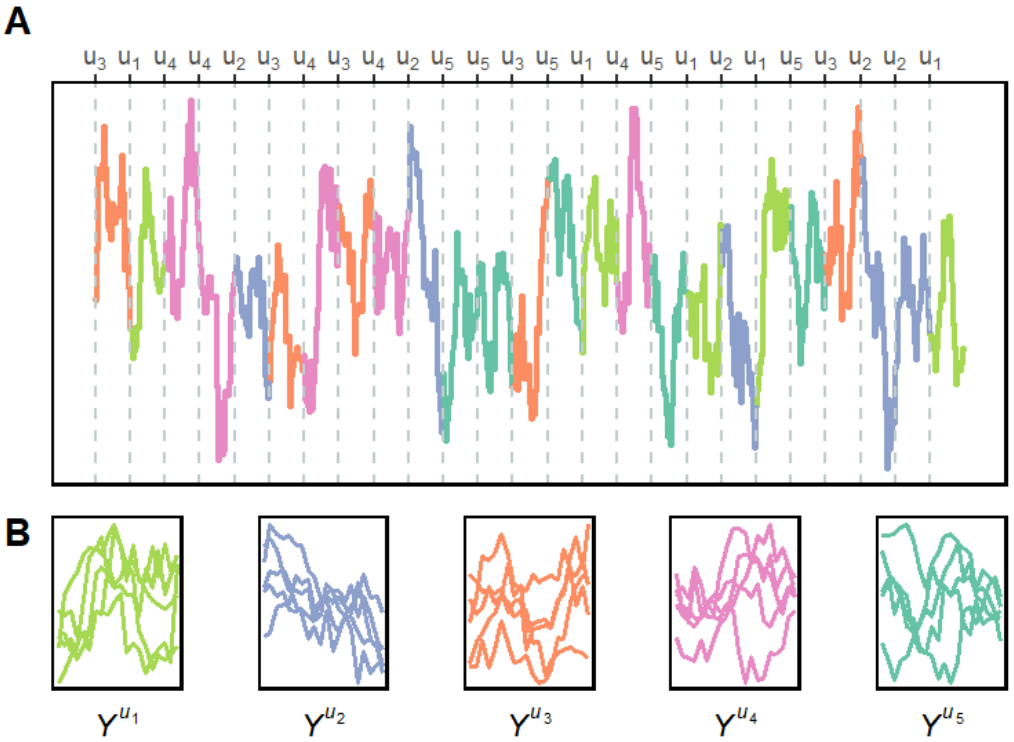

Figura 2: **A** Secuencia de estímulos y su correspondencia con la señal de EEG, a cada estímulo le corresponde un color diferente. **B** Segmentación del EEG para cada uno de los estímulos.

◎ *Nuestro objetivo es proponer un procedimiento de cluster para agrupar los conjuntos de funciones $\mathcal{Y}^u$ que hayan sido generados bajo la misma distribución de probabilidad.* En nuestro ejemplo, nos gustaría obtener los siguientes clusters $C_1 = \{\mathcal{Y}^{u_1}, \mathcal{Y}^{u_2}, \mathcal{Y}^{u_3}\}$ y $C_2 = \{\mathcal{Y}^{u_4}, \mathcal{Y}^{u_5}\}$.

2. El procedimiento de cluster

Para implementar un procedimiento de cluster jerárquico con enlace completo necesitamos considerar una *distancia entre medidas de probabilidad en* $L^2[0, T]$. Para esto consideraremos una adaptación infinito-dimensional del estadístico de Kolmogorov–Smirnov.

Sean Q^u y Q^v las medidas de probabilidad en $L^2[0, T]$ que generan las funciones dadas por el EEG correspondientes a los estímulos u y v. Se busca que los conjuntos de funciones $\mathcal{Y}_N^u$ e $\mathcal{Y}_N^v$ pertenezcan al mismo cluster si y solo si $Q^u = Q^v$. Siguiendo la propuesta de Cuesta-Albertos et al. [1] proponemos una distancia basada en proyecciones al azar.

Consideremos W una medida de probabilidad Gaussiana independiente de Q^u y de Q^v en $L^2[0, T]$. Sea h un elemento en $L^2[0, T]$ generado por W, llamemos Q_h^u a la distribución univariada de la variable aleatoria $\langle Y^u, h\rangle$ cuya función de distribución acumulada es $f^{u,h}$.

Luego, la distancia que proponemos entre dos medidas Q^u, Q^v en $L^2[0, T]$ es

$$D(u, v) = \int \|f^{u,h} - f^{v,h}\|_\infty dW(h). \tag{2.1}$$

Como $D(u, v)$ no es computable a partir de una muestra, la estimamos del siguiente modo. Para cada estímulo $u \in \mathcal{U}$, tenemos $\mathcal{Y}^u = \{Y_1^u, \dots, Y_N^u\}$ con $Y_N^u \in L^2[0, T]$ generada por la distribución Q^u. Así, $\mathcal{Y}^u$ es una muestra de funciones generadas bajo Q^u. Por simplicidad, asumimos que todos los conjuntos de datos tienen N funciones, aunque los argumentos pueden extenderse a casos con diferentes tamaños muestrales. La proyección de Y_n^u en la dirección h está dada por

$$R_n^{u,h} = \int_0^T h(t) Y_n^u(t) dt.$$

Luego, para cada estímulo u, al proyectar los elementos del conjunto $\mathcal{Y}_N^u$ en la dirección h tenemos el conjunto

$$\mathcal{Y}_N^{u,h} = \{R_n^{u,h} : Y_n^u \in \mathcal{Y}_N^u\}.$$

Siguiendo con el ejemplo sintético, la Figura 3 **A** muestra $\mathcal{Y}_N^{u_4,h}$ e $\mathcal{Y}_N^{u_5,h}$.

La *función de distribución acumulada empírica* de $\mathcal{Y}_N^{u,h}$ está dada por

$$\hat{f}_N^{u,h}(t) = \frac{1}{N} \sum_{R_n^{u,h} \in \mathcal{Y}_N^{u,h}} \mathbb{I}\{R_n^{u,h} \le t\}, \quad t \in \mathbb{R},$$

donde $\mathbb{I}\{A\}$ denota la función indicadora del conjunto A.

Estimamos $\|f^{u,h} - f^{v,h}\|_\infty$ en forma plug-in a partir de las muestras $\mathcal{Y}_N^u$ e $\mathcal{Y}_N^v$ considerando

$$D_N^{u,v}(h) = \sup_{t \in \mathbb{R}} \left\{ |\hat{f}_N^{u,h}(t) - \hat{f}_N^{v,h}(t)| \right\}. \tag{2.2}$$

La Figura 3 **B** muestra $\hat{f}^{u_4,h}$ y $\hat{f}^{u_5,h}$, la línea punteada indica $D_N^{u_4,u_5}(h)$ para el ejemplo. A partir de la ecuación (2.2) estimamos la ecuación(2.1): sean $B_1, \dots, B_M$, M realizaciones independientes de elementos de $L^2[0, T]$ generados con una medida Gaussiana, en nuestro caso consideramos el puente Browniano. Definimos en forma empírica la distancia entre los conjuntos $\mathcal{Y}_N^u$ y $\mathcal{Y}_N^v$ como

$$\hat{D}_{N,M}(u, v) = \frac{1}{M} \sum_{m=1}^{M} D_N^{u,v}(B_m). \tag{2.3}$$

Esta distancia es la que consideramos en el cluster jerárquico con enlace completo para determinar si dos elementos, que en este caso son conjuntos de observaciones $\mathcal{Y}_N^u$ con $u \in \mathcal{U}$, pertenecen o no al mismo cluster. En [2] vemos que $\widehat{D}_{N,M}(u, v)$ es un estimador consistente de $D(u, v)$. La Figura 3 **C** muestra como se construye la matriz de distancias entre los conjuntos $\mathcal{Y}_N^{u_1}, \dots, \mathcal{Y}_N^{u_1}$ que da lugar al dendrograma de la Figura 3 **D**.

3. Determinación del número de grupos

Finalmente, queda definir una cota para determinar el número de clusters. Comenzamos por notar que $\widehat{D}_{N,M}$ está estrechamente vinculado al estadístico tipo Kolmogorov–Smirnov

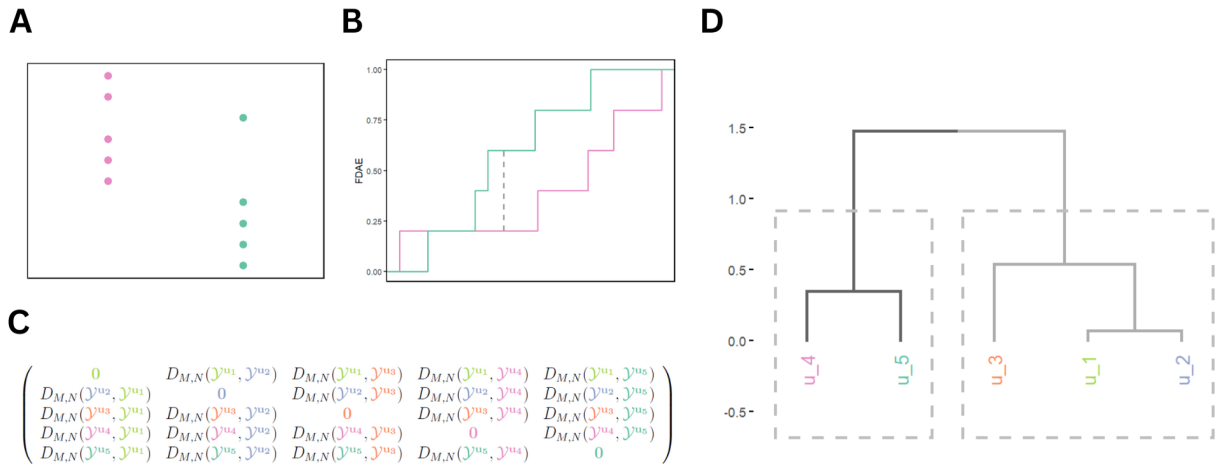

Figura 3: **A** El scatter plot muestra los datos $\mathcal{Y}^{u_4}$ en rosa e $\mathcal{Y}^{u_5}$ en verde proyectados en una dirección elegida al azar mediante un puente Browniano. **B** Función de distribución acumulada empírica de los conjuntos $\mathcal{Y}^{u_4}$ e $\mathcal{Y}^{u_5}$ proyectados en una dirección elegida al azar mediante un puente Browniano. **C** Matriz de distancia para los conjuntos de datos $\mathcal{Y}^{u_1}, \dots, \mathcal{Y}^{u_5}$ proyectados en M direcciones al azar. **D** Dendrogama indicando la conformación de grupos, en diferentes tonalidades de grises, se puede ver que la clusterización de los conjuntos de datos $\mathcal{Y}^{u_1}, \dots, \mathcal{Y}^{u_5}$ se corresponde con la clusterización de los estímulos $u_1, \dots, u_5$.

para realizar test de bondad de ajuste propuesto por Cuesta-Albertos et al [1]. Este test puede ser utilizado para determinar si dos muestras de datos multivariados o funcionales siguen la misma distribución. El procedimiento se basa en realizar proyecciones al azar de las observaciones y luego aplicar el test de Kolmogorov–Smirnov para datos univariados. Formalmente, para $\mathcal{Y}_N^u$ y $\mathcal{Y}_N^v$ dos muestras aleatorias en $L^2[0, T]$ se quiere determinar si fueron generadas por la misma distribución, i.e.,

$$H_0 : Q^u = Q^v \text{ vs } H_A : Q^u \neq Q^v.$$

El estadístico propuesto para el análisis es

$$KS\left(\hat{f}_N^{u,B}, \hat{f}_N^{v,B}\right) = \sqrt{\frac{N}{2}} D_N^{u,v}(B). \tag{2.4}$$

La hipótesis nula es rechazada a nivel α cuando $KS\left(\hat{f}_N^{u,B}, \hat{f}_N^{v,B}\right) > \eta_\alpha$. El valor crítico, η_α, se obtiene a partir de la distribución asintótica del estadístico (2.4). En nuestro caso buscamos rechazar la hipótesis nula cuando

$$\hat{D}_{M,N}(u, v) > \gamma_\alpha.$$

El valor crítico, γ_α, se puede obtener de manera asintótica como mostramos en [4].

Sin embargo, en general el tamaño muestral N no es suficientemente grande como para considerar resultados asintóticos. En el Teorema 1 de [2] hemos mostrado que la probabilidad de cometer un error grande entre la distancia estimada en la ecuación (2.3) y el verdadero valor dado por la ecuación (2.1) decae en forma exponencial. Basándonos en este resultado obtuvimos una cota de corte para el dendrograma que determina la partición buscada para el caso de muestra finita y que además tiene en cuenta la variabilidad del proceso. Esto nos permite concluir que dos conjuntos de funciones $\mathcal{Y}^u$ e $\mathcal{Y}^v$ no pertenecen al mismo cluster si

$$\hat{D}_{M,N}(u, v) > \gamma^*_{\alpha_N},$$

donde

$$\gamma^*_{\alpha_N} = \inf_{\delta \in (0,\alpha_N)} \left\{ \sqrt{\frac{2V^* \log(2/\delta)}{M}} + \sqrt{\frac{\log(\frac{e}{\alpha_N - \delta})}{N}} + \frac{7\log(2/\delta)}{3(M-1)} \right\} \tag{2.5}$$

siendo

$$V^* = \max_{(u,v)\in\mathcal{U}^2} \left\{ \hat{V}(u,v) \right\},$$

la máxima varianza entre todos los pares $(u, v) \in \mathcal{U}^2$, es decir,

$$\hat{V}(u,v) = \frac{1}{M-1} \sum_{m=1}^{M} \left[D_N^{u,v}(B_m) - \hat{D}_{M,N}(u,v) \right]^2.$$

El valor $\gamma^*_{\alpha_N}$ garantiza que el test tiene el nivel deseado y que, bajo condiciones de regularidad, el procedimiento recupera la partición real con alta probabilidad, esto puede verse en [2]. La Figura 3 **D** muestra los clusters que surgen de aplicar la cota que aparece en la ecuación (2.5).

4. Breve simulación

A continuación ilustraremos con una breve simulación. Generamos conjuntos de datos funcionales a partir del siguiente modelo que llamamos θ-scaled Brownian Bridge. Cada muestra fue generada del siguiente modo

$$Y_n^u(t) = \left(W(t) - \frac{t}{T} W(T) \right) \theta_u,$$

donde W es un proceso de Wiener, t se toma en forma equi-espaciada en una grilla de 80 puntos. Generamos 7 conjuntos de datos funcionales siguiendo 3 leyes de probabilidad, es decir, que tenemos tres grupos. Los parámetros θ elegidos para cada uno de los conjuntos de funciones son $(1, 1, 2, 2, 2, 4, 4)$. Para inicializar el procedimiento tenemos que fijar dos parámetros $\alpha_N = \sqrt{1/N}$, y $M_N = 30N$, consideramos conjuntos de datos de tamaño $N \in (40, 60, \ldots, 160)$, y realizamos 500 réplicas. Para $N = 40$ en más del $65\,\%$ de las réplicas obtuvimos la partición teórica, para tamaños muestrales mayores en más del $80\,\%$ de los casos conseguimos este resultado. Finalmente, alcanzamos un porcentaje superior al $90\,\%$ para conjuntos con 160 funciones.

Comentario final

Para concluir, en este breve artículo se encuentran los lineamientos generales de un par de trabajos realizados los últimos años junto a Antonio Galves, Fernando Najman y Claudia Vargas [4, 2], donde a partir de un interrogante de la neurociencia llegamos al problema de agrupar conjuntos de funciones en base a sus leyes de probabilidad. Bajo condiciones de regularidad en los procesos estocásticos hemos encontrado resultados

asintóticos y también para muestras finitas que nos permiten afirmar que con probabilidad alta el procedimiento propuesto recupera la estructura de la señal que lo generó.

Referencias

[1] Cuesta-Albertos, J. A., Fraiman, R., and Ransford, T. (2007). "Random projections and goodness-of-fit tests in infinite-dimensional spaces". *Bulletin of the Brazilian Mathematical Society, New Series* **37** 477-501.

[2] Galves, A., Najman, F., Svarc, M., and Claudia D. Vargas. "Clustering Sets of Functional Data by Similarity in Law." ArXiv, (2023). https://doi.org/10.48550/arXiv.2312.16656

[3] Hernández, N., Galves, A., García, J.E., Gubitoso M. D., and Vargas, C. "Probabilistic prediction and context tree identification in the Goalkeeper game". *Scientific Reports*, 14, 15467 (2024). https://doi.org/10.1038/s41598-024-66009-w.

[4] Najman, F. A., Galves, A., Svarc, M., and Vargas, C. D. "The Brain Uses Renewal Points to Model Random Sequences of Stimuli." ArXiv, (2023). https://doi.org/10.48550/arXiv.2311.07793

Ir al índice general