Teorema de clasificación de superficies topológicas

Junio 2020

Definición

S es superficie topológica si

- *Localmente homeomorfo a una bola en* $\mathbb{R}^2$.
- *Hausdorff, segundo numerable y conexo (*orientable).*

[IMAGEN DE SUPERFICIE COMPACTA Y NO COMPACTA]

Teorema de clasificación de superficies compactas orientables

- Suma conexa.
- Triangulación.

Definición

La suma conexa ($\#$) es un operador entre superficies.

$$S' = S_1 \# S_2$$

S' resulta de de retirar un disco abierto de cada superficie e identificarlas por el borde.

Ejemplo (EJEMPLO DE SUMA CONEXA)

- Es común tener que trabajar con datos categóricos, que no tienen un valor numérico. El **escalado multidimensional** permite transformar muestras de este tipo en puntos de un espacio euclídeo, tratando de conservar las distancias.
- Es necesario disponer de una matriz de distancias entre los individuos de una muestra:

$$\mathbf{\Delta} = \begin{pmatrix} \delta_{11} & \delta_{12} & \dots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \dots & \delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n1} & \delta_{n2} & \dots & \delta_{nn} \end{pmatrix}$$

- Si $\mathbf{\Delta}$ es compatible con una configuración euclídea, basta con tomar dicha configuración. Este es el **caso métrico**.
- **Teorema:** $\mathbf{\Delta}$ es compatible con una configuración euclídea si $\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{\Delta}^2\mathbf{H}$ es semidefinida positiva.
- Cualquier $\mathbf{Y}$ tal que $\mathbf{B} = \mathbf{Y}\mathbf{Y}^t$ sería una configuración euclídea compatible.
- Es MDS métrico se toma $\mathbf{Y} = \mathbf{U}\mathbf{\Lambda}^{1/2}$, que es valor que se obtendría al aplicar PCA sobre cualquiera de las configuraciones euclídeas compatibles con $\mathbf{\Delta}$.

- Si **B** no es semidefinida positiva, entonces no existe una configuración euclídea compatible con $\boldsymbol{\Delta}$.
- En este caso es necesario transformar $\boldsymbol{\Delta}$ conservando la relación entre distancias.
- Una transformación que permite obtener una **B** semidefinida positiva, si se elige una a adecuada, es la transformación q-aditiva:

$$\hat{\delta}_{ij}^2 = \begin{cases} \delta_{ij}^2 - 2a & \text{si } i \neq j \\ 0 & \text{si } i = j \end{cases}$$

- En el caso **caso no métrico** se aplica una transformación de este tipo para, posteriormente, poder utilizar la versión métrica.

- El escalado multidimensional también es compatible con matrices de similaridades.
- En general, una buena similaridad para un conjunto de datos con p_1 variables cuantitativas; p_2 variables binarias y p_3 variables categóricas es la de Gower:

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{p_1}(1 - \frac{|x_k - y_k|}{R_k}) + a + \alpha}{p_1 + (p_2 - d) + p_3}$$

  - R_k es el rango de la variable cuantitativa X_k, para $k = 1, ..., p_1$;
  - a es el número de coincidencias sobre 1 de las variables binarias;
  - d es el número de coincidencias sobre 0 de las variables binarias;
  - α es el número de coincidencias en las variables categóricas.

- **Regresión basada en distancias** es un método que permite incorporar la información de variables categóricas al modelo de regresión lineal.
- Consiste en aplicar regresión lineal por pesos sobre una configuración euclídea compatible con una matriz de distancias $\boldsymbol{\Delta}$.
- **Teorema:** La matriz de proyección del método de regresión no depende de la configuración empleada.
- Por tanto, la matriz de distancias determina totalmente los resultados de este método.

Regresión basada en distancias: Ejemplo (1)

- **Objetivo:** Predecir el precio de inmuebles en Iowa (EEUU).
- 66 variables regresoras: 36 categóricas y 30 numéricas.
- Errores obtenidos:

Método	Error cuadrático medio
Regresión lineal	2.15%
Regresión basada en distancias	1.27%

- Es un método de búsqueda de documentos en una base de datos $\mathcal{D}$, a partir de una serie de términos clave $\mathcal{T}$ y de una query de búsqueda $\mathbf{q}$.
- Utiliza SVD para capturar información implícita, lo que permite sobreponerse al problema de la sinonimia y relacionar conceptos cercanos.
- Parte de una matriz término-documento:

$$\mathbf{X} = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1p} \\ f_{21} & f_{22} & \dots & f_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \dots & f_{np} \end{pmatrix}$$

donde f_{ij} denota la "importancia" del término $t_i \in \mathcal{T}$ en el documento $\mathbf{d}_j \in \mathcal{D}$.

Indexación semántica latente (2)

- Si $\mathbf{X_k} = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^t$ es la SVD de $\mathbf{X}$ truncada a rango k:
  - $\mathbf{U}_k$ contiene representaciones vectoriales de k dimensiones para los términos de $\mathcal{T}$.
  - $\mathbf{V}_k$ contiene representaciones vectoriales de k dimensiones para los documentos de $\mathcal{D}$.
- Una vez obtenidas estas representaciones, se pueden medir las relaciones entre términos y documentos a partir de las distancia que los separa.
- Para proyectar la query $\mathbf{q}$ al mismo espacio, se halla $\mathbf{q}^t\mathbf{U_k}$.
- La búsqueda se puede realizar con la similitud coseno.

Indexación semántica latente: Ejemplo (1)

- LSI se aplica sobre los títulos de 15 libros de la biblioteca de la Escuela Politécnica y los de otros 15 de la Facultad de Psicología.
- Se seleccionan las palabras más relevantes y se construye la matriz término-documento empleando el bit de presencia como f. Extracto:

	D001	D002	D003	D004	D005	D006	D007	D008	D009	D010	D011	D012	D013	D014	D015
artificial	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0
cognition	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
human	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
intelligence	1	1	0	0	1	0	0	0	1	1	0	0	0	0	0
language	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0
learning	0	0	0	0	0	0	1	1	1	1	0	0	0	1	0
machine	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0
reasoning	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
robot	0	0	0	1	0	1	0	0	0	0	1	1	1	0	1

Matriz término-documento (Parte 1: Documentos EPS)

- La descomposición se trunca a 2 dimensiones.