Topological Data Analisys

Mauricio García Tec

Abstract

Análisis Topológico de Datos

1. Contenido

- Ayasdi - Análisis exploratorio. Extracting insights from the shape of complex data using topology
- Principal Component Analisys
- Maninfold Learning
- Regresión no Paramétrica con Teoría Morse
- Rankings Estadísticos con Teoría Hodge
- Teoría de Grafos

2. Topología

Se obtiene una idea a partir del análsis matemático o 'análsis situs' lo que se le llama análisis de cuerpos o locaciones, los cuales pueden ser objetos

Los espacios topológicos tienen propiedades, dentro de estas existe la noción de conjuntos abiertos y cerrados.

Lo que se busca es hacer un simil con la noción de continuidad como en r 2, objetos invariantes bajo transformaciones 'bonitas'

Definiciones

Un espacio topológico es un conjunto X y una familia τ de subconjuntos de X a los que les llamamos 'espacios abiertos' las propiedades que debe tener τ son las mismas que esperaríamos de los abiertos en RN

En $\Re$ los abiertos son intervalos de la forma (a, b) y las uniones entre ellos, pero en $\Re^N$ en general, los abiertos son uniones de las bolas de la forma $B(X_o, r) = \{x | d(X_o, X) < r\}$

Si tenemos una función $f : X \to Y$ entre espacios topológicos, se dice que es continua si $f^-1(u)$ es abierto *paratodou* abierto

En espacios métricos, la continuidad es:

$$\forall\ \epsilon > 0\ \exists\ \delta > 0\ \perp\ si ||x - x_o|| < \delta \Rightarrow ||f(x) - f(x_o)|| < \epsilon \tag{1}$$

'cosas cerca de X están cerca de Y' Si se rompen las cosas, se rompe la continuidad, pero si se pegan no La equivalencia en el mundo de la topología se llama Homeomorfismo (no rompe ni pega), cumple las condiciones si $f : X \to Y$ es un homeomorfismo si

- f es invertible
- f es contínua
- $f^(- 1)$ es continua

Entonces las propiedades topológocas que usamos son las invariantes bajo homeomorfismos

1) La más básica se llama componentes conexas El número de clusters no cambia -Conexidad, puedo arrastrar el lápiz para llegar de x a y 2) Homología: cantidad de hoyos $k - dimensionales$ Filosofia del maninfold (variedades) es la subclase favorita de los espacios topolócgicos Maninfold: es un espacio que localmente se parece a algún Rd donde d es la dimensión del maninfold.

3)Datos complejos X : no se sabe nada de ellos o se relacionan entre ellos

Ingredientes: i)Filtro $f : X \to Y$ donde $Y = R$ (usualmente) pensar en algo como cancer risk (puede ser cualquier variable de interés) ii) una partición del target space Y que me induce una cubierta en X iii) Método de Clustering que se aplica a cada método de la cubierta y conecto clusters que comparten individuos

iv) Método para convertir el resultado en lo anterior en un grafo

3. Principal Component Analysis PCA

-Limitación, es lineal -Queremos conocer la forma intrínseca de los datos 1)Supuestos: $E[X_i] = 0$ y $Var(X_i) = 1$ son lineas o planos que pasan por el origen, si los datos están enencontrados en 0 tiene sentido ajustar un subespacio lineal

Es una cuestión de modelado, se busca que la estructura de correlación entre los datos determine los componentes, no las unidades

2)El objetivo de PCA es maximizar la varianza $argmax\{var(XC)\}$ Matriz de varianzas y covarianzas

$$Var(X) = \frac{1}{n-1}(X^T X) \tag{2}$$

(si los datos están centrados en el origen)

$$Cov(X_i, X_j) = \frac{1}{n-1}\sum X_(k,i)X_(k,j) \tag{3}$$

por que la media es cero para todos los individuos pero si se cumple la observación $Var(X)c = \lambda c$ como

$$Var(Xc) = c^T Var(X)c = C^T \lambda C = \lambda C^T C = \lambda ||C||^2 \tag{4}$$

Se quiere el eigenvector del eigenvalor más grande, de hecho geométricamente la matriz $Var(X) = X^T X$ es una matriz simétrica, y si se grafican las curvas de nivel de la función tendrán forma de elipses -varianza explicada por cada componente Inercia o Varianza total =

$$\sum Var(X_j) = \sum diag[Var(x)] = traza[Var(X)] = \sum \lambda_j \tag{5}$$

-Diagnóstico

- Gráficas de varianza explicada
- Correlaciones entre componentes y variables originales para interpretar componentes
- Gráfico de componentes y se obtiene experienca de los datos originales
- Si es necesario se hace rotación (ya que se tienen los ejes, se rotan de manera que una componente esté asociada al menor número de variables posibles (facilita interpretación)

-Recordar el workflow 1 Datos 2 Auxiliar para partir los datos (filtro) 3 La 'cubierta' de los datos podría representarse 4 Clustering (numero de clusters, intervalos donde se comparten individuos, en las que si son las aristas, el número de individuos compartido es el peso 5 Matriz de adyacencia 6 Grafo 7 Interpretación de clusters traducido en dinero

4. Algoritmos de Clustering

a) db scan - Busca conjuntos densos en una vecindad de cada punto -captura idea de componentes conexas -es rápido -distingue formas -no se necesita saber el número de clusters -depende mucho de la métrica -sensible a los deltas de densidad

b) Self-organizing maps - algoritmos basados en redes neuronales, compiten entre ellos y el que gana se lleva a todos -sensibles a la topología -fácil de interpretar -alta dependencia de valores iniciales y métrica

c) Grid based algorithms -rapidez computacional -fácil de implementar -detecta clusters de diferentes tamaños y formas -problemas para detectar clusters que no tienen forma de rejilla

d) EM clustering for gaussian mixtures - varios centroides, asigna con una densidad gaussiana -distingue clusters que se empalman necesita número de clusters

e) Naive Bayes -No cae en la maldición de la dimensionalidad -No hace supuestos de la forma de la distribución -Es muy malo si las observaciones no son independientes -La máxima verosimilitu puede sobreajustar

5. Manifold Learning

Definición Motivación: Se tiene un ammnifold de dimensión d que está 'encajado' en un RT $\tau >> d$ se ve $\phi : \mu -> R$ además se tiene una 'muestra' $y_i = \bar{Y}$ de puntos de m que no se observa, pero si su imagen con ruido, entonces el problema principal de manifold es la reconstrucción entre dimensiones -distancoa geodésica

6. Kernel PCA

Como vimos anteriormente, el anlisis de componentes principales se basa en la transformación y diagonalización la matriz de covarianzas estimadas a partir de los datos x_k, $k = 1, \ldots, n$

Asumimos ahora que los datos que queremos mapear $\Phi(x_1), \ldots, \Phi(x_n)$ está centrado, es decir $\sum \Phi(x_n) = 0$ entonces hacemos PCS para la matriz de covarianzas

$$C = \frac{1}{n} \sum_{(} j = i)^n \Phi(x_j)\Phi(x_j)^T \tag{6}$$

las nuevas coordenadas en el Eigenvector, es decir las proyecciones ortogonales son llamadas componentes principales. Se puede generalizar para el caso no linear, suponemos mapear primero los datos en un espacio F como sigue:

$$\Phi : \Re^N \to F,\ x \to X \tag{7}$$

Se puede demostrar que aunque F tenga una dimensión arbitraria para diferentes casos de Φ, podemos siempre aplicar PCA en F. Esto basado en las diferentes funciones $Kernel$

7. Teoría Morse

Si se tiene una variedad M y una función $f : M- > R$ podemos estudiar la topología de M a traves de f (función morse), hay que fijarse en las curvas de nivel $f^-1(c)$ y de manera más general en los cinjuntos $f^-1((-inf, C))$

En términos de f, los puntos donde hay un cambio en topología donde el gradiente f es cero, son los puntos críticos de la función

funciones morse: Una función f es morse si sus ountos críticos son no degenerados. Un punto críticoes no degenerado si la hessiana en ese punto es invertible (si la hessiana induce una función cuadrática no degenerada en coordenadas locales)

Lema: toda función morse se puede expresar en coordenadas locales al rededor de un punto crítico p como

$$f(x) = f(p) - x_1^2 - x_2^2 - ... - x_j^2 + x_(j + 1)... + x_n^2 \tag{8}$$

j es el índice de la cuadrática, al rededor de un punto crítico no degenerado

corolario: Si f es una función morse y no tiene puntos críticos en $[a, b]$, entonces

$$f^-1((-inf, a))equivalentef^-1((-inf, b]) \tag{9}$$

Se mira como se parte en intervalos, de tal manera que los individuos que comparten los intervalos serán las aristas de los clusters

8. Homología Persistente

(de quí nace el TDA)

1) Complejos simpliciales -def: k-simplejo = $\{x \in R | \sum X_i = 1, x_i > 0\}$ 0-simplejo: punto 1-simplejo: triangulo lado 1

Complejo-Simplial = Simplejos de distintas dimensiones pegadas de manera bonita Un complejo simplicial abstracto es un conjunto finito de vŕtices $V = \{v_1, v_2..., v_\tau\}$ y una familia de subconjuntos de V tal que $\sigma \in \sum$ y $\tau subcon \sigma$ entonces $\tau \in \sum$ Dado un complejo cimplicial $(V, \sum)$

$$C_k = Free\{\sigma \in \sum, num(\sigma) = k\} \tag{10}$$

donde free es un grupo abeliano simple

-Grupo Libre de V con coeficientes en k (grupo abeliano) enteros, racionales, reales, cro-uno todas las comunidades lineales de objetos de V en coeficientes en k

va a resultar que los C_k están conectados por mapas d_k los cuales se llaman diferenciales

Se define así para $\sigma = \{v_i i, ..., v_i k\}$

$$d_k(\sigma) = \sigma(-1)^(i-1)\sigma\{vij\} \in C_(k-1) \tag{11}$$

Y en general se extiende linealmente

$$d_k(\sigma_1 + \sigma_2) = d_k\sigma_1 + d_k\sigma_2 \tag{12}$$

-Grupo cociente Sea G grupo, H¡G subgrupo, el grupo cociente G/H se obtiene de una relación de equivalencia $Xsimy$ si y solo si $x - y \in H$ ejemplo: $G = Znaturales = \{..., -2, -1, 0, 1, 2, ...\}$ $H = 5 - Znat = \{-10, -5, 0, 5, 10, ...\}$ $G/H = \{--0, 1, 2, 3, 4\}$ la suma es modulo 5

-def

$$kernel(\delta_k) = k - ciclos lm(\delta_k) = k - boundaries H_k = \frac{kernel(\delta_k}{lm(\delta_k h} = \frac{Z_k}{B_k} \tag{13}$$

es el mismo grupo de homología

def- B_k rank (H_k) es el k-esimo número de Betti, cuenta los hoyos k-dimensionales

honologia persistente -Sirve para contar hoyos (num de dimensiones) en un espacio topológico (triangulados) complejos simpliciales, esto por que se quiere conocer la 'forma' de los datos

-Output: Grupos Hk, k para cada dimensin, se reumen en los números de Betti k=num de vecinos -¿construir gráfica de k vecinos y para los faltantes 1) Elegir raíz entre faltantes 2) Encontrar componenete conexa 3) faltantes = faltantes - elementoe en componente conexa

```
BFS Algorithm
pseudoc\'odigo
bfs(G)
\{
List L = empty
Trre T = empty
choose a starting vertex x
search (x)
while (L none empty)
     remove edge (v.w) from start of L
         if w not yet visited
             \{
                 add(v,w) to T
                    search w
                    \}
             search(vertex v)
        visit (v)
    for each edge (v,w)
     add edge(v,w) to end of L
\}
```

1) Colas Son estructuras FIFO 'first in first out' (primero en tiempo primero en derecho)

2) Pilas –

3) Grafos: se quiere encontrar una componente conexa de un grafo, se quiere que dado un nodo raíz, resulten todos los nodos para los cuales existe un camino que empiece en el nodo raíz y termina en ese nodo.

Breadth -¿first search (usa cola) Depht -¿first search (usa pila)

Si lo aplicamos a datos: 1) Escoger una métrica y un ϵ 2) Convertir nube de puntos en un complejo simplicial 3) Cálculo de los $\beta_k = Rank(H_k)$ 4) Repetir para muchos ϵ, por eso se llama persistente 5) Se observan los features que persisten 6) Se crean gráficas de barras y gráficas de nacimiento y muerte

8.1. Historia

La topología y la Teoría de Grafos nacen al mismo tiempo -puentes de Koninsberg (1736) -Camino Eureliano (si empieza y termina igual es un circuito eureliano) -Entonces la representación gráfica (topologocamente equivalente) del problema nodos - componentes conexas -Euler: Solo existe solución a numero si a lo más hay 2 vértices de grado impar

[1] ...

[2]