Relación entre conjuntos de datos

Leonardo Andrés Jofré Flor

13 de septiembre de 2013

Resumen

El objetivo es encontrar la forma de relacionar n variables explicativas a una explicada de forma más robusta. Otro objetivo es el de encontrar funciones que relaciones variables aleatorias de forma alternativa a las sugeridas por la técnica de regresión lineal, considerando que si debe existir una relación entre variables aleatorias, esta debe conservar la transformación de variables aleatorias en término de las distribuciones de probabilidad. Esto quiere decir que conociendo la información de muestras, mediante este método se puede linealizar de tal manera que la matriz de correlación pueda encontrar relaciones no lineales entre variables aleatorias.

1. Preliminares

1.1. Caso elemental para dos variables

Consideremos variables aleatorias X e Y tal que una es una transformación de la otra

$$X \sim p_x(t)$$
$$Y = f(X) \sim p_y(x)$$

$$f(t) = F_Y^{-1}(F_X(t))$$

la función de densidad del error

$$\epsilon = F_Y^{-1}(F_X(X)) - Y = X - F_X^{-1}(F_Y(Y))$$

de lo que se deduce la primera medida de relación entre variables aleatorias

primera medida de relacion entre variables aleatorias Dos variables aleatorias están relacionadas si $X + Y = F_Y^{-1}(F_X(X)) + F_X^{-1}(F_Y(Y))$

teorema: $\mathbb{E}[\epsilon] = 0$

1.2. Relación entre dos variables dado un cambio de base

1.3. Considerando que f no es biyectiva

Si consideramos

$$V = \alpha X + \beta Y$$

y

$$W = \gamma X + \delta Y$$

Por lo que el error queda definido como

$$\epsilon = \|F_V^{-1}(F_W(W)) - V\|$$

Se pueden considerar como nuevas variables a relacionar mediante el procedimiento anterior.

2. Regresión lineal asociada

Es deseable saber que tan relacionadas están un conjunto de variables

$$\min_{\{\alpha\in\}} \sum_{i\neq j} \alpha_{i,j} \|F_{X_i}^{-1}(F_{X_j}(x_{jk})) - x_{ik}\|$$

por lo que ahora se ha de encontrar los valores de α_k que minimicen la función de error.

Con las siguientes restricciones $\alpha_{ij} = \alpha_{ji}$ y también $\alpha_{ii} = 0$, además debe ser una ponderación por lo que se debe cumplir que

$$\sum_{i,j} \alpha_{i,j} = 1$$

3. Grafo ponderado de relacion entre variables aleatorias

Consideremos la siguiente ecuación

$$\sum_{i,j} \alpha_{i,j} F_{X_i}^{-1}(F_{X_j}(t)) = 0$$

Se puede encontrar la matriz A de todos los coeficientes $\alpha_{i,j}$ que minimizan la l_2

4. Árbol generador máximo

El arbol generador máximo nos dice la dependencia entre las variables encontrando como raiz los elementos mas dependientes y las hojas las variables más independientes.

5. Linealización de las relaciones

6. Cambio de base

La idea ahora es almacenar la información relacionada entre las distintas variables en un conjunto mucho menor de variables.

7. versión combinacional

8. costo computacional

Una se las cosas interesantes de este método es el costo computacional, que es el costo de ordenar los datos

9. Convergencia a la ortogonalidad

Ecuaciones diferenciales Consideremos unas nuevas columnas con diferenciales de los datos reales, otras columnas también con las sumas hasta ese punto, o sea, las integrales. A partir de este metodo se pueden generar nuevas ecuaciones que serán diferenciales con coeficientes estocásticos.

Ejemplo

$X_1 \sim \mathcal{N}(\mu = 0, \sigma = 1)$ $X_2 = 3X_1^2 + 2X_1 + 5 + \mathcal{N}(\mu = 0, \sigma = 1)$ en la cual claramente no existe una relación lineal entre las variables, la pregunta es si se puede generar una relación entre las variables como una ecuación diferencial.

definición: Inversas parciales de una función

Existencia de las inversas parciales

Conjunto solución

Unicidad

Conversión del problema a una ecuación diferencial no homogenea

funcion de green de la ecuación diferencial