Problemas de relación entre atributos: Regresión

Maria-Amparo Vila
vila@decsai.ugr.es

Grupo de Investigación en Bases de Datos y Sistemas de Información Inteligentes https://idbis.ugr.es/
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada

Regresión: ideas básicas

Modelización de dependencias

- *Objetivo:* Describir dependencias significativas entre las variables incluidas en la base de datos Los modelos de dependencias pueden ser:
  - Cualitativas o cuantitativas (dependencias funcionales y análisis de regresión)
  - Dependencias parciales o completas

Cuando se trata de variables cuantitativas, y se espera la existencia de una relación $y = f(x_1, ..x_n)$ tenemos un **modelo predictivo** *normalmente de análisis de regresión*

Cuando no se tiene conocimiento previo, las variables son más generales y se buscan asociaciones entre valores tenemos un **modelo descriptivo**

Regresión: ideas básicas

La regresión como un modelo de predicción

Definición

Clasificación *es el proceso de aprender una función que aplica un conjunto de atributos* $X_1..X_n$ *en otro atributo* Y. *Si:*

- *Si* Y *es discreta, booleana, nominal etc. tenemos Modelos de clasificación propiamente dichos*
- *Si* Y *es continua tenemos Modelos de regresión*

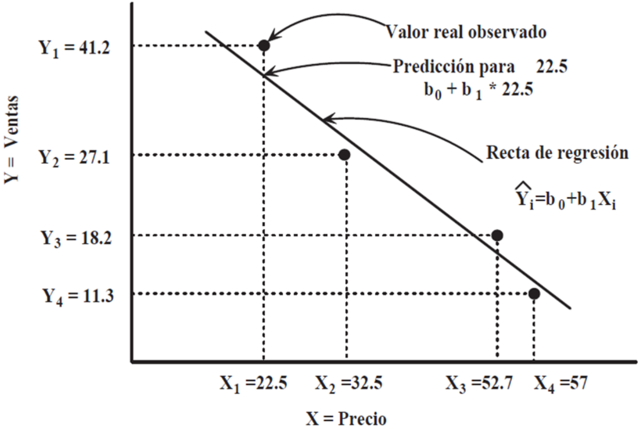
La funcion que se aprende se denomina también Modelo de clasificación en general

La regresión es un modelo de predicción de variables continuas, habitualmente las variables independientes tambien son continuas o al menos numéricas

Regresión lineal simple

Modelo básico

$$\hat{Y}_i = b_0 + b_1 X_i$$

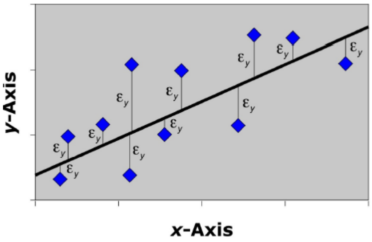


Regresión lineal simple

Modelo básico

Problema

- Dados $\{x_1..x_n\}$ $\{y_1,..,y_n\}$ obtener a y b de forma que $y = ax + b$ se ajuste a los datos.
- **Solución: mínimos cuadrados**



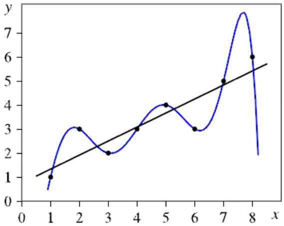
- Formalmente: $MinF(a,b) = \sum_{i=1}^{n}(ax_i + b - y_i)^2$

Extensiones de la regresión lineal

Regresión polinomial

Modelo

- Ajustar $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_n x^n$
- Es importante ajustar el valor de n, ya que dados m puntos se puede encontrar una curva de grado $m - 1$ que pase por todos ellos. (Sobreaprendizaje)



Extensiones de la regresión lineal

Regresión multivariante

Problema

- En este caso se trata de m variables in dependientes para y es decir tenemos el data set.

items\variables	x_1	x_2	$\cdots\cdots$	x_n	y
o_1	x_{11}	x_{12}	$\cdots\cdots$	x_{1n}	y_1
$\vdots$	$\vdots$	$\vdots$	$\ldots\ldots$	$\vdots$	$\vdots$
o_M	x_{M1}	x_{M2}	$\cdots\cdots$	x_{Mn}	y_n

y queremos encontrar $a_0, a_1 ... a_n, b$ tal que:

$$y = a_0 + a_1 x_1 + ... a_n x_n$$

- **Solución: mínimos cuadrados**

$$MinF(a_0, a_1, ..a_n) = \sum_{i=1}^{M} (a_0 + \sum_{j=1}^{n} a_j x_{ij} - y_i)^2$$

Extensiones de la regresión lineal

Regresión multivariante

- Una vez ajustada la recta de regresión los coeficientes de la misma se pueden intepretar en términos del **peso** que tiene cada variable independiente en la variable dependiente.
- En regresión lineal $y = a_0 + a_1 x_1$, a_1 mide lo que debe aumentar/disminuir x_1 para que Y aumente una unidad

Bondad de la regresión lineal

Coeficientes de correlación

Coeficiente de correlación simple *usado para medir la relación lineal entre dos variables*

$$r_{xy} = \frac{\sum_{i=1}^{M}(x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{M}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{M}(y_i - \bar{y})^2}}$$

- $r_{xy} \in [-1, 1]$
- Si $r_{xy} \simeq 1$ existe *correlacion lineal positiva entre x e y*
- Si $r_{xy} \simeq -1$ existe *correlacion lineal negativa entre x e y*

- Un valor más ajustado de la la correlación es (r_{xy}^2). Este es un caso particular del *coeficiente de determinación de la regresión múltiple*

Bondad de la regresión lineal

Coeficientes de correlación

Coeficiente de correlación múltiple *usado Para medir la relación lineal entre una variable dependiente y n variables independientes*

Coeficientes de correlación parcial
Dadas $1, 2, .., n$ variables, supongamos que la variable dependiente es la 1, $r_{1k \cdot 2345.(k-1)(k+1)..n}$ representa el coeficiente de correlación que obtendríamos si fijamos todas las variables y calculamos la correlacion entre las variables 1 y k

El *coeficiente de correlacion múltiple* $R_1 \cdot 234..n$ se calcula de forma iterativa a partir de la regresión parcial

Bondad de la regresión lineal

Coeficientes de correlación

$$R_{1\cdot 23} = \sqrt{(r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23})/(1 - r_{23}^2)}$$

$$R_{1\cdot 234} = \sqrt{1 - [(1 - r_{12}^2)(1 - r_{13\cdot 2}^2)(1 - r_{14\cdot 23}^2)}$$

$$\vdots$$

$$R_{1\cdot 234..n} = \sqrt{1 - [(1 - r_{12}^2)(1 - r_{13\cdot 2}^2)(1 - r_{14\cdot 23}^2)...(1 - r_{1n\cdot 23..(n-1)}^2)}$$

R^2 es el coeficiente de determinación

R^2 mide la proporción varianza de de y explicada por $x_1..x_n$
debe ser próximo a 1

$\bar{R}^2 = 1 - \frac{M-1}{M-n}(1 - R^2)$ es el coeficiente de determinación ajustado

Se utiliza cuando el número de variables es muy grande

El proceso de regresión lineal

Cuando entre las variables independientes hay una relación lineal aparece la *Multicolinealidad*. *Se debe suprimir una de las variables relacionadas*

Para hacer un análisis de regresión correcto:

1. Análisis exploratorio: utilizar los diagramas de "scatter" para visualizar una cierta "dependencia lineal parcial" entre la variable dependiente y las independientes
2. Buscar posibles relaciones de multicolinealidad entre las variables independientes, $R^2 \succeq 0.65$ es un buen límite. Eliminar variables relacionadas dejando sólo una de ellas
3. Realizar el análisis de regresión, calculando R^2 y $\bar{R}^2$, nuevamente $0,65$ es un buen valor de ajuste.

Extensiones de la regresión lineal

Tranformaciones numéricas

- Si hacemos: $x^2 = x_2, ..x^n = x_n$ transformamos :

  $$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_n x^n \Longrightarrow y = a_0 + a_1 x_1 + ... a_n x_n$$

- Si tenemos $y = ax^b$ como posible función se puede ajustar $lny = ln + blnx$ Hay que tener en cuenta que realment se ajustan por mínimos cuadrados en el espacio transformado; pero puede ser una buena aproximación.

Extensiones de la regresión lineal

Regresión logística

- Otra función importante es la *función logística* donde se ajusta:

$$Z = ln(p/(1-p)) = B_0 + B_1x_1 + ...B_nx_n$$

- La regresión logística se usa para predecir la probabilidad de ocurrencia de una variable binaria. Es decir para predecir la probabilidad de pertenencia o no a una determinada clase (fallo no-fallo, pago impago etc.) es muy usada en control de calidad y en análisis de riesgo.
- Las medidas de calidad de la regresión logística se basan en la diferencia al cuadrado de las predicciones y los datos reales ya que en este caso las medidas basadas en correlación no funcionan bien, puesto que trabajamos realmente con funciones no lineales.

Extensiones de la regresión lineal

Regresión logística

- Teniendo en cuenta que p es probabilidad de éxito $p/(1-p)$ es la razón entre la probabilidad de éxito y de fracaso. Este valor se denomina $odds$, y por ejemplo si $odds = 4$ esto quiere decir que tenemos cuatro veces más probabilidad de éxito que de fracaso.
- En realidad lo que se estima es $ln(odds)$, por lo que, para interpretar los coeficientes de regresión, $\{B_i\}$ debemos considerar $exp(B_i)$ (EB_i) en algunas salidas. EB_i representa el peso que tiene la variable X_i en $odds$ es decir en el valor de la razón entre el éxito y el fracaso.

Extensiones de la regresión lineal

Regresión logística

- Es posible incluso extender el modelo para variables catagóricas no binarias utilizando la **regresión multinomial** que extiende los distintos valores de la variable objetivo transformándolos en una variable binaria multivariante.
  Por ejemplo

$$Y = 1, 2, 3, 4 \Leftrightarrow (Y_1, Y_2) = \{(0,0), (1,0), (0,1), (1,1)\}$$