Curso Propedéutico en Ciencias de Datos ITAM

Mauricio García Tec

2017-07-14

Tarea 3: Regresión Lineal

A continuación están los detalles de la tarea 3. Deben entregarla en formato PDF (pueden hacer un Markdown, Latex o un documento de Word y guardarlos como pdf) y ponerla con el resto de sus tarea en su carpeta del Github con el nombre `tarea3.pdf` . Aceptaré el último pull request el martes 18 a las 2:00pm, no hay excepciones pues debo entregar calificaciones ese mismo día.

Parte teórica. Esta parte del proyecto será sobre regresión lineal. Supongamos que quieren explicar una variable estadística Y (por ejemplo altura) utilizando la información de p variables $X^1, \ldots, X^p$ (peso, ancho de huesos, etc.). Si se toma una muestra de N individuos, cada variable está representada por un vector de tamaño N . La información de las variables explicativas se pueden juntar en una matriz

$$X = \left[X^1 \mid \ldots \mid X^p\right]$$

de tamaño $n \times p$ donde cada columna es una variable y cada fila uno de los individuos de la muestra. Tienen que contestar lo siguiente:

- Plantear el problema de regresión como un problema de mínimos cuadrados, encontrar el vector $\hat{\beta} = [\hat{\beta}_1, \ldots, \hat{\beta}_p]^\top$ que resuelva

  $$\hat{\beta} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2$$

  y encontrar la solución teórica. ¿Por qué este planteamiento nos da un ajuste lineal a nuestros datos? ¿Podríamos usarlo para ajustar polinomios (ej $y = x^2$)?
- Argumentar la relación entre la solución encontrada y un problema de proyección en subespacios vectoriales de álgebra lineal. ¿Cuál es la relación particular con el teorema de Pitágoras?
- ¿Qué logramos al agregar una columna de unos en la matriz X ? Es decir, definir mejor

  $$X = \left[\mathbf{1}_n \mid X^1 \mid \ldots \mid X^p\right]$$

  con $\mathbf{1}_n = [1, 1, \ldots, 1]^\top$.
- Plantear el problema de regresión ahora como un problema de estadística

  $$Y_i = \beta_0 + \beta_1 X_i^1 + \ldots + \beta_p X_i^p + \epsilon_i$$

  donde los errores son no correlacionados con distribución

  $$\epsilon_i \sim N(0, \sigma^2)$$

- ¿Cuál es la función de verosimilitud del problema anterior? **Hint:** empiecen por escribir el problema como

$$Y = X\beta + \epsilon$$

con

$$\epsilon \sim N(0, \sigma^2 I_n)$$

con I_n la matriz identidad. Y concluyan entonces que

$$Y \sim N(X\beta, \sigma^2 I_n)$$

Escriban entonces la verosimilitud como $L(\beta, \sigma^2) = f(Y \mid \beta, \sigma^2, X)$.

- Mostrar que la solución de máxima verosimilitud es la misma que la del problema de mínimos cuadrados.
- Investiga el contenido del Teorema de Gauss-Markov sobre minimos cuadrados.

Parte aplicada. Para esta parte pueden usar la base de datos `diamonds` que sugieron, aunque hay puntos adicionales si usan alguna base original interesante.

Cargar la base `diamonds` que se encuentra en el paquete `ggplot2` . Los comandos que pueden usar para cargar la base `diamonds` a su ambiente de trabajo en `R` son:

```
# install.packages("ggplot2") # solo si necesario...
library(ggplot2)
data(diamonds)
head(diamonds)
```

```
## # A tibble: 6 × 10
##   carat       cut color clarity depth table price     x     y     z
##   <dbl>     <ord> <ord>   <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23     Ideal     E     SI2  61.5    55   326  3.95  3.98  2.43
## 2  0.21   Premium     E     SI1  59.8    61   326  3.89  3.84  2.31
## 3  0.23      Good     E     VS1  56.9    65   327  4.05  4.07  2.31
## 4  0.29   Premium     I     VS2  62.4    58   334  4.20  4.23  2.63
## 5  0.31      Good     J     SI2  63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good     J    VVS2  62.8    57   336  3.94  3.96  2.48
```

Posteriormente deben hacer una regresión lineal. Su objetivo es explicar la variable `price` usando las demás variables. Noten que algunas variables no son numéricas, por lo que no pueden incluirse en un análisis crudo de regresión lineal. Para este proyecto *NO* es necesario saber transformar las variables no numéricas para poder usarlas en la regresión; hacerlo es optativo, de hecho, las paqueterías lo hacen por ustedes pero deben ser cuidadosos. Pueden usar la función `lm` de `R` para su análisis de regresión.

- ¿Qué tan bueno fue el ajuste? Una buena respuesta incluye argumentaciones teóricas y visualizaciones. Puntos adicionales si investigan como usar alguna de las librerias `ggplot2` o `plotly` para sus gráficas.
- ¿Qué medida puede ayudarnos a saber la calidad del ajuste? ¿Cuál fue el valor de σ^2 que ajustó su modelo y que relación tiene con la calidad del ajuste?
- ¿Cuál es el ángulo entre Y y $\hat{Y}$?. Hint: usen la R^2 y el arcocoseno.
- Defininan una funcion que calcule la **log**verosimilitud de unos parámetros β y σ^2 .
- Utilicen la función `optim` de `R` para numéricamente el máximo de la función de verosimilitud. Si lo hacen correctamente, su solución debe coincidir con la del método `lm` .

¡Buena suerte!