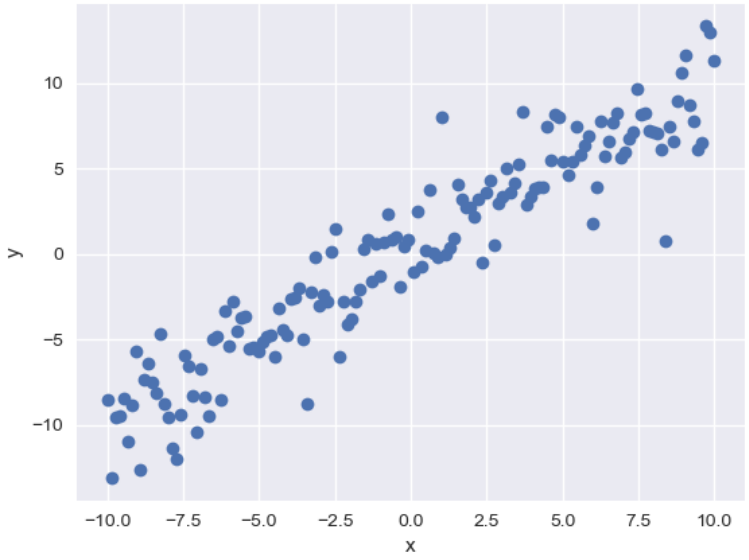
Regresión Lineal

Carlos Malanche

13 de marzo de 2018

Ok, regresemos a lo que se tenía la clase pasada. El problema es encontrar una función estimadora $\hat{f}$ que aproxime f dadas las parejas de datos $\{x_i, y_i = f(x_0) + \epsilon\}_{i=1}^{n}$ donde, sin conocimiento previo, ϵ es ruido de distribución Gaussiana con varianza σ^2 y primer momento $\mu = 0$. Bien, pues supongamos que tomamos nuestra serie de pares de datos, y la graficamos, obteniendo lo siguiente:



Pues, a reserva de abstenciones, yo voto que es una línea con ruido, sí o no raza?

1. Caso de una variable

Por eso mismo sería una buena idea proponer $\hat{f}(x) = mx + b$. Ahora, dada la serie de datos, cómo estimamos los parámetros m y b?

En la clase anterior vimos que la función de pérdidas cuadrática (para funciones escalares) está definida como

$$J(\hat{f}) = \frac{1}{2n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2 \tag{1}$$

En nuesro caso ya no es necesario escribir J como un funcional, ya sabemos de qué variables va a depender el costo

$$J(m, b) = \frac{1}{2n}\sum_{i=1}^{n}(mx_i + b - y_i)^2 \tag{2}$$

Suena como una buena idea derivar la función de costo respecto a cada variable e igualar a cero para encontrar un *mínimo*. Primero respecto a m

$$\frac{\partial J}{\partial m}(m, b) = \frac{1}{n}\sum_{i=1}^{n}(mx_i + b - y_i)x_i = 0 \tag{3}$$

Y ahora respecto a b

$$\frac{\partial J}{\partial b}(m,b) = \frac{1}{n}\sum_{i=1}^{n}(mx_i + b - y_i) = 0 \tag{4}$$

Esto es fácil de resolver, pues tenemos un sistema de dos ecuaciones con dos incógnitas, siendo la ecuación lineal en ambas variables. Abriremos la suma y factorizaremos para que sea más obvio (además de multiplicar ambas ecuaciones por n).

$$(\sum_{i=1}^{n} x_i^2)m + (\sum_{i=1}^{n} x_i)b = \sum_{i=1}^{n} x_i y_i \tag{5}$$

$$(\sum_{i=1}^{n} x_i)m + (n)b = \sum_{i=1}^{n} y_i \tag{6}$$

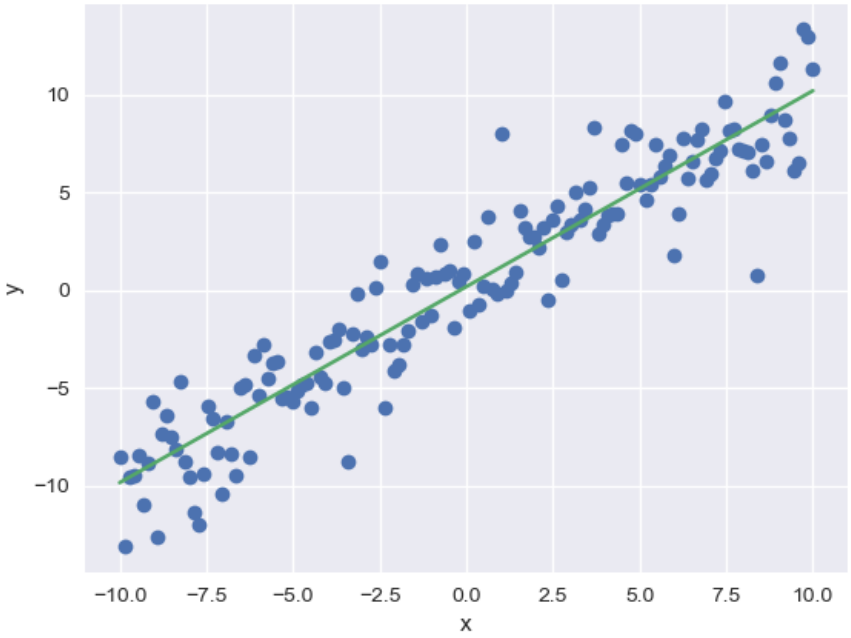
Pueden usar el método que más les guste para resolver este sistema. Por determinantes por ejemplo

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{Cov(x,y)}{Var(x)} \tag{7}$$

Qué curioso resultado. Ahora veamos que pasa con b

$$b = \frac{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\overline{(x \odot x)}\bar{y} - \bar{x}\overline{(x \odot y)}}{Var(x)} \tag{8}$$

Con eso, podemos construír la línea de predicciones del modelo, la cual se ve así



2. Caso multivariado

Antes de seguir, hay que definir las derivadas de escalares respecto a vectores

Definición. *La derivada de un función vectorial* $\boldsymbol{y}(x)$ *respecto a un escalar* x *será un vector renglón*

$$\frac{\partial \boldsymbol{y}}{\partial x} = \left[\frac{\partial y_1}{\partial x}, \frac{\partial y_2}{\partial x}, \cdots, \frac{\partial y_m}{\partial x}\right]$$

Definición. *La derivada de un función escalar $y(\boldsymbol{x})$ respecto a un vector $\boldsymbol{x}$ será un vector columna*

$$\frac{\partial y}{\partial \boldsymbol{x}} = \Big[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \cdots, \frac{\partial y}{\partial x_m}\Big]^T$$

Definición. *De las anteriores dos definiciones, se sigue que la derivada de una función vectorial $\boldsymbol{y}(\boldsymbol{x})$ respecto a un vector $\boldsymbol{x}$ es una matriz cuyas entradas están dadas por*

$$(\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}})_{ij} = \frac{\partial y_j}{\partial x_i}$$

Una implicación directa de esta convención (conocida como *denominator layout*, o formulción Hessiana) es que al derivar un producto de matrices, aparece una operación de transposición:

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x}) = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}}\boldsymbol{A}^T = \boldsymbol{A}^T$$

Bien, lo que sigue es escribir la generalización del problema para $\boldsymbol{x}$ un vector en $\mathbb{R}^m$ en lugar de un escalar. Podemos escribir un plano en m dimensiones como el siguiente producto interior

$$\hat{f}(\underset{\rightarrow}{x}) = \underset{\rightarrow}{c} \cdot \underset{\rightarrow}{x} = \boldsymbol{c}^T\boldsymbol{x} \tag{9}$$

donde $\underset{\rightarrow}{c}$ contiene los coeficientes a determinar. Falta decir algo sobre esta función, pero lo arreglaremos después de hacer un poco de trabajo.

Nuestra función de costo nuevamente ya tiene variables de las que depende, y podemos escribirla como

$$J(\underset{\rightarrow}{c}) = \frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{c}^T\boldsymbol{x}_i - y_i)^2 \tag{10}$$

Con un poco de imaginación, podemos ver que la suma de la función de costo es cuadrado de la norma de un vector, un vector cuyas entradas son

$$\Big[\boldsymbol{c}^T\boldsymbol{x}_1 - y_1, \boldsymbol{c}^T\boldsymbol{x}_2 - y_2, \cdots, \boldsymbol{c}^T\boldsymbol{x}_n - y_n\Big]$$

Pues, si inteligentemente definimos la matrix $\boldsymbol{X}$ como

$$\boldsymbol{X} = \Big[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\Big]^T$$

entonces, definiendo $\boldsymbol{y} \coloneqq [y_1, y_2, \cdots, y_n]^T$ tenemos que la función de costo se puede escribir como

$$J(\underset{\rightarrow}{c}) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{c} - \boldsymbol{y}\|^2 = \frac{1}{2n}(\boldsymbol{X}\boldsymbol{c} - \boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{c} - \boldsymbol{y}) \tag{11}$$

Vamos ahora a buscar el mínimo de esta función, encontrando el gradiente $\partial J_{\boldsymbol{c}}$ e igualándolo a cero. Para esto necesitamos la regla de la cadena. Noten porfavor que por la formulación Hessiana, se tiene que para una función $u : \mathbb{R}^n \to \mathbb{R}$, una función $\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^n$ y una variable $\boldsymbol{c} \in \mathbb{R}^m$, la regla de la cadena es:

$$\frac{\partial}{\partial \boldsymbol{c}}u(\boldsymbol{g}(\boldsymbol{c})) = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\frac{\partial u}{\partial \boldsymbol{g}} \tag{12}$$

Ahora basta con aplicar la regla de la cadena para encontrar el gradiente de J

$$\frac{\partial J}{\partial \boldsymbol{c}} = \frac{1}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{c} - \boldsymbol{y}) \tag{13}$$

Al igualar esta expresión a cero se puede resolver para $\boldsymbol{c}$

$$\boldsymbol{c} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{14}$$

Al conjunto de ecuaciones descritas por la ecuación matricial (14) se les conoce como **ecuaciones normales de la regresión lineal**.

2.1. Intepretación geométrica de las ecuaciones normales

Desde haber escrito la función de costo en forma matricial, el problema se convirtio en minimizar el tamaño del vector $\boldsymbol{Xc} - \boldsymbol{y}$, es decir, encontrar un vector ortogonal al espacio generado por las columnas de $\boldsymbol{X}$, cuya longitud está dada por la longitud de $\boldsymbol{y}$ y su ángulo a dicho espacio. Cada vector columna de la matriz bsX es un vector correspondiente a una de las variables que estamos midiendo, y todas sus entradas son las múltiples mediciones que hicimos.

2.2. La matriz de Gram

Como una curiosa observación, pongámosle un poco más de atención a la matriz $\boldsymbol{X}^T\boldsymbol{X}$, la cual es conocida como *matriz de Gamm*. Esta matriz puede dar problemas si resulta no ser invertible, por lo que diremos lo siguiente

Teorema. *La matriz de Gramm es invertible si y sólo si* $\boldsymbol{X}$ *es de rango completo.*

Demostración. ($\leftarrow$) Suponga que $\boldsymbol{X}$ no es de rango completo, entonces existe un vector $\boldsymbol{v} \neq \boldsymbol{0}$ tal que $\boldsymbol{Xv} = \boldsymbol{0}$, con lo que se tiene que $\boldsymbol{X}^T\boldsymbol{Xv} = \boldsymbol{0}$, lo que implica que $\boldsymbol{X}^T\boldsymbol{X}$ tampoco es de rango completo y por lo tanto no es invertible.

($\rightarrow$) Suponga que $\boldsymbol{X}^T\boldsymbol{X}$ no es invertible, siendo ese el caso existe un vector $\boldsymbol{v} \neq \boldsymbol{0}$ tal que $\boldsymbol{X}^T\boldsymbol{Xv} = \boldsymbol{0}$. Se sigue entonces que

$$\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{Xv} = \|\boldsymbol{Xv}\|^2 = 0$$

Esto implica que hay dependencia lineal en las columnas de $\boldsymbol{X}$ y por tanto es de rango deficiente □

El resultado de esto es que las columnas de la matriz $\boldsymbol{X}$ deben ser linealmente independientes. Ojo que las columnas de la matriz no son los vectores de cada medición; hay un vector por cada variable en la que se hacen mediciones. Independencia lineal de estos vectores quiere decir que necesitamos que ni una de las variables que medimos sea la combinación lineal de otras. Aquí viene la importancia de la correlación! Y aunque la correlación entre columnas no sea 1, columnas con alta correlación volveran computacionalmente inestable a la inversión de la matriz de Gram. Cuidado!

Por último, esta matriz resulta ser positiva semidefinida. La prueba la pueden hacer ustedes.

2.3. Una manera de arreglar la inestabilidad

Cuando la inversión de la matriz de Gram es inestable, a veces se recurre a añadir términos de regularización a la función de costo. Es para que se queden *picados*, viene en las siguientes clases. Ah,también nunca verificamos que el problema es *convexo* y de minimización. Eso queda pendiente!