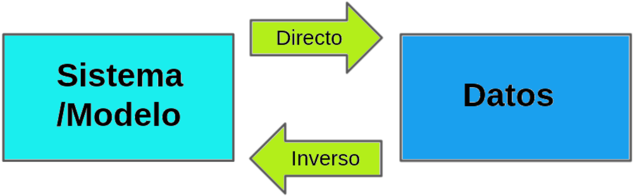
Regresión Lineal: Un Problema de Aprendizaje Estadístico

Mario I. Caicedo

17 de mayo de 2022

Aprendizaje Estadístico

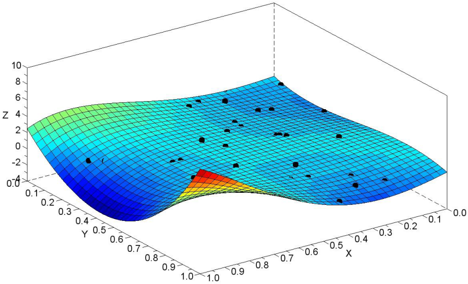
$$\boldsymbol{Y} = \boldsymbol{H}(\boldsymbol{\Theta}|\boldsymbol{X})$$



$$\boldsymbol{Y}^{(obs)} = \boldsymbol{H}(\boldsymbol{\Theta}|\boldsymbol{X})$$

Regresión Lineal

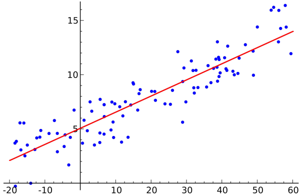
$$\boxed{z_i = z_0 + a_i\, x_i + b_i y_i + c_i x_i^2 + d_i y_i^2 + e_{ij} x_i y_j + [\ldots] + \epsilon_i}$$

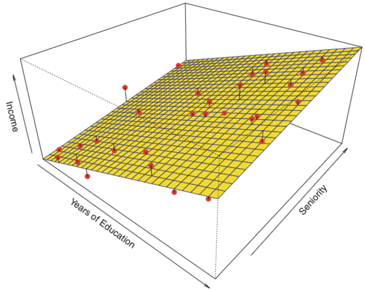


- $\mathbf{y}^{(obs)} = \mathbf{H}(\boldsymbol{\Theta}|\mathbf{X})$
- $\mathbf{y}^{(obs)} = \boldsymbol{\Theta}^T\mathbf{X}$

Casos Más Sencillos

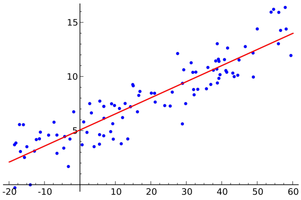
$$\boxed{\begin{array}{l} y_i = b + m\, x_i + \epsilon_i \\ z_i = z_0 + a_i x_i + b_i y_i + \epsilon_i \end{array}}$$





Primer Contacto

La **Regresión Lineal Simple** es el caso más elemental de entre los problemas de regresión lineal.
DATOS: N pares $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$



Los valores y_i tienen ciertos errores de medición (ϵ_i) y en un gráfico cruzado (crossplot) $y\, vs\, x$ encontramos que parece haber una relación de la forma

$$y_i = \theta_1\, x_i + \theta_0 + \epsilon_i$$

Nomenclatura

En la notación estándar de **ML**, la **hipótesis/modelo** se resume en las N fórmulas

$$\boxed{y_i = h_{\mathbf{\Theta}}(x_i) = \theta_0 + \theta_1\, x_i + \varepsilon_i}$$

donde los valores y_i y x_i son conocidos como **valores objetivo** y **predictores** respectivamente y las cantidades θ_0 y θ_1 **pesos**.

Notación

En notación matricial, la hipótesis se escribe como

$$\mathbf{Y} = \mathbf{H}(\mathbf{\Theta}|\mathbf{X}) = \mathbf{X}\,\mathbf{\Theta}$$

Donde:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \quad \text{y} \quad \mathbf{\Theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}$$

Planteamiento Del Problema

- El problema de regresión (ajuste) ó **aprendizaje** consiste en encontrar (calcular/estimar) el vector de pesos $\mathbf{\Theta}$
- Para atacar el problema de **aprendizaje** se introduce un problema de minimización en el espacio de los pesos
- Una vez que se alcanza el **aprendizaje**, es decir, se encontraron los pesos θ_0 y θ_1, estos valores se pueden utilizar junto con la hipótesis para llevar a cabo predecciones/generalizaciones (extrapolaciones)

El Problema de Minimización

- La función de pérdida/costo (error cuadrático medio) se define por:

$$L(\boldsymbol{\Theta}) = \left[\boldsymbol{Y}^{(obs)} - \mathbf{H}(\boldsymbol{\Theta}|\mathbf{X})\right]^T \left[\boldsymbol{Y}^{(obs)} - \mathbf{H}(\boldsymbol{\Theta}|\mathbf{X})\right] =$$
$$= \left[\boldsymbol{Y}^{(obs)} - \mathbf{X}\boldsymbol{\Theta}\right]^T \left[\boldsymbol{Y}^{(obs)} - \mathbf{X}\boldsymbol{\Theta}\right]$$

- La solución al problema de aprendizaje se reduce a minimizar la función de costo con respecto al vector de pesos, esto es, resolviendo la ecuación

$$\nabla_{\boldsymbol{\Theta}^T} L(\boldsymbol{\Theta}) = 0\,.$$

Solución Analítica

- El gradiente de la función de pérdida es

$$\nabla_{\boldsymbol{\Theta}^T} L(\boldsymbol{\Theta}) = -\boldsymbol{X}^T(\boldsymbol{Y}^{(obs)} - \boldsymbol{X}\boldsymbol{\Theta})$$

- Al resolver para los puntos críticos de L, se obtiene de inmediato

$$\boxed{\hat{\boldsymbol{\Theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\,\boldsymbol{X}^T\boldsymbol{Y}^{(obs)}}$$

Solución Analítica
Comentarios

1. La solución analítica requiere el cálculo de la inversa de la matriz $\boldsymbol{X}^T\boldsymbol{X}$, una matriz 2×2, lo que convierte al problema en casi trivial [la existencia de la inversa no está garantizada]
2. El problema de aprendizaje también puede resolverse minimizando la función de costo a través de la técnica de **Descenso por Gradiente**.

Fórmulas Explícitas Para La Solución Analítica

$$\mathbf{\Theta} = \begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum (x_i)^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum y_i x_i \end{pmatrix}$$

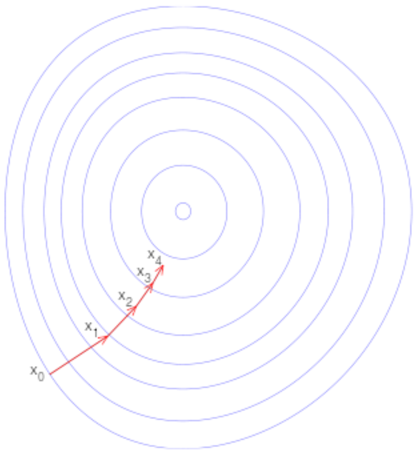
$$\theta_0 = intercepto = \frac{N \sum y_i - m \sum x_i}{N}$$

$$\theta_1 = pendiente = \frac{N \sum x_i y_i - \sum x_i \sum y_j}{N \sum x_i^2 - (\sum x_i)^2}$$

Descenso por Gradiente: Intuición

El método de descenso por gradiente intenta emular el comportamiento de un arroyo de montaña, que siempre fluye “hacia abajo” siguiendo la ruta de mayor pendiente posible.

Descenso por Gradiente I



- El método de descenso por gradiente es un **algoritmo**. de minimización **iterativo**
- El algoritmo utiliza la propiedad geométrica fundamental del gradiente.

Descenso por Gradiente II

1. El algoritmo comienza dando una semilla (valor inicial) $\boldsymbol{\Theta}^{(0)}$
2. A partir de la semilla se itera (el superíndice k indica la iteración) para conseguir nuevos valores de los parámetros

$$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\Theta}^{(k)} - \mu \nabla_{\boldsymbol{\Theta}^T} L(\boldsymbol{\Theta}^{(k)})$$

3. La iteración se detiene con algún criterio. [Un criterio típico consiste en la estabilización de la función de costo].

Pseudocódigo

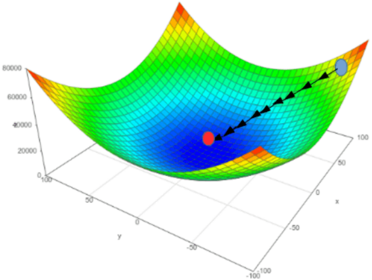
procedure Gradient Descent
Input $\mathbf{\Theta}^{(0)}$: Semilla
$\mathbf{\Theta}^{(k)} = \mathbf{\Theta}^{(0)}$
While $\epsilon \geq criterio$
$L^{(k)} = L(\mathbf{\Theta}^{(k)})$
Paso de Gradiente:

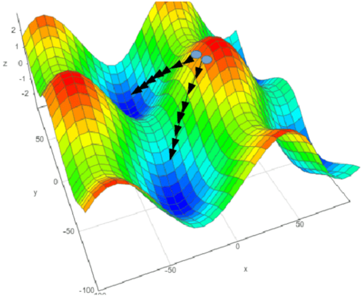
$$\mathbf{\Theta}^{(k+1)} = \mathbf{\Theta}^{(k)} - \mu \nabla_{\mathbf{\Theta}^T} L(\mathbf{\Theta}^{(k)})$$

$L^{(k+1)} = L(\mathbf{\Theta}^{(\mathbf{k+1})})$
$\epsilon = |L^{(k+1)} - L^{(k}|$
Output: return $\mathbf{\Theta}^{(k+1)}$

Descenso POr Gradiente

Hay que tener cuidado con la posibiidad de multimodalidad





Regresión Lineal

Definición

Un problema de Regresión Lineal es un problema de Aprendizaje Automático en que la Hipótesis $\mathbf{Y} = \mathbf{H}(\mathbf{\Theta}|\mathbf{X})$ *tiene la forma matricial*

$$\mathbf{Y} = \mathbf{F}[\mathbf{X}]\mathbf{\Theta}\,,$$

donde $\mathbf{F}[\mathbf{X}]$ *es una matriz que depende de los atributos predictores*

Observación

La solución a un problema de regresión lineal se consigue siguiendo exactamente los mismos pasos utilizados en el caso de la Regresión Lineal Simple, es decir, utilizando el método de mínimos cuadrados

Regresión Lineal: Observaciones I

- La solución analítica del problea ya es conocida:

$$\boxed{\hat{\mathbf{\Theta}} = (\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F}^T \boldsymbol{Y}^{(obs)}}$$

- Desafortunadamente, no hay garantías acerca de la inversibilidad o buen condicionamiento de $\boldsymbol{F}^T \boldsymbol{F}$
- Para $dim(\mathbf{\Theta}) = N$, $dim(\boldsymbol{F}^T \boldsymbol{F}) = N \times N$, en consecuencia, y a pesar de las lindas propiedades de $\boldsymbol{F}^T \boldsymbol{F}$ e incluso si la matriz es inversible y bin consicionada, la inversión puede ser un problema extremadamente pesado de álgebra lineal numérica
- Para las regresiones lineales, la función de pérdida de mínimo cuadrados es convexa lo que garantiza que el método de descenso encontrará un mínimo (se alcanza el aprendizaje)

Descenso por Gradiente
Aplicación A Mínimos Cuadrados

- Función de pérdida (loss). Necesaria para seguir la evolución del algoritmo y detener la corrida ($\mathbf{F} = \mathbf{F}[\mathbf{X}]$).

$$\boxed{L(\boldsymbol{\Theta}) = \left[\boldsymbol{Y}^{(obs)} - \mathbf{F}\boldsymbol{\Theta}\right]^T \left[\boldsymbol{Y}^{(obs)} - \mathbf{F}\boldsymbol{\Theta}\right]} \qquad (1)$$

- Gradiente

$$\boxed{\nabla_{\boldsymbol{\Theta}^T} L(\boldsymbol{\Theta}) = -\boldsymbol{F}^T(\boldsymbol{Y}^{(obs)} - \boldsymbol{F}\boldsymbol{\Theta})} \qquad (2)$$

Regresión Lineal: Observaciones II

- ¿De donde sale la función de costo? [Máxima Verosimilitud].
- Por diversas razones conviene "regularizar" la función de costo

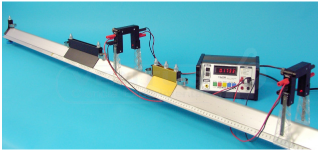
$$L(\mathbf{\Theta}) = \left[\boldsymbol{Y}^{(obs)} - \mathbf{F}\mathbf{\Theta}\right]^T \left[\boldsymbol{Y}^{(obs)} - \mathbf{F}\mathbf{\Theta}\right] + \lambda^2\,\mathbf{\Theta}^T\mathbf{\Theta}$$

- La regularización implica modificaciones, por ejmplo, la solución exacta:

$$\boxed{\hat{\mathbf{\Theta}} = [\boldsymbol{F}^T\boldsymbol{F} + \lambda^2\boldsymbol{I}]^{-1}\boldsymbol{F}^T\mathbf{Y}^{(obs)}}$$

- ¿Como saber que tan "buena" es la regresión? [validación].

Ejemplo I



En un experimento de medición de posición en función del tiempo se encuentra que los datos parecen alinearse ($x = x_0 + v_0 t$)

$$\boldsymbol{X}^{(obs)} = \boldsymbol{F}\boldsymbol{\Theta} \Longleftrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{pmatrix} = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix} \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}$$

Ejemplo II

Para este problema, las fórmulas generales 1 y 2 quedan como sigue

$$\boxed{\nabla_{\mathbf{\Theta}^T} L(\mathbf{\Theta}) = -\mathbf{F}^T(\boldsymbol{X}^{(obs)} - \mathbf{F}\mathbf{\Theta})}$$

$$\nabla_{\mathbf{\Theta}^T} L(\mathbf{\Theta}) = -\begin{pmatrix} 1 & 1 & 1\dots & 1 \\ t_1 & t_2 & t_3\dots & t_N \end{pmatrix} \left[\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{pmatrix} - \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ \vdots & \vdots \\ 1 & t_N \end{pmatrix} \begin{pmatrix} x_0 \\ v_0 \end{pmatrix} \right]$$

Ejemplo III

Y en definitiva, el gradiente tiene la forma

$$\boxed{\nabla_{\boldsymbol{\Theta}^T} L(\boldsymbol{\Theta}^{(k)}) = -\begin{pmatrix} \sum x_i \\ \sum t_i x_i \end{pmatrix} + \begin{pmatrix} N & \sum t_i \\ \sum t_i & \sum (t_i)^2 \end{pmatrix} \begin{pmatrix} x_0^{(k)} \\ v_0^{(k)} \end{pmatrix}}$$

Resumen

1. Hemos explorado un problema (regresión lineal) muy general de aprenmdizaje automático
2. Mostramos el rango de aplicablidad del problema y discutimos sus soluciones analítica e iterativa.

Recursos

- **Lenguajes**
  1. Python
- **Bibliotecas**
  1. Numpy, Python
  2. Sklearn, Python
- **Recurso Cloud**
  1. Colab