Regresión lineal y logística

Víctor Mijangos

Facultad de Ingeniería

Regresión lineal

Regresión lineal

La **regresión lineal** [1] se puede considerar un método de aprendizaje de máquina: Predice valores reales a partir de datos observables.

La idea general es encontrar una **función lineal** que describa el **comportamiento** de los datos.

Dado un conjunto de datos X, descrito por variables $X_1, X_2, \ldots, X_d$, queremos predecir la dependencia de una variable de salida Y.

Buscamos encontrar una función f tal que:

$$f(x) = y$$

Donde x es un vector de entrada, y y es un valor continuo de Y.

Fundamentos de la regresión

El problema de regresión busca determinar una función continua, de la forma:

$$f : X \subseteq \mathbb{R}^d \rightarrow Y \subseteq \mathbb{R}$$

Donde X es un conjunto de datos observados. La variable Y es una variable **continúa**.

Esta función puede determinarse como:

$$f(x_1, ..., x_d) = p(Y = y | x_1, ..., x_d)$$

Donde el vector $x = (x_1, ..., x_d) \in X$ es un dato observado.

Aproximación lineal

Se busca una **función linea** que describa estos datos. Una función lineal es aquella que cumple:

$$f(\lambda x + x') = \lambda f(x) + f(x')$$

Y puede verse que las funciones lineales de $\mathbb{R}^d$ a $\mathbb{R}$ son de la forma:

$$f(x) = wx$$

Tal que $w \in (\mathbb{R}^d)^*$.

En la práctica se utiliza un parámetro de sesgo o **bias**, tal que las funciones son:

$$f(x) = wx + b$$

con $w \in (\mathbb{R}^d)^*, b \in \mathbb{R}$.

Regresión lineal y media

La relación de esta función lineal con la distribución de probabilidad es la **media**:

$$\begin{aligned}\mu &:= \mathbb{E}[Y|x_1, \ldots, x_d] \\ &= w_1 x_1 + \ldots + w_d x_d + b = wx + b\end{aligned}$$

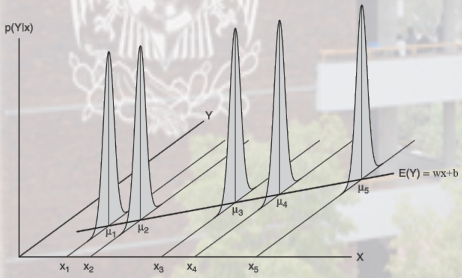


Figura: Suposición: las medias condicionales muestran dependencia linealmente

Ejemplo de regresión lineal

Ejemplo

Supóngase que se tiene una lista de casas, tal que se conoce el número promedio de cuarto (X) y el precio de la casa (Y):

	X	**Y**
Casa 1	*6*	*22.9*
Casa 2	*7.14*	*36.2*
Casa 3	*6.4*	*21.6*
Casa 4	*6.7*	*30.5*
Casa 5	*5.1*	*16.3*
Casa 6	*7.15*	*37.3*
Casa 7	*8.3*	*50*
Caa 8	*8.2*	*48*

Ejemplo de regresión lineal

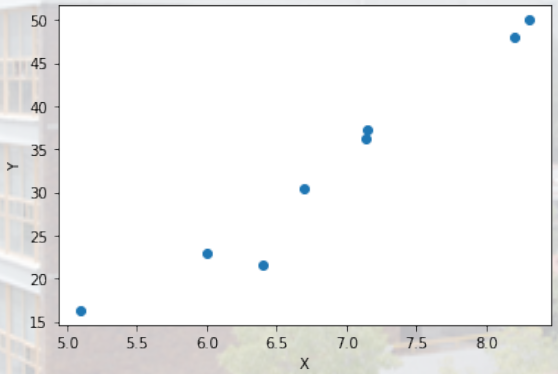


Figura: Visualización de los datos de los precios de casas según el número de cuartos.

Regresión lineal y dependencia lineal

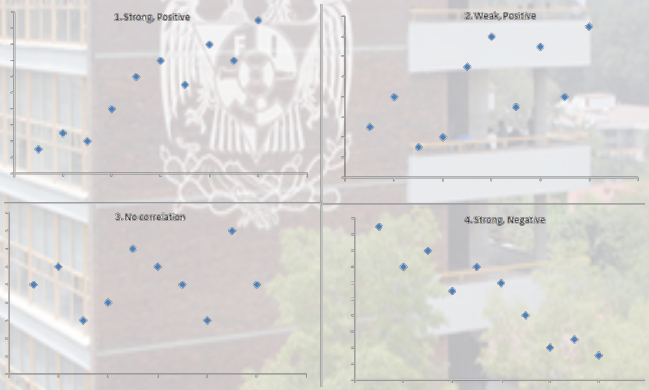
Pueden plantearse la preguntas:

- ¿Existe una correlación (lineal) entre las variables X e Y.
- Si existe ¿cómo se comporta esta correlación?

A partir de la regresión lineal, podemos estimar la correlación que existe entre una y otra variable.

Regresión lineal y correlación lineal

Una correlación estima la dependencia entre dos variables aleatorias. Una alta correlacion asegura que la regresión lineal se acople a los datos adecuadamente.



Error y estimación

Residuos en regresión lineal

El objetivo es entonces predecir el valor de salida y de un valor de entrada x. Podemos estimar que el valor y es [1]:

$$y = wx + b + \epsilon_y$$

Aquí, ϵ_y es un residuo.

Residuos en regresión lineal

Despejando la función anterior, obtenemos que los residuos se calculan como:

$$\epsilon_y = y - wx + b$$

O bien, como:

$$\epsilon_y = y - f(x)$$

Función de riesgo en regresión

Para garantizar que la regresión es adecuada, se busca que la suma de los residuos sea pequeña. Surgen dos aproximaciones:

- **Least-absolute value (LAV)**:

$$R(f) = \sum_{\mathcal{S}} |y - f(x)|$$

- **Least-squares**:

$$R(f) = \sum_{\mathcal{S}} (y - f(x))^2$$

El método de least-squares representa mayor sencillez para resolverse.

Regresión lineal como problema de aprendizaje

Hasta ahora contamos con los siguientes elementos:

1. Un conjunto de datos supervisados $\mathcal{S} = \{(x, y) : x \in \mathbb{R}^d, y \in \mathbb{R}\}$ (se espera que exista una correlación lineal entre X y Y).
2. Una función que define la ML:

$$f(x) = wx + b$$

tal que $w \in \mathbb{R}^d$ y $b \in \mathbb{R}$ son los parámetros a aprender.

3. Una función de riesgo:

$$R(f) = \frac{1}{2} \sum_{\mathcal{S}} (y - f(x))^2$$

(El factor $\frac{1}{2}$ ayuda a simplificar la derivación).

Regresión lineal como problema de aprendizaje

El objetivo es encontrar una función $\hat{f}$ (dependiente de w y b) que minimice la función de riesgo. Asumimos que R es convexa, entonces buscamos su punto de inflexión tal que:

$$\nabla_w R(f) = 0$$

De aquí que:

$$\nabla_w \frac{1}{2} \sum_S (y - f(x))^2 = 0$$

$$\nabla_w \frac{1}{2} ||Y - Xw||^2 = 0$$

$$X^T Y - X^T X w = 0$$

$$(X^T X)^{-1} X^T Y = w$$

Regresión lineal como problema de aprendizaje

Se ha incorporado el bias al vector w ($[w; b]$). Y es el vector de valores esperados y X la matriz cuyos renglones son los ejemplos.
Tenemos, entonces que:

$$\arg\min_{w} \frac{1}{2}||Y - Xw||^2 = (X^T X)^{-1} X^T Y$$

Regresión lineal y distribución normal

Regresión como modelo paramétrico

La regresión lineal es un **modelo paramétrico**, en tanto asume propiedades de la distribución de los datos.

Desde una perspectiva probabilística, la regresión lineal busca estimar $\mathbf{p}(\mathbf{y}|\mathbf{x})$ asumiendo que:

$$y \sim N(\mu, 1)$$

Es decir, asume una distribución normal.

Distribución normal

En este sentido, la función de probabilidad depende de dos parámetros: **media** y **varianza**. Esta última se asume igual a 1, tal que:

$$p(y|x) = \frac{1}{\sqrt{2\pi}} exp[-\frac{1}{2}(y-\mu)^2]$$

El objetivo entonces es estimar la media μ con respecto a los datos X de entrada determinada como:

$$\mu = \mathbb{E}[Y|X] = wx + b$$

Relación con la entropía

Dada la función de probabilidad (que depende de μ, ergo de w) se busca minimizar la entropía:

$$R(w) = -\frac{1}{N}\sum_{y} \ln p(y|x)$$

Es decir, encontrar la mejor función de distribución $p(y|x)$ se reduce a encontrar los valores w que mejor describen la media.

Objetivo: Minimizar la función $R(w)$, que depende de w.

Deducción de la función de riesgo

$$
\begin{aligned}
\arg\min R(w) &= \arg\min -\sum_{y} \ln \frac{1}{\sqrt{2\pi}} exp[-\frac{1}{2}(y-\mu)^2] \\
&= \arg\min -\sum_{y} [\ln \frac{1}{\sqrt{2\pi}}] - \ln exp[\frac{1}{2}(y-\mu)^2] \\
&= \arg\min -\sum_{y} [\ln \frac{1}{\sqrt{2\pi}}] - \frac{1}{2}(y-\mu)^2 \\
&= \arg\min \frac{1}{2} \sum_{\mathcal{S}} (y-\mu)^2 \\
&= \arg\min \frac{1}{2} \sum_{\mathcal{S}} (y-(wx+b))^2
\end{aligned}
$$

Regresión logística

Fundamentos de regresión logística

Muy ligado a la regresión logística, otro modelo de estimación es la **regresión logística** [2].

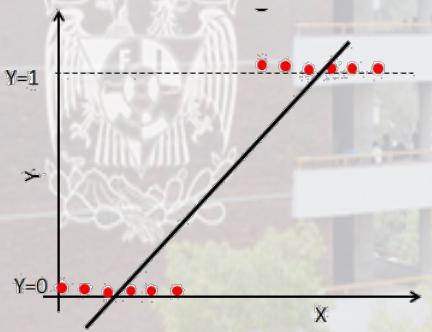
En este caso, se busca estimar una variable Y discreta. Esto es, buscamos una función f:

$$f : X \subseteq \mathbb{R}^d \to \{0, 1\}$$

Se trata de una función de **clasificación**. ¿Se puede aproximar por una función lineal?

Aproximación lineal

Se puede realizar una estimación lineal; sin embargo, no es del todo satisfactoria.

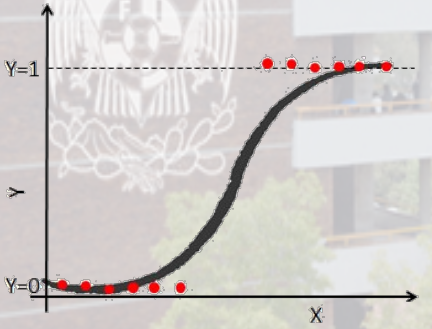


Por tanto, una función lineal no es suficiente para el problema de clasificación.

Aproximación logística

Una mejor aproximación a los datos de clasificación se da al “doblar” la recta.

Una función de este tipo, toca más puntos en las clases observadas.



Calculo de la probabilidad

Se asume que se tienen dos clases, 1 y 0. Si denotamos $p := p(Y = 1|X = x)$, entonces, el logaritmo del cociente de las probabilidades es:

$$\ln \frac{p}{1-p} = b + \sum_i w_i x_i = wx + b$$

De aquí podemos observar que:

$$\frac{p}{1-p} = e^{wx+b}$$

Despejando p obtenemos la probabilidad de que se obtenga el valor 1.

Función logística

Se define la **función logística** como:

$$f(\mu) = \frac{1}{1 + e^{-\mu}} \tag{2}$$

$$= \frac{e^{\mu}}{e^{\mu} + 1} \tag{3}$$

Donde:

$$\mu = wx + b$$

Ya que esta función depende de x, podemos denotar la función logística como $f(x)$.

Regresión logística y probabilidad

Dentro de las propiedades de la función logística, vemos se que cumple que para toda $x \in X$:

$$0 \leq f(x) \leq 1$$

Se puede ver esta función, como una **probabilidad** determinada como:

$$p(Y = 1|x) = \frac{1}{1 + e^{-wx-b}} = f(x)$$

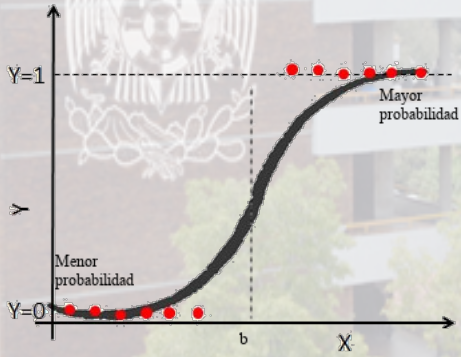
Y por otra parte:

$$p(Y = 0|x) = 1 - p(Y = 1|x) = \frac{1}{1 + e^{wx+b}} = f(-x)$$

Regresión logística y probabilidad

Puede observarse que esta función cumple que:

- $\lim_{x\to\infty} f(x) = 1$
- $\lim_{x\to-\infty} f(x) = 0$



References

John Fox.
Applied regression analysis and generalized linear models.
Sage Publications, 2015.

David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein.
Logistic regression.
Springer, 2002.

The End