Tema 2 : Regresión Lineal

En este tema se provee una introducción a al tema de regresión que consiste en encontrar un expresión matemática que describa el comportamiento de una variable de salida, la cual depende de una o varias variables. Considérese como ejemplo los datos mostrados en la Figura 1. Se tiene un total de diez muestras ($m = 10$) identificadas cada una por un par de valores (x, T). La intención entonces, es encontrar una función $y = f(x)$ tal que pueda con la que sea posible estimar T para cualquier punto de x. Nótese que la distribución de las muestras en el eje de las abscisas (x) es homogénea, sin embargo esto no siempre es así, solo por simplicidad del ejemplo, esta distribución se muestras homogénea.

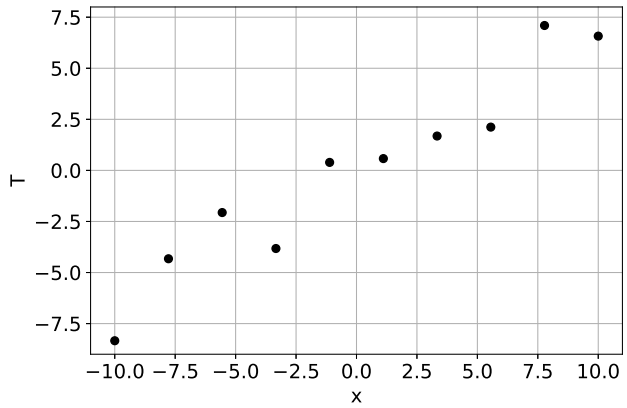


Figure 1: Ejemplo de un conjunto de diez muestras tomado como referencia para un problema de regresión.

El procedimiento para realizar lo anterior mencionado se suelen seguir los siguientes pasos:

1. **Proponer una hipótesis**. Una hipótesis (hablando en un contexto matemático) es la propuesta de una función matemática $y = f(x)$ que se considera aquella función que mejor podría describir el comportamiento de las muestras. La hipótesis mas sencilla que se puede proponer es $y = f(x) = \omega x$, la cual es una ecuación que describe una linea recta que cruza por el origen. El valor de ω es entonces el *coeficiente* de la hipótesis que tendrá que ser encontrado por medio de un método de optimización.

2. **Inicializar el valor de los coeficientes**. Al proponer la hipótesis es importante determinar el valor inicial ve búsqueda de los coeficientes, en este caso, solo se debe definir el valor de ω. Una estrategia común es asignar el valor de cero $\omega_0 = 0$, sin embargo no es una regla general. En posteriores ejercicios de discutirá mas a detalle este proceso de inicialización.

3. **Definir una función de costo** Suponga que el valor inicial del coeficiente es 0.1 $\omega_0 = 0.1$. Esto seria una selección arbitraria que en principio nos permitiría tener una función $y = f(x) = 0.1x$ que no podría predecir el valor de T de la mejor manera. Por ello se debe de *cuantificar* cuan alejado resulta y evaluado en el punto x_i con respecto a la amplitud de cada muestra i ($i \in [1, m]$ y señala la i-esima muestra para todas las variables T, y, etc.). La Figura 2 ilustra el conjunto de muestras junto a la función $y = 0.1x$, para cada muestra se debe evaluar un costo c_i que podría ser simplemente la diferencia de amplitud. No obstante, la expresión para evaluar el costo $c_i = y_i - T_i$ representa una función que no tiene un mínimo global (un punto cuya derivada se cero).

   Pensando en un algoritmo numérico de optimización el cual permite encontrar el valor de un coeficiente donde se encuentre un mínimo local, la expresión mencionada para evaluar el costo por muestra no es aplicable. Por ello, se debe utilizar una función que sea optimizable, es decir que exista dentro de su dominio al menos un mínimo. El error medio cuadrático es una opción común para evaluar el costo:

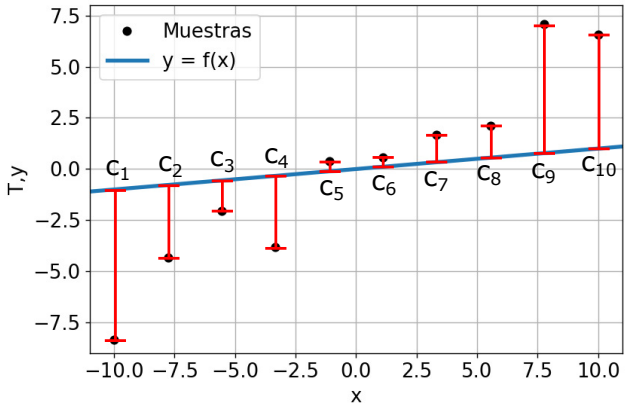


Figure 2

$$c_i(\omega) = (y_i(x_i) - T_i)^2 = (\omega x_i - T_i)^2 \tag{1}$$

Ahora considere que se están evaluando m costos ya que existen m muestras. En un problema de optimización solo debe existir una función a optimizar. Esta función puede ser aquella que represente el promedio de los m costos, tal y como se expresa a continuación:

$$C(\omega) = \frac{1}{m}\sum_{i=1}^{m} c_i(\omega) = \frac{1}{m}\sum_{i=1}^{m}(y_i(x_i) - T_i)^2 = \frac{1}{m}\sum_{i=1}^{m}(\omega x_i - T_i)^2 \tag{2}$$

La Ecuación 2 entonces sera nuestra función de costo a optimizar

4. **Obtener la derivada de la función de costo**. El procedimiento consiste en analíticamente encontrar la expresión matemática de la función de costo (Ecuación 2). Tenemos entonces:

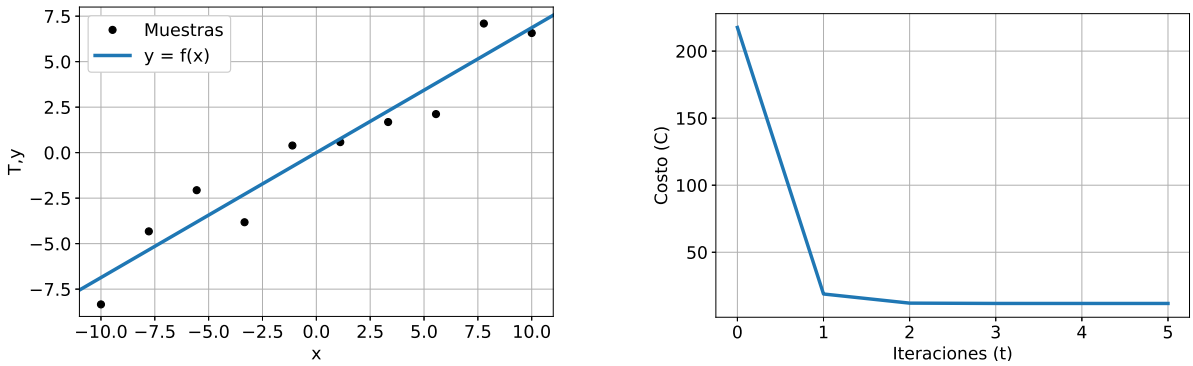
   Derivando con respecto a ω

$$\begin{aligned}
\frac{\partial C(\omega)}{\partial \omega} &= \frac{1}{m}\frac{\partial \sum_{i=1}^{m}(y_i(x_i) - T_i)^2}{\partial \omega} \\
&= \frac{1}{m}\frac{\partial \sum_{i=1}^{m}((\omega x_i) - T_i)^2}{\partial \omega} \\
&= \frac{1}{m}\sum_{i=1}^{m} 2((\omega x_i) - T_i)\frac{\partial((\omega x_i) - T_i)}{\partial \omega} \\
&= \frac{1}{m}\sum_{i=1}^{m} 2((\omega x_i) - T_i)(x_i) \\
&= \boxed{\frac{2}{m}\sum_{i=1}^{m}(y_i(x_i) - T_i)(x_i)}
\end{aligned} \tag{3}$$

5. **Implementar el algoritmo de optimización** El algoritmo de optimización puede ser Gradiente Descendente (GD) la cual sigue el siguiente regla de actualización:

$$\omega_{t+1} = \omega_t - \alpha\frac{\partial C}{\partial \omega_t} \tag{4}$$

El procedimiento de actualización (GD) permite actualizar sucesivamente el coeficiente ω. con este método el valor obtenido para ω es 0.71 aproximadamente empleando un valor $\alpha = 0.001$. la Figura 3 muestra los resultados del este paso.



(a) (b)

Figure 3: a) Resultado del algoritmo de optimización ejecutado cinco iteraciones ($y = 0.71x$). b) Valor del costo calculado con la Ecuación 2 en cada iteración

6. **Comprobación del resultado (Paso opcional)**. Una posible opción para verificar que efectivamente se encontró el valor correcto de ω donde se encuentra un mínimo, es graficar la expresión del costo (Ecuación 2) para un rango arbitrario de ω. Este rango es una selección con criterio de donde se estima que se encuentra el mínimo. Generalmente hablando, no hay manera de conocer a priori la ubicación aproximada de un mínimo. Téngase en cuneta que al trabajar con una expresión de costo que es una función dependiente de una sola variable es posible generar una grafica de dos dimensiones, como es el caso de este ejemplo (Figura 4). Para el caso de una expresión de costo que sea una función dependiente de dos variables se debería generar entonces una grafica de tres dimensiones. Para aquellas funciones de costo dependientes de mas de dos variables resultará imposible generar una grafica y por consecuencia este paso no se podrá realizar.

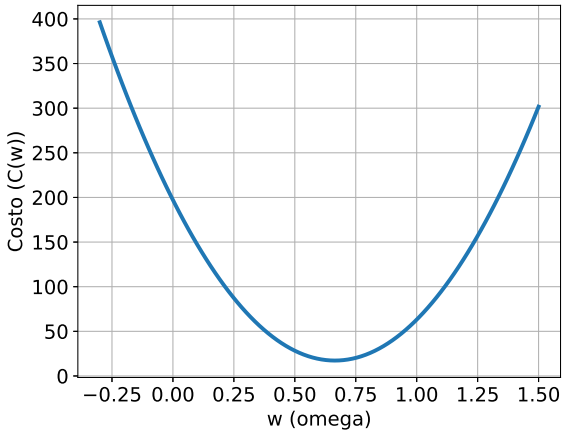


Figure 4: Valor del costo (Ecuación 2) para diferentes valores de ω.

Ejercicios

1. Realizar el ajuste de una hipótesis propuesta, por ejemplo $y = f(x) = \omega x$ a un conjunto de datos (conjunto de entrenamiento) generado artificialmente (La estrategia para generar estas muestras es libre).

   Los puntos para elaborar son:

   - Generar el conjunto de entrenamiento (X, T) ($X = x_1, x_2, \ldots, x_m$, $T = T_1, T_2, \ldots, T_m$), tal que $x_i \in [-10, 10]$ y se cuente con al menos 100 muestras ($m \geq 100$). Se debe procurar que la tendencia que siguen mas muestra siga una linea recta y crucen cerca al origen.
   - Proponer una hipótesis $y = f(x)$
   - Seleccionar una función de costo ($C(\omega)$) y obtener su derivada. Se deberá explorar al menos tres funciones de costo:
     (a) $C(\omega) = \frac{1}{m}\sum_{i=1}^{m}(y_i(x_i) - T_i)^2$ (MSE)
     (b) $C(\omega) = \frac{1}{m}\sum_{i=1}^{m}|(y_i - T_i)|$ (MAE)
     (c) $C(\omega) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - T_i)^2}$ (RMSE)
   - Programar algoritmo de optimización basado en GD generando como resultado tres graficas
     (a) T vs X y sobre la misma grafica $f(x)$ vs X
     (b) $C(\omega)$ vs Iteraciones
     (c) $C(\omega)$ vs ω

   Se deberá discutir acerca de las diferencias (ventajas o desventajas) de emplear cada una de las funciones de costo.

2. Repetir el ejercicio anterior pero esta vez empleando como hipótesis $y = \omega_0 + \omega_1 x$