Tema 3 : Regresión polinomial (Univariable)

En este tema se brindara una descripción del uso de polinomios de grado g ($g \geq 2$) para describir comportamiento de observaciones dependientes de una sola variable x. Como referencia se tiene el conjunto de muestras A ($A = (x_0, T_0), (x_1, T_1), \ldots, (x_{m-1}, T_{m-1})$) ilustrado en la Figura 1.

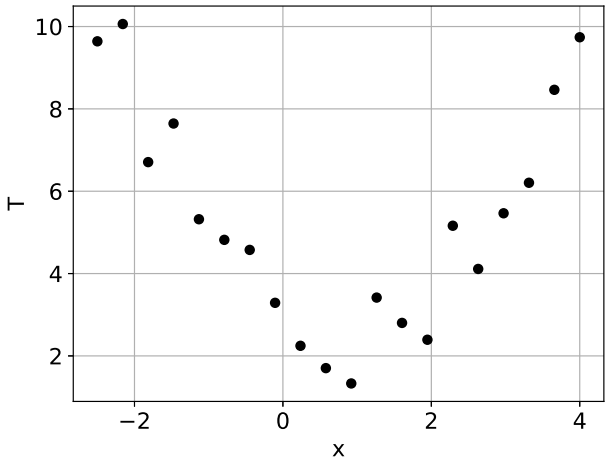


Figure 1: Ejemplo de un conjunto de veinte muestras tomado como referencia para un problema de regresión polinomial.

Evidentemente, una función (hipótesis) de grado 1, por ejemplo $y = f(x) = \omega_0 + \omega_1 x$ no representara correctamente la tendencia seguida por las muestras. Por esta razón, es necesario proponer una hipótesis de mayor complejidad. De manera general, la idea es agregar términos adicionales elevando cada nuevo termino a una potencia superior, teniendo así el polinomio de grado g:

$$y(x) = \omega_0 x^0 + \omega_1 x^1 + \omega_2 x^2 + \ldots + \omega_g x^g \tag{1}$$

En el que existen $g + 1$ coeficientes (ω) de los cuales deberá ser encontrado su valor por medio de un algoritmo de optimización, por ejemplo el basado en gradiente descendente.

Realizando una recapitulación, son necesarios cinco pasos básicos para obtener la ecuación final que describir el comportamiento del conjunto A:

1. **Proponer una hipótesis:** Considérese como propuesta de hipótesis para el conjunto A un polinomio de grado 2:

   $$y_i(x_i) = \omega_0 + \omega_1 x_i + \omega_2 x_i^2 \tag{2}$$

   Las operaciones por esta hipótesis pueden también expresarse como el producto punto de dos vectores, que en su forma general queda como:

   $$\mathbf{x}_i^\intercal \cdot \mathbf{w} = \begin{bmatrix} x_i^0 & x_i^1 & x_i^2 & \ldots & x_i^g \end{bmatrix} \cdot \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \vdots \\ \omega_g \end{bmatrix} = \begin{bmatrix} y_i \end{bmatrix} \tag{3}$$

   Donde las letras minúsculas en negritas $\mathbf{x}$ y $\mathbf{w}$ representan vectores columna que contienen respectivamente las potencias de x para la muestra i y los coeficientes ω_j ($\forall j \in [0, g]$). Asimismo, es posible

representar el producto matricial considerando las m muestras, para ello, es necesario formar un nuevo arreglo $\mathbf{Q}$ que alojará en cada columna las m muestras. Estas muestras estarán elevadas a un potencia igual a la columna en la que están alojadas, siendo que la primera columna (al igual que la primera fila) corresponde a la columna cero.

$$\mathbf{Q} \cdot \mathbf{w} = \begin{bmatrix} x_0^0 & x_0^1 & x_0^2 & \dots & x_0^g \\ x_1^0 & x_1^1 & x_1^2 & \dots & x_1^g \\ x_2^0 & x_2^1 & x_2^2 & \dots & x_2^g \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m-1}^0 & x_{m-1}^1 & x_{m-1}^2 & \dots & x_{m-1}^g \end{bmatrix} \cdot \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \vdots \\ \omega_g \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_{m-1} \end{bmatrix} \tag{4}$$

2. **Inicializar el valor de los coeficientes:** Típicamente, inicializar los valores de $\omega_i = 0$ $(i = [0, 1, 2])$ es una selección adecuada

3. **Definir una función de costo:** Para este ejemplo se utilizara como función de costo el Error Medio Cuadrático (MSE):

$$C(\mathbf{w}) = \frac{1}{2m} \sum_{i=1}^{m} (y_i(x_i) - T_i)^2 \tag{5}$$

   Las razones para elegir esta expresión son las siguientes:

   (a) **Forma convexa del hiperplano:** La forma convexa del hiperplano de $C(W)$ se caracteriza por tener derivadas de magnitud cada vez menor a medida que se está mas cerca de un mínimo local, esto implica que la aproximación a este mínimo en un algoritmo de búsqueda sea suave y se logre la convergencia.

   (b) **Simplicidad computacional:** Considerando el proceso de optimización, es conveniente emplear expresiones matemáticas con el menor número de operaciones implicadas, si se sigue este criterio es posible obtener una reducción en tiempo de procesamiento en comparación al obtenido en implementaciones que usan funciones matemáticamente mas complejas. En el caso del MSE, tanto su misma expresión como la de su derivada son las expresiones mas simples.

4. **Obtener la derivada de la función de costo:** Es necesario obtener la derivada de la función de costo (Ecuación 5) con respecto a cada uno de los coeficientes. Ya que se toma como ejemplo un polinomio de grado dos, son tres las derivadas que de deben de obtener, esto es:

   Derivando con respecto a ω_0

$$\begin{aligned} \frac{\partial C(\mathbf{w})}{\partial \omega_0} &= \frac{1}{2m} \frac{\partial \sum_{i=1}^{m} (y_i(x_i) - T_i)^2}{\partial \omega_0} \\ &= \frac{1}{2m} \frac{\partial \sum_{i=1}^{m} ((\omega_0 + \omega_1 x_i + \omega_2 x_i^2) - T_i)^2}{\partial \omega_0} \\ &= \frac{1}{2m} \sum_{i=1}^{m} 2((\omega_0 + \omega_1 x_i + \omega_2 x_i^2) - T_i) \frac{\partial ((\omega_0 + \omega_1 x_i + \omega_2 x_i^2) - T_i)}{\partial \omega_0} \\ &= \frac{1}{m} \sum_{i=1}^{m} ((\omega_0 + \omega_1 x_i + \omega_2 x_i^2) - T_i)(1) \\ &= \boxed{\frac{1}{m} \sum_{i=1}^{m} (y_i(x_i) - T_i) = \frac{1}{m} \sum_{i=1}^{m} (y_i(x_i) - T_i) x_i^0} \end{aligned} \tag{6}$$

Siguiendo este procedimiento se tiene que:

$$\boxed{\frac{\partial C(\mathbf{w})}{\partial \omega_1} = \frac{1}{m}\sum_{i=1}^{m}(y_i(x_i) - T_i)x_i^1} \tag{7}$$

y

$$\boxed{\frac{\partial C(\mathbf{w})}{\partial \omega_2} = \frac{1}{m}\sum_{i=1}^{m}(y_i(x_i) - T_i)x_i^2} \tag{8}$$

Naturalmente, si se propone un hipótesis de grado mayor se tendrán que obtener mas expresiones para las derivadas, sin embargo es posible obtener una expresión general para el calculo de la derivada de la función de costo con respecto al coeficiente ω_j:

$$\boxed{\frac{\partial C(\mathbf{w})}{\partial \omega_j} = \frac{1}{m}\sum_{i=1}^{m}(y_i(x_i) - T_i)x_i^j} \tag{9}$$

Nótese como a la Ecuación 5 se le agrego un 2 que multiplica a m, éste simplifica en una operación a la ecuación de la derivada, lo cual es una ligera contribución a la reducción de tiempo de ejecución del algoritmo de optimización puesto que es la expresión de la derivada la que se emplea un mayor número de veces durante la búsqueda de un mínimo.

5. **Implementar el algoritmo de optimización:** La regla de actualización de los coeficientes del algoritmo de Gradiente Descendente (GD) debe ser aplicada a cada uno de los $g + 1$ coeficientes en cada iteración, esto es:

$$\omega_{j,t+1} = \omega_{j,t} - \alpha\frac{\partial C}{\partial \omega_{j,t}} \tag{10}$$

donde j es el indice del coeficiente y t indica la iteración actual del algoritmo de búsqueda .

En la Figura 2 se muestra el resultado del algoritmo de búsqueda GD después de 100 iteraciones con un factor de aprendizaje $\alpha = 0.01$.

Ejercicios

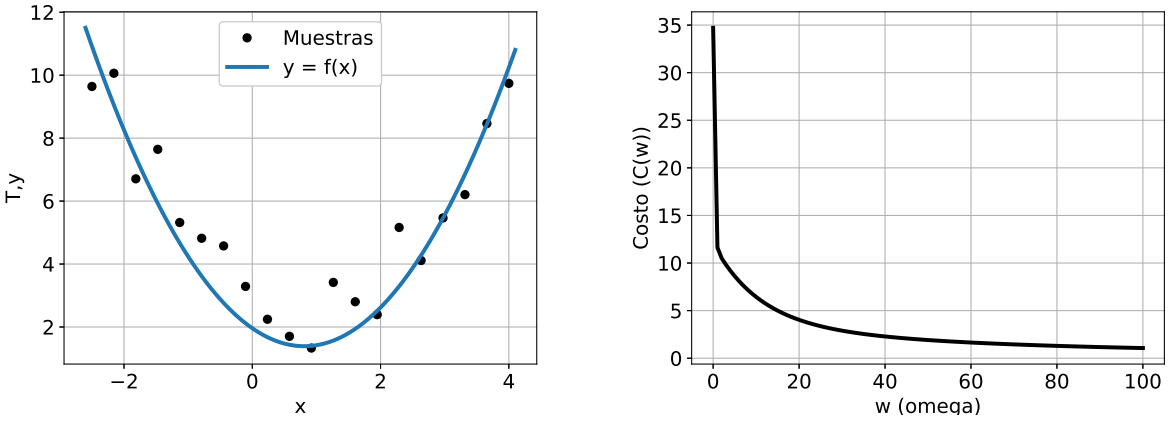
1. Generar un conjunto de muestras similar al conjunto A mostrado en la Figura 1. Posteriormente, se deberá seguir el procedimiento descrito en este documento para obtener la ecuación $f(x)$ que describa el comportamiento de las muestras generadas.

   Queda a criterio del estudiante explorar como hipótesis, polinomios de un grado mayor a dos. Sin embargo, se debe dar prioridad a entender y formular una explicación del por que los resultados mostrados en la Figura 2 (o los obtenidos por el estudiante) están alejados de lo que debería haber arrojado el algoritmo de búsqueda.

2. A partir del conjunto de muestras descrita por los siguientes vectores:

   $x = [1, 2, 3, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 18, 19, 21, 22]$
   $T = [100, 90, 80, 60, 60, 55, 60, 65, 70, 70, 75, 76, 78, 79, 90, 99, 99, 100]$

   Elabore un algoritmo de regresión polinomial que ajuste una hipótesis de grado tres a la tendencia seguida por este conjunto. Los pasos a seguir son los mismos del inciso anterior (punto 1).



(a) (b)

Figure 2: a) Resultado del algoritmo de optimización. b) Costo obtenido en cada iteración del GD

3. Elaborar un algoritmo de regresión lineal que ajuste una hipótesis propuesta (por ejemplo $p = f(x) = \omega x$ o $p = \omega_0 + \omega_1 x$) a los puntos generados por la siguiente función :

$$y = -0.3x + 20 + 3R; \tag{11}$$

donde R es un número generado de manera aleatoria cuya valor probable respeta un distribución uniforme en el rango $[-1, 1]$.

Se deberán tener en consideración los siguientes puntos

- Proponer una hipótesis $p = f(x)$
- Seleccionar una función de costo ($C(\mathbf{w})$). Se sugiere el uso del error cuadrático medio como función de costo:

$$C(\mathbf{w}) = \frac{1}{2m}\sum_{i=1}^{m}(P_i(x_i) - T_i)^2 \qquad \text{(MSE)} \tag{12}$$

- Las derivadas de la función de costo respecto a los parámetros ($\frac{\partial C(\mathbf{w})}{\partial \omega_j}$) se deberá obtener como una aproximación, es decir, el valor de la derivada se obtendrá como el calculo de las pendientes d_j de la función $C(\mathbf{w})$ en el punto ω_j ($j \in [0, g-1]$ ya que es una pendiente d por cada parámetro), siendo $d_j = \frac{yb_j - ya_j}{\omega b_j - \omega a_j}$, $\omega b_j = \omega_j + \epsilon$, $\omega a_j = \omega_j - \epsilon$, $yb_j = C(\omega_j + \epsilon)$ y $ya_j = C(\omega_j - \epsilon)$
- Elaborar el algoritmo de entrenamiento basado en GD generando como resultado tres graficas: 1) T vs x y sobre la misma grafica $P(x)$ vs x 2) $C(\mathbf{w})$ vs Iteraciones y 3) $C(\mathbf{w})$ vs parámetros ω_0 y ω_1 (ω_1 existe solo si la hipótesis es $p = \omega_0 + \omega_1 x$)