Universidad del Altiplano

FACULTAD ESTADISTICA E INFORMATICA

Descenso del Gradiente

Alumno: Melissa Jessica Macedo Ramos
Materia: METODOS DE OPTIMIZACION
Fecha: January 27, 2025

Ejercicio 1

Minimizar la función:

$$g(x) = (x-5)^2$$

Iteraciones

La derivada es:

$$\frac{dg}{dx} = 2(x-5).$$

Partimos de $x_0 = 10$ con tasa de aprendizaje $\eta = 0.2$. Calculamos 5 iteraciones.

- **Iteración 1:**

$$\begin{aligned}
\text{Gradiente: } & \frac{dg}{dx}\Big|_{x=10} = 2(10-5) = 10. \\
\text{Actualización: } & x_1 = x_0 - 0.2 \cdot 10 = 10 - 2 = 8. \\
\text{Valor de } g(x): & \; g(10) = (10-5)^2 = 25.
\end{aligned}$$

- **Iteración 2:**

$$\begin{aligned}
\text{Gradiente: } & \frac{dg}{dx}\Big|_{x=8} = 2(8-5) = 6. \\
\text{Actualización: } & x_2 = x_1 - 0.2 \cdot 6 = 8 - 1.2 = 6.8. \\
\text{Valor de } g(x): & \; g(8) = (8-5)^2 = 9.
\end{aligned}$$

- **Iteración 3:**

$$\begin{aligned}
\text{Gradiente: } & \frac{dg}{dx}\Big|_{x=6.8} = 2(6.8-5) = 3.6. \\
\text{Actualización: } & x_3 = x_2 - 0.2 \cdot 3.6 = 6.8 - 0.72 = 6.08. \\
\text{Valor de } g(x): & \; g(6.8) = (6.8-5)^2 = 3.24.
\end{aligned}$$

- **Iteración 4:**

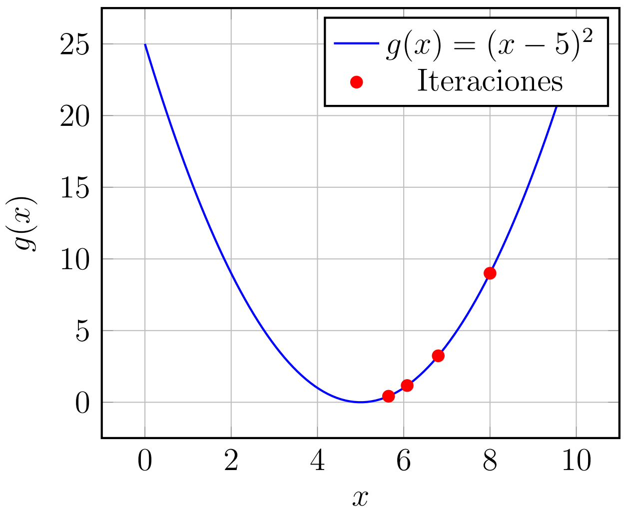
$$\begin{aligned}
\text{Gradiente: } & \frac{dg}{dx}\Big|_{x=6.08} = 2(6.08-5) = 2.16. \\
\text{Actualización: } & x_4 = x_3 - 0.2 \cdot 2.16 = 6.08 - 0.432 = 5.648. \\
\text{Valor de } g(x): & \; g(6.08) = (6.08-5)^2 = 1.1664.
\end{aligned}$$

- **Iteración 5:**

$$\begin{aligned}
\text{Gradiente:}\quad & \frac{dg}{dx}\Big|_{x=5.648} = 2(5.648-5) = 1.296. \\
\text{Actualización:}\quad & x_5 = x_4 - 0.2\cdot 1.296 = 5.648 - 0.2592 = 5.3888. \\
\text{Valor de } g(x):\ & g(5.648) = (5.648-5)^2 = 0.419904.
\end{aligned}$$

El valor de x converge hacia 5, el mínimo de la función.

Gráfico del descenso del gradiente



Ejercicio 2

Dado el conjunto de datos (x_i, y_i):

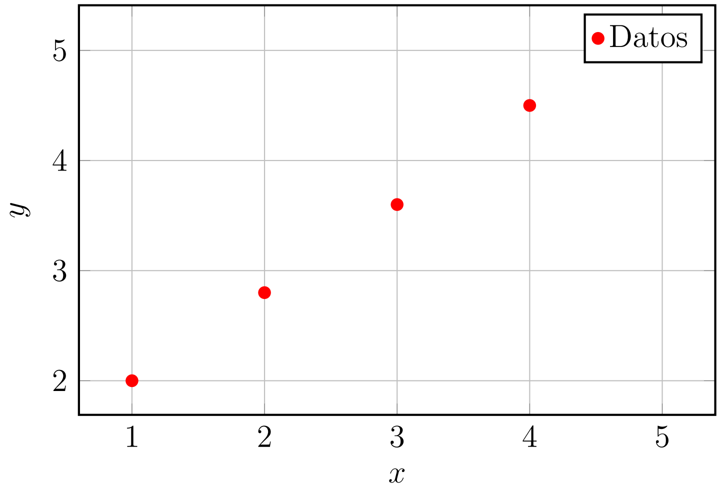
$$(x_i, y_i) \in \{(1,2), (2,2.8), (3,3.6), (4,4.5), (5,5.1)\}$$

Queremos ajustar la recta $h(x) = \beta_0 + \beta_1 x$ minimizando:

$$J(\beta_0, \beta_1) = \sum_{i=1}^{5} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2.$$

Usamos descenso del gradiente con $\eta = 0.01$. Partimos de $\beta_0 = 0$, $\beta_1 = 0$ y realizamos 3 iteraciones.

Visualización del conjunto de datos



Iteraciones

- **Iteración 1:**

$$\begin{aligned}
\text{Gradientes:}\quad & \frac{\partial J}{\partial \beta_0} = -\frac{2}{5}\sum_{i=1}^{5}\big(y_i - (\beta_0 + \beta_1 x_i)\big),\\
& \frac{\partial J}{\partial \beta_1} = -\frac{2}{5}\sum_{i=1}^{5} x_i\big(y_i - (\beta_0 + \beta_1 x_i)\big).\\
\text{Cálculos:}\quad & \frac{\partial J}{\partial \beta_0}\Big|_{\beta_0=0,\beta_1=0} = -7.2,\\
& \frac{\partial J}{\partial \beta_1}\Big|_{\beta_0=0,\beta_1=0} = -30.8.\\
\text{Actualización:}\quad & \beta_0 = 0 - 0.01(-7.2) = 0.072,\\
& \beta_1 = 0 - 0.01(-30.8) = 0.308.
\end{aligned}$$

- **Iteración 2:**

$$\begin{aligned}
\text{Gradientes:}\quad & \frac{\partial J}{\partial \beta_0} = -\frac{2}{5}\sum_{i=1}^{5}\big(y_i - (0.072 + 0.308 x_i)\big),\\
& \frac{\partial J}{\partial \beta_1} = -\frac{2}{5}\sum_{i=1}^{5} x_i\big(y_i - (0.072 + 0.308 x_i)\big).\\
\text{Cálculos:}\quad & \frac{\partial J}{\partial \beta_0}\Big|_{\beta_0=0.072,\beta_1=0.308} \approx -2.952,\\
& \frac{\partial J}{\partial \beta_1}\Big|_{\beta_0=0.072,\beta_1=0.308} \approx -12.664.\\
\text{Actualización:}\quad & \beta_0 = 0.072 - 0.01(-2.952) = 0.10152,\\
& \beta_1 = 0.308 - 0.01(-12.664) = 0.43464.
\end{aligned}$$

- **Iteración 3:**

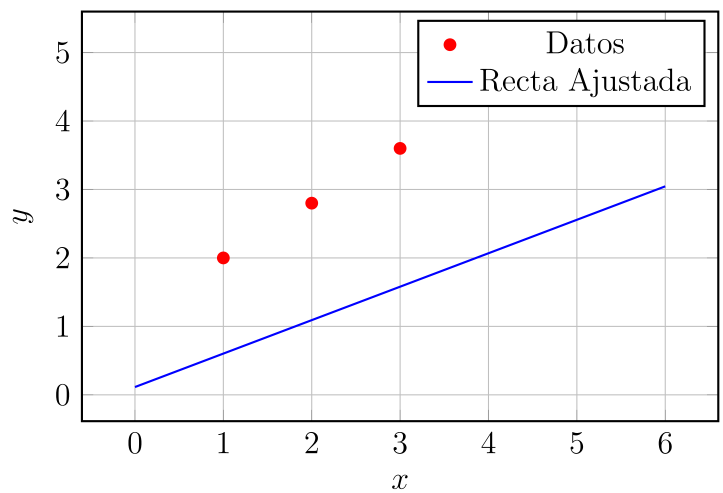
$$
\begin{aligned}
\text{Gradientes:} \quad & \frac{\partial J}{\partial \beta_0} = -\frac{2}{5}\sum_{i=1}^{5}\big(y_i - (0.10152 + 0.43464x_i)\big), \\
& \frac{\partial J}{\partial \beta_1} = -\frac{2}{5}\sum_{i=1}^{5} x_i\big(y_i - (0.10152 + 0.43464x_i)\big). \\
\text{Cálculos:} \quad & \frac{\partial J}{\partial \beta_0}\Big|_{\beta_0=0.10152,\beta_1=0.43464} \approx -1.253, \\
& \frac{\partial J}{\partial \beta_1}\Big|_{\beta_0=0.10152,\beta_1=0.43464} \approx -5.396. \\
\text{Actualización:} \quad & \beta_0 = 0.10152 - 0.01(-1.253) = 0.11405, \\
& \beta_1 = 0.43464 - 0.01(-5.396) = 0.4886.
\end{aligned}
$$

Recta Ajustada

Tras 3 iteraciones, obtenemos una aproximación para la recta ajustada:

$$h(x) \approx 0.11405 + 0.4886x.$$

Visualizamos los datos originales y la recta ajustada:



EJERCICIO 3

Dado el conjunto de datos:

Muestra	x_1	x_2	y
1	0.5	1.0	0
2	1.5	2.0	0
3	2.0	2.5	1
4	3.0	3.5	1

Queremos entrenar un modelo de clasificación logística $\sigma(\mathbf{w}^T\mathbf{x})$ mediante el descenso del gradiente. Inicializamos los pesos $\mathbf{w}_0 = (0, 0, 0)$ e incluimos el sesgo como componente adicional.

La función sigmoide es:

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

La función de costo logístico es:

$$J(\mathbf{w}) = -\frac{1}{m}\sum_{i=1}^{m}\Big[y^{(i)}\log(\sigma(\mathbf{w}^T\mathbf{x}^{(i)})) + (1 - y^{(i)})\log(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}))\Big].$$

Iteración 0: Inicialización

Pesos iniciales:

$$\mathbf{w}_0 = (0, 0, 0).$$

Tasa de aprendizaje: $\eta = 0.1$.

Iteraciones del Descenso del Gradiente

Iteración 1

- **Predicciones:**

$$z^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} = 0, \quad \sigma(z^{(i)}) = 0.5 \quad \forall i.$$

- **Gradientes:**

$$\frac{\partial J}{\partial \mathbf{w}} = \frac{1}{m}\sum_{i=1}^{m}\big(\sigma(z^{(i)}) - y^{(i)}\big)\mathbf{x}^{(i)} = [0.0, -0.375, -0.375].$$

- **Actualización:**

$$\mathbf{w}_1 = \mathbf{w}_0 - \eta\frac{\partial J}{\partial \mathbf{w}} = (0.0, 0.0375, 0.0375).$$

- **Costo:**

$$J(\mathbf{w}_1) = 0.6684.$$

Iteración 2

- **Predicciones:**

$$z^{(i)} = \mathbf{w}_1^T\mathbf{x}^{(i)} \quad \text{(calculado)}.$$

- **Gradientes:**

$$\frac{\partial J}{\partial \mathbf{w}} = [-0.0374, -0.2944, -0.2757].$$

- **Actualización:**

$$\mathbf{w}_2 = \mathbf{w}_1 - \eta \frac{\partial J}{\partial \mathbf{w}} = (-0.0037, 0.0669, 0.0651).$$

- **Costo:**

$$J(\mathbf{w}_2) = 0.6538.$$

Iteración 3

- **Predicciones:**

$$z^{(i)} = \mathbf{w}_2^T \mathbf{x}^{(i)} \quad \text{(calculado).}$$

- **Gradientes:**

$$\frac{\partial J}{\partial \mathbf{w}} = [-0.0644, -0.2360, -0.2038].$$

- **Actualización:**

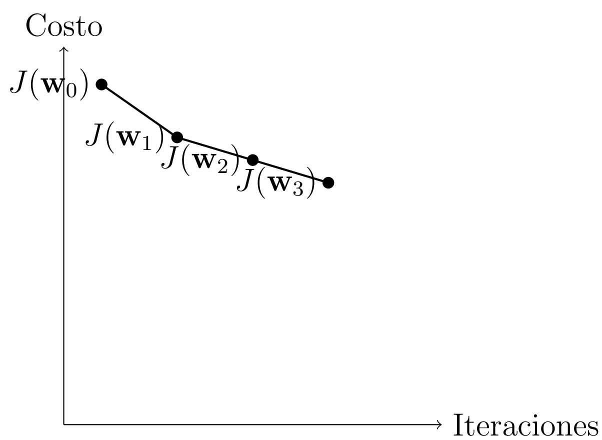
$$\mathbf{w}_3 = \mathbf{w}_2 - \eta \frac{\partial J}{\partial \mathbf{w}} = (-0.0102, 0.0905, 0.0855).$$

- **Costo:**

$$J(\mathbf{w}_3) = 0.6401.$$

Visualización

A continuación, mostramos cómo evolucionan los pesos y cómo disminuye el costo logístico en cada iteración.



Ejercicio 4

Dado un conjunto de 1000 observaciones $\{(\mathbf{x}_i, y_i)\}_{i=1}^{1000}$ para un problema de regresión multivariable, aplicamos el método de Descenso Estocástico del Gradiente (SGD), dividiendo el conjunto en minibatches de tamaño $B = 50$. La función de costo utilizada es:

$$J(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2,$$

donde $\mathbf{w}$ representa el vector de parámetros y N es el número total de observaciones. Se utiliza una tasa de aprendizaje $\eta = 0.01$. A continuación, se describe el algoritmo y se analiza cómo los parámetros $\mathbf{w}$ se actualizan en cada iteración.

Algoritmo: Descenso Estocástico del Gradiente (SGD)

El procedimiento general del algoritmo es:

1. Dividir el conjunto de datos en minibatches de tamaño $B = 50$.
2. Seleccionar aleatoriamente un minibatch $\{(\mathbf{x}_i, y_i)\}_{i=1}^{B}$.
3. Calcular el gradiente del costo asociado al minibatch:
$$\nabla J(\mathbf{w}) = -\frac{1}{B}\sum_{i=1}^{B}\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)\mathbf{x}_i.$$
4. Actualizar los parámetros $\mathbf{w}$:
$$\mathbf{w} \leftarrow \mathbf{w} - \eta\nabla J(\mathbf{w}).$$
5. Repetir los pasos 2-4 para todos los minibatches, iterando hasta alcanzar la convergencia.

Iteraciones del Algoritmo SGD

Iteración 1

- **Minibatch:** Seleccionamos aleatoriamente un subconjunto de 50 observaciones.
- **Gradiente:** Calculamos el gradiente promedio para este minibatch:
$$\nabla J(\mathbf{w}_0) = -\frac{1}{50}\sum_{i=1}^{50}\left(y_i - \mathbf{w}_0^T\mathbf{x}_i\right)\mathbf{x}_i.$$
- **Actualización:** Ajustamos los parámetros:
$$\mathbf{w}_1 = \mathbf{w}_0 - \eta\nabla J(\mathbf{w}_0).$$

Iteración 2

- **Minibatch:** Seleccionamos un nuevo subconjunto de 50 observaciones.
- **Gradiente:** Calculamos el gradiente:
$$\nabla J(\mathbf{w}_1) = -\frac{1}{50}\sum_{i=1}^{50}\left(y_i - \mathbf{w}_1^T\mathbf{x}_i\right)\mathbf{x}_i.$$
- **Actualización:** Actualizamos los parámetros:
$$\mathbf{w}_2 = \mathbf{w}_1 - \eta\nabla J(\mathbf{w}_1).$$

Iteración 3

- **Minibatch:** Seleccionamos otro subconjunto de 50 observaciones.
- **Gradiente:** Calculamos el gradiente:

$$\nabla J(\mathbf{w}_2) = -\frac{1}{50}\sum_{i=1}^{50}\big(y_i - \mathbf{w}_2^T\mathbf{x}_i\big)\mathbf{x}_i.$$

- **Actualización:** Ajustamos los parámetros:

$$\mathbf{w}_3 = \mathbf{w}_2 - \eta\nabla J(\mathbf{w}_2).$$

Ventajas del SGD con Minibatches

- **Convergencia más rápida:** Al realizar actualizaciones frecuentes con cada minibatch, se logra una mejora más rápida en las primeras iteraciones.
- **Costos computacionales reducidos:** Trabajar con minibatches en lugar de todo el conjunto de datos reduce significativamente el costo de cálculo en cada paso.
- **Mayor capacidad de exploración:** La aleatoriedad en la selección de los minibatches introduce ruido beneficioso, ayudando a evitar mínimos locales y a explorar mejor el espacio de soluciones.

Visualización del Proceso

El gráfico a continuación ilustra cómo el costo $J(\mathbf{w})$ disminuye con cada iteración: