Regresión Lineal

4 de junio de 2025

Descenso de Gradiente: Conceptos Fundamentales

- El **Descenso de Gradiente** es un método para la **optimización matemática sin restricciones**.
- Es un **algoritmo iterativo de primer orden** diseñado para **minimizar una función multivariante diferenciable**.
- La idea central es dar **pasos repetidos en la dirección opuesta al gradiente** (o gradiente aproximado) de la función en el punto actual, ya que esta es la dirección de **mayor descenso**.
- Es particularmente útil en **aprendizaje automático** para la **minimización de la función de costo o pérdida**.
- Históricamente, se atribuye a Augustin-Louis Cauchy, quien lo sugirió por primera vez en 1847.

Descenso de Gradiente: El Proceso Iterativo

- El proceso comienza con una estimación inicial $\mathbf{x}_0$ para un mínimo local de la función F.
- La secuencia de puntos se define mediante la siguiente regla de actualización:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \nabla F(\mathbf{x}_n), \ \ n \geq 0.$$

- Aquí, γ_n representa el **tamaño del paso** o la **tasa de aprendizaje** (learning rate). Este valor puede variar en cada iteración.
- El término $\gamma \nabla F(\mathbf{a})$ se resta de $\mathbf{a}$ porque el objetivo es **moverse en contra del gradiente**, hacia el mínimo local.
- Este método garantiza una secuencia **monótona no creciente** de valores de la función: $F(\mathbf{x}_0) \geq F(\mathbf{x}_1) \geq F(\mathbf{x}_2) \geq \cdots$.
- Para funciones convexas, el descenso de gradiente puede converger a la **solución global**, ya que todos los mínimos locales son también mínimos globales.

Aplicación a la Regresión Lineal

- La **Regresión Lineal** es un modelo estadístico que estima la relación entre una variable de respuesta (dependiente) y una o más variables explicativas (independientes).
- En este modelo, las relaciones se representan mediante **funciones lineales de predicción**, cuyos parámetros desconocidos se estiman a partir de los datos.
- El modelo de regresión lineal se expresa como $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$ o en forma matricial $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- Para ajustar un modelo lineal, se estiman los coeficientes de regresión $\boldsymbol{\beta}$ de manera que el **término de error** $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ se minimice.
- Comúnmente, la medida a minimizar es la **suma de los errores al cuadrado** (MSE): $\|\boldsymbol{\varepsilon}\|_2^2$. Esta es la **función de costo** para el descenso de gradiente.
- El vector óptimo de parámetros $\vec{\hat{\beta}}$ es aquel que **minimiza esta suma de cuadrados**.

- En el caso de 2 dimensiones, escribimos:

$$y_i = mx_i + b$$

- El gradiente de esta función de pérdida L es:

$$\frac{\partial L(D, \vec{\beta})}{\partial \vec{\beta}} = -2X^{\mathsf{T}}Y + 2X^{\mathsf{T}}X\vec{\beta}.$$

- Al igualar el gradiente a cero, se obtiene la solución de **mínimos cuadrados** : $\hat{\vec{\beta}} = \left(X^{\mathsf{T}}X\right)^{-1} X^{\mathsf{T}}Y.$
- El descenso de gradiente **itera para encontrar este mínimo** actualizando β en la dirección opuesta al gradiente de la función de pérdida. Para distribuciones de error normales, los resultados de mínimos cuadrados y máxima verosimilitud son idénticos.

Transformación Box-Muller: Generación de Números Aleatorios Normales

- La **Transformación Box-Muller**, formulada por George Edward Pelham Box y Mervin Edgar Muller, es un **método de muestreo de números aleatorios**.
- Su propósito es generar **pares de números aleatorios independientes y distribuidos normalmente estándar** (con expectativa cero y varianza unitaria).
- Para funcionar, requiere una **fuente de números aleatorios distribuidos uniformemente**.
- La necesidad de esta transformación surgió como una **alternativa computacionalmente más eficiente** al método de muestreo por transformada inversa.

Transformación Box-Muller: Forma Básica

- Esta es la forma original dada por Box y Muller.
- Dadas dos muestras independientes, U_1 y U_2, de una **distribución uniforme en el intervalo (0,1)**.
- Luego, las mapea a dos muestras normalmente distribuidas (Z_0 y Z_1) utilizando las siguientes ecuaciones:
  - $Z_0 = \sqrt{-2 \ln U_1} \cos(2\pi U_2)$
  - $Z_1 = \sqrt{-2 \ln U_1} \sin(2\pi U_2)$
- Las variables resultantes Z_0 y Z_1 son **independientes y tienen una distribución normal estándar**.
- La derivación se basa en la propiedad de un sistema cartesiano bidimensional donde las coordenadas X e Y son variables normales independientes.

Transformación Box-Muller: Forma Polar

- Esta forma fue propuesta inicialmente por J. Bell y modificada por R. Knop.
- Toma dos valores, u y v, **independientes y distribuidos uniformemente en el intervalo cerrado** $[-1, +1]$.
- Se calcula $s = u^2 + v^2$. Si $s = 0$ o $s \geq 1$, se descartan u y v, y se elige otro par.
- La principal ventaja es que **evita el uso directo de las funciones seno y coseno**, reemplazándolas por divisiones. Esto puede ser beneficioso cuando las funciones trigonométricas son costosas de calcular.
- Las dos desviaciones normales estándar se producen de la siguiente manera:
  - $z_0 = u \cdot \sqrt{\frac{-2 \ln s}{s}}$
  - $z_1 = v \cdot \sqrt{\frac{-2 \ln s}{s}}$
- La forma polar es un tipo de **muestreo por rechazo**. Descarta aproximadamente el $1 - \pi/4 \approx 21{,}46\,\%$ de los pares de números uniformes de entrada.

Transformación Box-Muller: Comparación y Generalización

- La **forma polar puede ser más rápida** que la forma básica en algunos casos, ya que las divisiones pueden ser computacionalmente menos costosas que las funciones trigonométricas, a pesar de que desecha algunos números generados.
- La forma básica requiere dos multiplicaciones, $1/2$ logaritmo, $1/2$ raíz cuadrada y una función trigonométrica por cada variable normal generada.
- La forma polar requiere $3/2$ multiplicaciones, $1/2$ logaritmo, $1/2$ raíz cuadrada y $1/2$ división por cada variable normal generada, reemplazando una multiplicación y una función trigonométrica con una división y un bucle condicional.
- **Generalización a cualquier distribución normal**:
  - La transformación Box-Muller estándar genera valores de una **distribución normal estándar** ($\mu = 0$, $\sigma = 1$).
  - Para obtener valores de cualquier distribución normal con **media** μ **y varianza** σ^2, se puede aplicar la fórmula: $X = Z\sigma + \mu$, donde Z es un valor normal estándar generado por Box-Muller.

¡Gracias por su atención!