Tarea 06 - Optimización

Giovanni Gamaliel López Padilla

Resumen

abstrac abstract

1. Introducción

La optimización es procedimiento que con sus resultados se toman de decisiones en el análsis de sistemas físicos. Para realizar es la optimización de un sistema o situación se debe de contemplar un objetivo, el cual debe ser caracterizado por una función cuantitativa.
El resultado de realizar la optimización a un sistema puede representarse como un ahorro de tiempo, energía o cualquier objeto que pueda ser reflejado en un número. El objetivo del proceso de una optimización es obtener un conjunto de números o caracteristicas que representen un mínimo o máximo de objeto el cual esta siendo caracterizado. Esto puede ser representado como se encuentra en la ecuación 1.

$$\min_{x \in \mathbb{R}} f(x) \tag{1}$$

donde f es la caracterización cuantitativa del problema y x son los objetos que interactuan con el sistema.
Una optimización local es una solución al problema de optimización en una vecindad alrededor del valor de x encontrado. En cambio una optimización global es aquella solución que es menor o mayor con respecto a todas las demás. La solución de un proceso de optimización no siempre encontrará los valores en que el sistema se situe una optimización global.

2. Marco teórico

2.1. Convexidad

El concepto de convexidad en los problemas de optimización es de gran importancia. Este concepto aporta una mayor facilidad al resolver un problema. La convexidad puede ser aplicado a conjunto o funciones. Se dice que S es un conjunto convexo si un segmento de linea conecta cualquier par de puntos en S. Sean $x, y \in S$, entonces $\alpha x + (1-\alpha)y \in S$ donde $\alpha \in [0,1]$. Se dice que una función es convexa si su dominio S es convexo y cumplen con la propiedad escrita en la ecuación 2.

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) \tag{2}$$

Donde $x, y \in S$ y $\alpha \in [0,1]$.

2.2. Condiciones necesarias

Suponiendo que $f : \mathbb{R}^n \to \mathbb{R}$ y que es continuamente diferenciable y $p \in \mathbb{R}^n$. Entonces, tenemos que

$$f(x+p) = f(x) + \nabla f(x+tp)^T p$$

donde $t \in (0,1)$. De igual manera, si f es doblemente continua diferenciable, entonces se tiene que:

$$\nabla f(x+p) = \nabla f(x) \int_0^1 \nabla^2 f(x+tp)pdt$$

por lo tanto

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla f(x+tp)p$$

Con estas hechos, se puede demostrar que si x^* es un punto estacionario entonces el gradiente y el hessiano de f tiene las caracteristicas mostradas en la ecuación 3.

$$\nabla f(x^*) = 0 \qquad \nabla^2 f(x^*) \succ 0 \tag{3}$$

También se puede demostrar que si $\nabla^2 f$ es continua en una vecindad alrededor de x^* y que $\nabla f(x^*) = 0$ y $\nabla^2 f(x^*)$ es positiva definida. Entonces x^* es un mínimo local de f.

2.3. Direcciones de búsqueda

La dirección mas eficiente usando un método de descenso es usar una dirección p descrita en la ecuación 4 en cada paso.

$$p_k = -\nabla f_k \tag{4}$$

En cada iteración de la linea de búsqueda se implementa un cambio en la posición siguiendo la ecuación 5.

$$x_{k+1} = x_k + \alpha_k p_k \tag{5}$$

Donde α_k es un escalar positivo llamado tamaño de paso.

2.4. Condiciones de Wolfe

Existen maneras de verificar si el α_k elegido para el k-esimo paso es el óptimo para seguir en la dirección p_k. La condición de decrecimiento suficiente esta descrita en la ecuación 6.

$$f(x_k + \alpha_k p_k) \leq f(x_k) c_1 \alpha_k \nabla f_k^T p_k \tag{6}$$

La interpretación de este resultado es que la función f debe ser proporcional al tamaño de paso α_k y la derivada direccional $\nabla f_k^T p_k$ para una constante $c_1 \in (0,1)$. Esta condición también es conocida como la condición de Armijo.
La condición de decrecimiento suficiente no es la única que se tiene que contemplar, esto debido a que existen α_k muy pequeñas que satisfacen a la desigualdad. Es por ello que se tiene que contemplar la condición de curvatura. La condición de curvatuva esta definida en la ecuación 7.

$$\nabla f(x_k + \alpha_k p_k)^T \geq c_2 \nabla f_k^T p_k \tag{7}$$

Donde $c_2 \in (0,1)$.
La búsqueda de un α_k que cumpla las condiciones de las ecuaciones 6 y 7 esta descrito en el algoritmo 1.

Algorithm 1: Búsqueda de un α que cumpla las condiciones de las ecuaciones 6 y 7

$\alpha_0 \leftarrow 0$ $\quad \alpha_i \leftarrow 1$ $\quad \beta \leftarrow \infty$
repeat
  if *armijo conditon*(α_i) *or curvature contidion*(α_i) **then**
    if *armijo condition*(α_i) **then**
      $\beta \leftarrow \alpha_i$
      $\alpha_i = \frac{\beta+\alpha}{2}$
      else if *curvature condition* (α_i) **then**
        $\alpha \leftarrow \alpha_i$
        if β *equals* ∞ **then**
          $\alpha_i \leftarrow 2\alpha$
          else
            $\alpha_i = \frac{\beta+\alpha}{2}$
  else
    break
return α_i

2.5. MNIST

El conjunto de datos MNIST consta de 70,000 números escritos a mano. Este conjunto de datos fue dividido en datos de entrenamiento (50,000), datos de prueba (10,000) y datos de validación (10,000). Cada imagen esta constitiuda de 28x28 pixeles. En la figura 1 se muestran algunos de ellos.

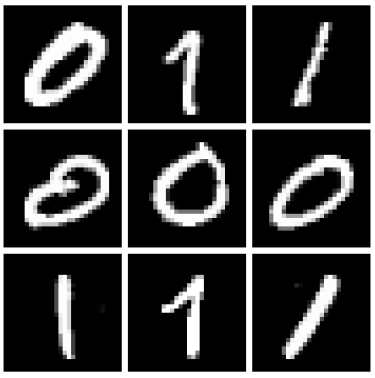

Figura 1: Algunas imagenes contenidas en el conjunto MNIST.

2.6. Función Log likelihood

Definimos como la función log likelihood en la ecuación 8.

$$h(\beta, \beta_0) = \sum_{i=1}^{n} y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \qquad \pi_i(\beta, \beta_0) = \frac{1}{1 + exp(-x_i^T \beta - \beta_0)} \tag{8}$$

donde $x_i \in \mathbb{R}^n$ y $y_i \in \{0, 1, \ldots, 9\}$. En nuestro caso $n = 784$ y $y \in \{0, 1\}$. Para obtener una reducción de parametros en la ecuación 8 aplicaremos un aumento de dimensión al vector x de tal forma que

$$x = [x_0, x_1, \ldots, x_{784}, 1]^T$$

Entonces

$$x^T\beta - \beta_0 \to x^T\beta$$

donde

$$\beta = [\beta_1, \beta_2, \ldots, \beta_{785}, \beta_0]^T$$

por lo que el vector x es ahora elemento del conjunto $\mathbb{R}^{785}$. Entonces la ecuación 8 puede escribirse como en la ecuación .

$$h(\beta) = \sum_{i=1}^{n} y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i) \qquad \pi_i(\beta) = \frac{1}{1 + exp(-x_i^T\beta)} \tag{9}$$

Calculando el gradiente de la función con respecto a β se obtiene lo siguiente:

$$\frac{\partial h}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}\frac{\partial \pi_i}{\partial \beta_j} + \frac{1 - y_i}{1 - \pi_i}\frac{\partial \pi_i}{\partial \beta_j}$$

calculando $\frac{\partial \pi_i}{\partial \beta_j}$ se obtiene lo siguiente:

$$\begin{aligned}\frac{\partial \pi_i}{\partial \beta_j} &= \frac{x_j exp(-x_i^T\beta)}{(1 + exp(-x_i^T\beta))^2} \\ &= x_j exp(-x_i^T\beta)\pi_i^2 \\ &= x_j(1 - \pi_i)\pi_i\end{aligned}$$

entonces

$$\begin{aligned}\frac{\partial h}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i}{\pi_i}\frac{\partial \pi_i}{\partial \beta_j} + \frac{1 - y_i}{1 - \pi_i}\frac{\partial \pi_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{y_i}{\pi_i}(x_j(1 - \pi_i)\pi_i) + \frac{1 - y_i}{1 - \pi_i}(x_j(1 - \pi_i)\pi_i) \\ &= \sum_{i=1}^{n} x_j y_i(1 - \pi_i) - x_j(1 - y_i)\pi_i \\ &= \sum_{i=1}^{n} x_j(y_i - \pi_i y_i + \pi_i y_i - \pi_i) \\ \frac{\partial h}{\partial \beta_j} &= \sum_{i=1}^{n} x_j(y_i - \pi_i)\end{aligned}$$

por lo tanto

$$\frac{\partial h}{\partial \beta_j} = \sum_{i=1}^{n} x_j(y_i - \pi_i) \tag{10}$$

Con su función y gradientes definidos podemos llegar a aplicar el método de descenso de gradiente con una busqueda lineal empleando el algoritmo 1.

3. Resultados

En la tabla 1 se encuentran los parámetros ingresados para el algoritmo de descenso del gradiente para la función log-likelihood.

c_1	c_2	Tolerancia
$1x10^{-4}$	0.9	$1x10^{-6}$

Tabla 1: Parámetros usados para el algoritmo 1 con la función log-likelihood.

En la figura 2 se muestran las evaluaciónes de las ecuación 9 y la norma de la ecuación 10 en cada iteración realizada.

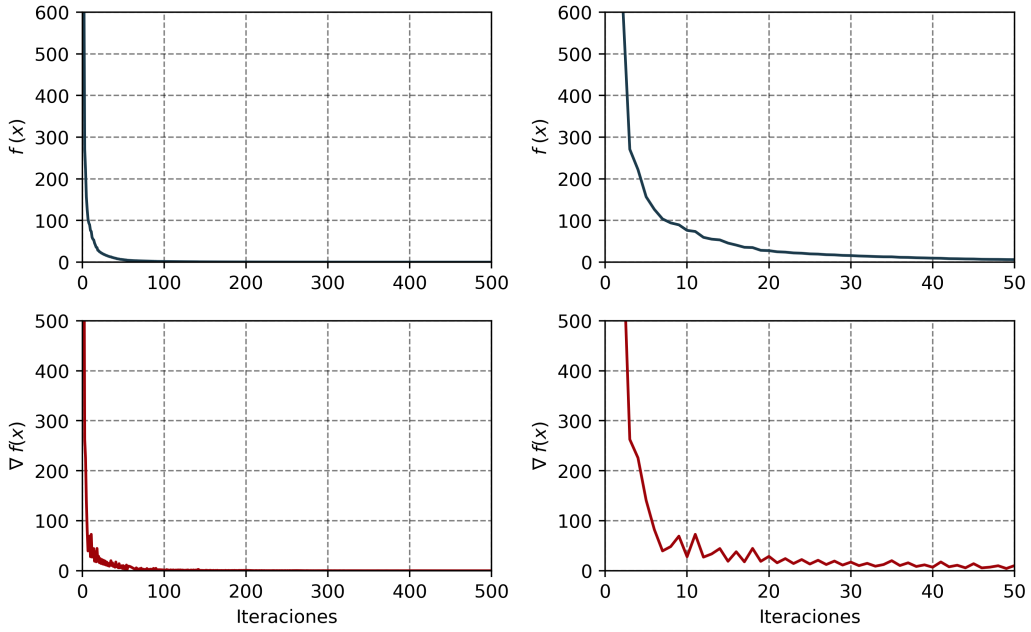

Figura 2: Resultados de la evaluación de la ecuación 9 y la norma de la ecuación evaluada 10 en cada iteración realizada. De lado izquierdo se muestran los resultados hasta las 500 iteraciones y en en el lado derecho hasta las 50 iteraciones.

Realizando el calculo del error que esta definido en la ecuación 11. Se obtuvo que para el conjutno de datos de prueba es de 0.00047281.

$$\text{error} = \frac{1}{n}\sum_{i=1}^{n}\left|1_{\pi_i(\beta)>0.5(x_i)-y_i}\right| \tag{11}$$

4. Conclusiones

El método de descenso del gradiente con el tamaño de paso obtenido a partir del algoritmo de bisección es bueno para este conjunto de datos seleccionado debido a que llegan a un óptimo global de la función con un pequeño número de iteraciones comparado con el tamaño de valores diferentes que se dio como entrenamiento. A pesar que en algunos puntos el gradiente parecia tomar valores oscilantes estos ayudaron para que el punto fuera encontrado de forma exitosa. Se intento utilizar el algoritmo de back traking pero este tardo alrededor de 2 días en llegar al mismo punto que el algoritmo de bisección.