Problemas Convexos

Carlos Malanche

13 de marzo de 2018

1. Optimización

1.1. Búsqueda en retícula (*grid search*)

La búsqueda en retícula es el método más simple (y más ineficiente) para encontrar un punto óptimo. Consiste en discretizar el dominio de los parámetros (muchas veces en puntos equidistantes) y evaluar con fuerza bruta el valor de la función objetivo para poder elegir el *mínimo*. Puesto de manera formal, si la patición de la i-dimensión en el espacio de parámetros de $\underset{\rightarrow}{c}$ la denotamos por P_{c_i} el resultado será

$$\underset{\rightarrow}{c}_0 = \text{mín}\{\underset{\rightarrow}{c} | c_i \in P_{c_i} \forall i = 1, \cdots, m\} \tag{1}$$

En el modestísimo caso en el que hay $m = 32$ parámetros (imagínen que podemos pasar una imágen, con millones de pixeles, siendo cada uno una *medición*), tendríamos que realizar 2^{32} evaluaciones de la función objetivo para buscar en los extremos de las particiones nadamás, y esto sólo dará buenos resultados bajo la suposición de que la función objetivo es lineal en todos los parámetros (nunca pasará).

1.1.1. La versión estocástica

La versión estocástica de este problema consiste en tomar puntos al azar, evaluarlos y quedarnos con el mejor. El panorama no es alentador. Todo es por algo conocido como *la maldición de la dimensionalidad*. Tiene un nombre demasiado alarmista para lo que es. Sólo consiste en observar el crecimiento exponencial del espacio de muestreo con el crecimiento lineal de los parámetros de los que está compuesto.

Sin embargo, es curioso ver que una búsqueda aleatoria tiene una *alta* probabilidad de quedar cerca de un óptimo: Imaginemos el espacio de búsqueda, creado a partir de los extremos de las particiones de cada variable en la búsqueda en malla. Asumiremos que vamos a tomar puntos de manera aleatoria bajo una distribución uniforme en cada dimensión.

Si esto es así, la probabilidad de que un punto quede en el 5 % del espacio que rodea al óptimo sería 0.05. Ahora, nos preguntamos: Después de k intentos, cuál es la probabilidad de que **ninguno** de los k puntos se encuentre en el 5 % al rededor del óptimo? Debería resultar claro que la respuesta es

$$(1 - 0{,}05)^k$$

Si pedimos que su complemento en probabilidad (es decir, que al menos uno haya caído en el 5 %) sea al menos 95 % (considerablemente bueno, digamos), podemos resolver para k, lo que nos da el bellísimo resultado de

$$k \geq \frac{\log(0{,}05)}{\log(0{,}95)} \approx 60 \tag{2}$$

Es decir, independientemente del número de dimensiones, en sólo 60 evaluaciones hay un %95 de probabilidades de estar cerca del máximo/mínimo. Léase: Es mejor utilizar una búsqueda aleatoria para encontrar al menos un buen punto de partida.

1.2. Descenso de gradiente (*gradient descent*)

De lo que se trata el algoritmo es de iterativamente caminar en dirección contraria al gradiente, en algún paso moderado. Si denotamos con $\nabla\mathcal{L}(\underset{\rightarrow}{c})$ el gradiente de la función objetivo evaluado en $\underset{\rightarrow}{c}$, y suponemos $\underset{\rightarrow}{c}_0$ una *buena adivinanza* de punto de comienzo, la iteración número $i + 1$ estará dada por

$$\underset{\sim}{c}_{i+1} = \underset{\sim}{c}_i - \eta \nabla \mathcal{L}(\underset{\sim}{c}_i) \tag{3}$$

Al parámetro η se le conoce como el *paso* del algoritmo. Hay varios estudios al respecto de este parámetro, pues de él dependerá el ritmo de convergencia y la estabilidad del descenso de gradiente.

1.3. Convergencia

El análisis de convergencia del algoritmo es algo complicado, es por ello que veremos el caso en el que se utiliza un tamaño de paso constante.

Definición. *Sea $f : U \to \mathbb{R}$ una función. Se dice que la función es Lipschitz continua si existe un valor L tal que*

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\|$$

Esto acota en cierta medida el ritmo de cambio de una función. Tomemos $f \coloneqq \nabla\mathcal{L}$ y supongamos que el gradiente es Lipschitz-continuo de constante L, además de suponer que nuestra función es *convexa*. Vamos a probar que el ritmo de convergencia va como el inverso del tamaño del paso.

Al suponer que el gradiente tiene constante de Lipschitz L, ocurre que el Hessiano queda acotado por L, que se puede escribir de otro modo como

$$\nabla^2\mathcal{L} - LI \geq 0 \tag{4}$$

(una manera de probarlo es utilizar el teorema del valor medio y la definición de Lipschitz continuo). Usando la definición de semidefinido positivo, tomamos convenientemente el vector $(x - y)$ para escribir

$$(x-y)^T(\nabla^2\mathcal{L} - LI)(x-y) \leq 0 \tag{5}$$
$$(x-y)^T\nabla^2\mathcal{L}(x-y) \leq L\|x-y\|^2 \tag{6}$$

Usando el teorema del residuo de Taylor, podemos encontrar $z \in [x, y]$ (donde estoy representando con ese intervalo la linea que une $\boldsymbol{x}$ con $\boldsymbol{y}$) tal que

$$\begin{aligned}\mathcal{L}(y) &= \mathcal{L}(x) + \nabla\mathcal{L}(x)^T(y-x) + \frac{1}{2}(x-y)^T\nabla^2\mathcal{L}(z)(x-y)\\ &\leq \mathcal{L}(x) + \nabla\mathcal{L}(x)^T(y-x) + \frac{L}{2}\|y-x\|^2\end{aligned}$$

hagamos que $y \coloneqq x^+ = x - \eta\nabla\mathcal{L}(x)$ (es decir, y es el paso inmediato del descenso de gradiente al haber evaluado x).

$$\begin{aligned}\mathcal{L}(x^+) &\leq \mathcal{L}(x) + \nabla\mathcal{L}(x)^T(x - \eta\nabla\mathcal{L}(x) - x) + \frac{L}{2}\|x - \eta\nabla\mathcal{L}(x) - x\|^2\\ &= \mathcal{L}(x) - \eta\nabla\mathcal{L}(x)^T\nabla\mathcal{L}(x) + \frac{\eta^2 L}{2}\|\nabla\mathcal{L}(x)\|^2\\ &= \mathcal{L}(x) - (1 - \frac{\eta L}{2})\eta\|\nabla\mathcal{L}(x)\|^2\end{aligned}$$

Tomemos $0 < \eta < 1/L$, con lo que podemos ver que $-(1 - \frac{\eta L}{2}) = \frac{\eta L}{2} - 1$ que en el mejor caso (es decir $\eta = 1/L$) el valor queda acotado por $1/2$.

$$\mathcal{L}(x^+) \leq \mathcal{L}(x) - \frac{\eta}{2}\|\nabla\mathcal{L}(x)\|^2 \tag{7}$$

Primer observación: bajo las condiciones establecidas sobre $\mathcal{L}$ y sobre η, el descenso de gradiente siempre conduce a un valor mejor (más pequeño). Recordemos, además, que $\mathcal{L}(x)$ es convexa. Esto implica que

$$\mathcal{L}(x) \leq \mathcal{L}(x^*) + \nabla\mathcal{L}(x)^T(x - x^*) \tag{8}$$

(la expresión es la misma que la de la clase pasada pero multiplicada por -1, y hemos bautizado a x^* como el valor óptimo). Bien, usemos esta definición dentro del resultado anterior

$$\begin{aligned}\mathcal{L}(x^+) &\leq \mathcal{L}(x) - \frac{\eta}{2}\|\nabla\mathcal{L}(x)\|^2 \\ &\leq \mathcal{L}(x^*) + \nabla\mathcal{L}(x)^T(x - x^*) - \frac{\eta}{2}\|\nabla\mathcal{L}(x)\|^2 \\ &= \mathcal{L}(x^*) + \frac{1}{2\eta}(\|x - x^*\|^2 - \|x - x^* - \eta\nabla\mathcal{L}(x)\|^2) \\ &= \mathcal{L}(x^*) + \frac{1}{2\eta}(\|x - x^*\|^2 - \|x^+ - x^*\|^2)\end{aligned}$$

Como última observación, ya que el descenso de gradiente es no creciente, podemos concluir que en el paso número k, la distancia al óptimo será mejor que el promedio de todos los pasos anteriores

$$\mathcal{L}(x^{(k)}) - \mathcal{L}(x^*) \leq \frac{1}{k}\sum_{i=1}^{k}(\mathcal{L}(x^{(i)}) - \mathcal{L}(x^*)) \tag{9}$$

Esa suma se convierte en una suma telescópica, pues

$$\begin{aligned}\sum_{i=1}^{k}(\mathcal{L}(x^{(i)}) - \mathcal{L}(x^*)) &\leq \sum_{i=1}^{k}\frac{1}{2\eta}(\|x^{(i-1)} - x^*\|^2 - \|x^{(i)} - x^*\|^2) \\ &= \frac{1}{2\eta}(\|x^{(0)} - x^*\|^2 - \|x^{(i)} - x^*\|^2) \\ &\leq \frac{1}{2\eta}\|x^{(0)} - x^*\|^2\end{aligned}$$

Entonces, el ritmo de convergencia depende inversamente del tamaño del paso, y de la distancia al óptimo en un principio (además del número de iteraciones)

$$\mathcal{L}(x^{(k)}) - \mathcal{L}(x^*) \leq \frac{\|x^{(0)} - x^*\|^2}{2\eta k} \tag{10}$$

Noten porfavor que este ritmo de convergencia salió de poner condiciones bastante estrictas a la función objetivo.

1.3.1. Subgradiente

Increíblemente, podemos seguir aplicando nuestro descenso de gradiente aún cuando la función no es diferenciable. Pero para ello es necesario definir un *subgradiente*.

Definición. *Sea $f : U \to \mathbb{R}$ una función definida sobre un conjunto convexo y abierto, subconjunto de $\mathbb{R}^m$. Un vector $\underset{\rightarrow}{v}$ es un subgradiente de f en el punto $\underset{\rightarrow}{x}_0$ si para toda $\underset{\rightarrow}{x} \in U$*

$$f(\underset{\rightarrow}{x}) - f(\underset{\rightarrow}{x}_0) \geq \underset{\rightarrow}{v} \cdot (\underset{\rightarrow}{x} - \underset{\rightarrow}{x}_0)$$

Hay un gran parecido entre la definición de función convexa diferenciable y la de un subgradiente. Como última nota, si la función es diferenciable, el gradiente común y corriente es el único subgradiente de f.

Ya podemos escribir nuestro algoritmo, que queda de la siguiente manera (denotando con $\hat{\nabla}\mathcal{L}$ el subgradiente de la función objetivo)

$$\underset{\rightarrow}{c}_{i+1} = \underset{\rightarrow}{c}_i - \mu\hat{\nabla}\mathcal{L}(\underset{\rightarrow}{c}_i)$$

1.4. Descenso de gradiente estocástico (*stochastic gradient descent*)

Muchas veces los cantidad de puntos de *entrenamiento* será muy grande y afectará mayormente el desempeño de nuestros algoritmos. En el caso del descenso del gradiente, calcular el gradiente puede tomar mucho tiempo. Una opción es tirar parte de la información que tenemos para hacer el cómputo del gradiente más rápido, pero si no sabemos *cuál es la información más útil* esto es una mala idea. Llega el descenso de gradiente estocástico al rescate.

Simplemente consiste en tomar uno de los datos de manera aleatoria y avanzar en la dirección del gradiente que este dato proporciona. De esta manera, para un punto fijo

$$E[\nabla_s\mathcal{L}] = \nabla\mathcal{L} \tag{11}$$