Entregable 5: Descenso por gradiente estocástico

18 de noviembre de 2020

Ejercicio 1 - Optimización estocástica (*Convergencia a una bola*)

Se desea hallar el valor $\theta^\star$ que minimiza la función

$$u(\theta) = E_{X,y}[||X\theta - y||^2]$$

La matrix X es aleatoria, según el modelo de ruido aditivo $X = A + N$ con

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

y

$$N = \begin{bmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{bmatrix}$$

donde $n_{ij} \sim \mathcal{N}(0, \sigma^2)$ independientes entre si para $i = 1, 2,\ j = 1, 2$, de varianza $\sigma^2 = 1$.
El vector y tambien es aleatorio, siguiendo el modelo $y = X * \theta_0 + w$ con

$$\theta_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

y

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

donde $w_i \sim \mathcal{N}(0, \sigma^2)$ independientes entre si para $i = 1, 2$, de varianza $\sigma^2 = 1$.

a) Muestre que $f(\theta) = \frac{du}{d\theta} = 2(A^T A + \sigma^2 I)\theta - 2(A^T A + \sigma^2 I)\theta_0$, anulándose en $\theta = \theta_0$.

b) Dadas muestras i.i.d. $\xi_k = (X_k, y_k)$ distribuidas como X e y, obtenga la secuencia de desenso por gradiente estocástico

$$\theta_{k+1} = \theta_k - \alpha_k F(\theta_k, X_k, y_k)$$

especificando la función $F(\theta_k, X_k, y_k)$..

c) Implemente $K = 10000$ pasos de SGD según la iteración hallada en la parte anterior. Grafique la secuencia de valores de θ_k en el plano. Pruebe con paso constante $\alpha_k = 0{,}1$, $\alpha_k = 0{,}01$, y $\alpha_k = 0{,}001$, y con paso decreciente $\alpha_k = 0{,}1/k$

d) (opcional) Obtenga las constantes c_1 y c_2 de las hipótesis del teorema de Robbins Monró, y represente la bola de radio $B = \frac{c_1 \alpha}{2c_2}$ alrededor de θ_0.

Ejercicio 2 - Entrenamiento de una neurona artificial.

Considere los pares (x, y) donde los vectores $x \in \mathbb{R}^N$ pertenecen a dos clases indicadas por la variable de activación $y \in 0, 1$. Se desea obtener el parámetro $a \in \mathbb{R}^N$ solución de

$$\min_{a \in \mathbb{R}^N} E_{x,y}[G(a; x, y)] = \min_{a \in \mathbb{R}^N} E_{x,y}\left[(y - \mathrm{Relu}(a'x))^2\right] \tag{1}$$

con

$$\mathrm{Relu}(z) = \begin{cases} \epsilon z, & z \leq 0 \\ z, & z \geq 0 \end{cases} \tag{2}$$

paramétrico en $\epsilon \in [0, 1)$.

a) Halle el subgradiente de $G(a; x, y)$ respecto a a.

b) Escriba la iteración del algoritmo SGD que resuelve (1) a partir de muestras (x_k, y_k).

En las siguientes partes se programará SGD para clasificar imágenes de gatos y conejos. Las imágenes de $M = 30$ gatos están contenidas en el archivo *Gatos.asc* y $M = 30$ conejos en *Conejos.asc*, respectivamente. Estos archivos contienen matrices $\mathbf{G} \in \mathbb{R}^{N \times M}$ y $\mathbf{C} \in \mathbb{R}^{N \times M}$ cuyas columnas tienen dimensión $N = (256)(256)(3) + 1$ y corresponden a una imagen de tamaño 256×256 píxeles codificadas en tres canales de color RGB y con un elemento adicional siempre igual a uno para lograr una función afín. Los vectores se pueden visualizar como imagen en Matlab usando la función mostrar_imagen$(x, y, 256)$.

c) Corra el algoritmo diseñado en la parte anterior para los $K = 40$ vectores x_k contenidos en los las primeras $K = 20$ columnas de $\mathbf{C}$ y $\mathbf{R}$. Utilice las etiquetas $y = 0$ para los gatos y $y = 1$ para los conejos. Seleccione un paso constante $\alpha = 1e^{-9}$ y $\epsilon = 0{,}1$.

d) A partir de la solución a_K obtenida en la parte anterior clasifique los vectores x en las últimas 10 columnas de $\mathbf{C}$ y $\mathbf{R}$. Presente una gráfica del valor de $z = a'_K x$ contra el número de muestra, observando la clasificación obtenida para las muestras de entrenamiento y validación.

e) (opcional) Estudie la convexidad de $g(a) = E[G(a; x, y)]$.

f) (opcional) Seleccione distintos grupos de entrenamiento y validación y reevalue los resultados.

g) (opcional) Evalúe el error promedio en la muestra de validación contra el número de iteración $k = 1, \ldots, K$ confirmando que el clasificador aprende progresivamente de los datos. Para ello intercale gatos y conejos en el entrenamiento.

h) (opcional) Clasifique esta muestra con la SVM desarrollada en el Entregable 3.

Ejercicio 3 - *Martingala para SGD sin convexidad (opcional)*

Se quiere minimizar la función $g(\theta) = E_\xi[G(\theta, \xi)]$ para lo cual se considera el algoritmo SGD

$$\theta_{k+1} = \theta_k - \alpha_k \nabla G_\theta(\theta_k, \xi_k). \tag{3}$$

No asumiremos que la función es convexa, pero si que se cumplen las siguientes hipóteis

H1: El gradiente es Lipschitz, respecto a θ, es decir que $\exists L > 0$ tal que $\|\nabla_\theta g(\theta) - \nabla_\theta g(\theta')\| \leq L\|\theta - \theta'\|$.

H2: Se puede cambiar el orden de derivación e integración en

$$\nabla_\theta g(\theta) = \nabla_\theta E_\xi[G(\theta, \xi)] = \nabla_\theta \int_\xi G(\theta, \xi)p(\xi)d\xi = \int_\xi \nabla_\theta G(\theta, \xi)p(\xi)d\xi = E[\nabla_\theta G(\theta, \xi)].$$

H3: $E[\|\|\nabla_\theta G(\theta, \xi)\|^2] \leq C$.

H4: $\sum_{k=0}^\infty \alpha_k^2 \leq \infty, \quad \sum_{k=0}^\infty \alpha_k^2 = \infty$.

a) Utilizando el teorema del valor medio y la condición de Lipschitz, muestre que

$$g(\theta_{k+1}) \leq g(\theta_k) - \alpha_k \nabla_\theta g(\theta_k)^T \nabla_\theta G(\theta_k, \xi_k) + L\|\theta_{k+1} - \theta_k\|^2.$$

b) Pruebe a partir de ello que $S_k = g(\theta_k) + LC \sum_{j=k}^\infty \alpha_k^2$ es una supermartingala, mostrando que cumple $E_\xi[S_{k+1}|\xi_0, \ldots, \xi_{k-1}] \leq S_k - \alpha_k\|\nabla_\theta g(\theta_k)\|$.

c) Muestre usando la parte anterior que $\sum_{k=0}^\infty \alpha_k \|\nabla_\theta g(\theta_k)\|^2 \leq \infty$.

d) Concluya que $\liminf_k \|\nabla_\theta g(\theta_k)\| = 0$, luego θ_k acumula en un los puntos estacionarios de $g(\theta)$.

Ejercicio 4 - *Filtro adaptativo para ECG (opcional)*

Se busca diseñar un filtro adaptativo para obtener una secuencia electrocardiográfica (ECG) fetal $\{e(n)\}_{n\in\mathbb{N}}$, a partir de señales adquiridas del abdomen y el corazón maternos.

Como datos se tiene la señal adquirida del abdomen $z(n) = e(n) + y(n) + u(n)$, la cual se modela como la señal deseada $e(n)$ con interferencia aditiva $y(n)$ proveniente del corazón materno y ruido de adquisición $u(n)$ independiente. También se mide la señal de referencia $x(n)$ adquirida del torax materno. La transferencia de los latidos maternos $x(n)$ hacia el ambodmen se modela como una función lineal $y(n) = a_0x(n) + a_1x(n-1) + \ldots + a_{m-1}x(n-m+1)$.

Abreviando $a = (a_0, a_1, \ldots, a_{m-1})^T$ y $x_n = (x(n), x(n-1), \ldots, x(n-m+1))^T$, se busca el estimador $\hat{e}(n) = y(n) - x_n'a^*$, donde

$$a^* = \arg\min_{a\in\mathbb{R}^m} \left\{ E[\varphi(a; x, z)] = E_{x,z}\left[\frac{1}{2}(z - x'a)^2\right] \right\}$$

a) Derive el optimizador estocástico que para cada instante $n \in \mathbb{N}$ incorpora un nuevo par $(x(n), < (n))$ y actualiza la solución

$$a^{n+1} = a^n - \delta\nabla_a\varphi(a_n; x_n, z_n)$$

b) Implemente el algoritmo con $m = 15$ y paso δ constante a partir de los datos en los archivos *xECG.asc, zECG.asc.*

c) Estudie la dependencia del error contra el paso y relacione con el ejercicio 1.

Ejercicio 5 - *Hipótesis de convexidad fuerte*(opcional)

Considere el siguiente problema de optimización estocástica

$$\theta_0 = \arg \min_{\theta \in \Theta} \{g(\theta) = E_\xi[G(\theta, \xi)]\}$$

donde la variable determinística $\theta \in \mathbb{R}^n$ está restringida al conjunto convexo $\Theta \subset \mathbb{R}^n$.

a) Pruebe que la convexidad fuerte de $g(\theta)$ implica la siguiente hipótesis del teorema de Robbins-Monro

$$c\|\theta - \theta^\star\|^2 \leq (\theta - \theta^\star)^T \nabla g(\theta)$$

Sugerencia: recordar que la condición de optimalidad para $\theta^\star \in \Theta$ es $\nabla g(\theta^\star)(\theta - \theta^\star) \geq 0, \ \forall \theta \in \Theta$.

Ejercicio 6 - Aproximación estocástica con paso constante (opcional)

En la prueba del Teorema de Robbins Monro vista en clase se deduce la cota

$$e_{k+1} \leq (1 - 2c_2\alpha_k)e_k + c_1\alpha_k^2 \tag{4}$$

donde $\mathcal{E}_k = E\left(\|\hat{\theta}^\star - \theta^\star\|^2\right)$, c_1 y c_2 son constantes que provienen de las hipótesis, y α_k es el paso de optimización. Con paso constante $\alpha_k = \alpha$, a cota (4) puede reescribirse como

$$e_{k+1} \leq \gamma e_k + c_1\alpha^2 \tag{5}$$

$$\gamma := 1 - 2c_2\alpha \tag{6}$$

a) A partir de (5) muestre que

$$e_k \leq \gamma^k e_0 + \frac{1 - \gamma^k}{1 - \gamma} c_1\alpha^2 = B_k(\alpha) \tag{7}$$

b) Halle el límite $B(\alpha) = \lim_{k \to \infty} B_k(\alpha)$ de la cota hallada en (7). El resultado dependerá de c_1 y c_2 que aparecen junto con α al definir γ en (6).

c) Dado $B > 0$ halle el valor de α tal que $B(\alpha) = B$.