به نام خدا

تمرین دوم درس یادگیری عمیق
دکتر محمدی
مجتبی نافذ 96431335

سوال یک:
تفاوت stochastic , mini-batch , batch :
در حقیقت الگوریتم پایه ی بهینه سازی lost function الگوریتم Gradient Descent است که همان batch gradient descent نامیده میشود
و stochastic , min-batch به نوعی variation هایی از batch میباشند.
تفاوت این الگوریتم ها در مقدار دیتایی ایست که در هر epoch میگیرند.
الگوریتم batch : هر epoch روی کل دیتاست می باشد.
الگوریتم mini-batch: هر epoch رو یک batch و در واقع یک بخشی از دیتا است (معمولا 16و 32و 64و ...)
الگوریتم stochastic : هر epoch روی یک داده از دیتاست می باشد.

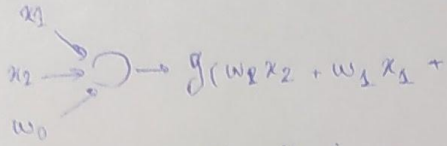
مقایسه SGD با GD :
الگوریتم batch :
رو دیتاست بزرگ میتواند خیلی کند باشد چون در هر epoch فقط یک آپدیت دارد.
در دیتاست بزرگ دیتا قابلیت fit شدن در حافظه را ندارد. چون حجم دیتا بیشتر از حافظه است.
در سطوح non-convex ممکن است در در مینیمم محلی گیر کند.
الگوریتم SGD:
در کل SGD شانس بیشتری از GD برای فرار مینیمم محلی را دارد.
و قطعا سریع تر به دقت های بالا خواهد رسید و همگرا خواهد شد.
به دلیل عملکرد greedy خود تخمینی از گرادیان میزند.
به دلیل فرکانس نوسانات در داده ها ، نزدیک نقطه ی مینیمم overshooting خواهد داشت.

مشکلات SGD :
به دلیل عملکرد greedy خود تخمینی از گرادیان میزند.
به دلیل فرکانس نوسانات در داده ها ، نزدیک نقطه ی مینیمم overshooting خواهد داشت.
در مناطقی که شیب کم است سرعت همگرایی به شدت کند خواهد بود و در شیب صفر حرکت نخواهد کرد. و انگونه بیشتر در مینیمم محلی هم گیر هواد کرد.

رفع مشکلات SGD با استفاده از Momentum :
رفع مشکل سرعت همگرایی در SGD : در روش momentum در واقع ما علاوه بر گرادیان در این نقطه به سرعت حرکت در نقاط قبلی هم باید توجه کنیم یعنی رفته رفته با توجه به سرعت (گرادیان) مکان های قبلی شتاب بگیریم و step size مان را بیشتر کنیم و با گام های بلندتری حرکت کنیم. که این گونه ما در جاهایی که شیب کم است سرعت همگرایی بالاتری خواهیم داشت و گاها میتوانیم از مینیمم های محلی فرار کنیم.

سوال ۲

الف:



$$g(w_2x_2 + w_1x_1 + w_0) = S_i$$

use MSE

$$L = \frac{1}{N}\sum_{i=1}^{N} L^{(i)}(s^i, y^i) = \frac{1}{N}\left(\sum_{i=1}^{N}(y_{(i)} - s_{(i)})^2\right) =$$

$$L = \frac{1}{N}\sum_{i=1}^{N}\left[y_{(i)} - \max\left(0, (w_2x_2^{(i)} + w_1x_1^{(i)} + w_0)\right)\right]^2 \Rightarrow$$

$$\frac{\partial L}{\partial w_1} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial w_1}\left[y_{(i)} - \max\left(0, (w_2x_2^{(i)} + w_1x_1^{(i)} + w_0)\right)\right]^2 =$$

$$\frac{\partial L}{\partial w_1} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial w_1}\left[y_{(i)} - (w_2x_2^{(i)} + w_1x_1^{(i)} + w_0)\right]^2 & : w_2x_2 + w_1x_1 + w_0 > 0 \\ \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial w_1}[y_i]^2 & : w_2w_2 + w_1x_1 + w_0 \le 0 \end{cases}$$

$$\frac{\partial L}{\partial w_1} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} -2\overbrace{\left[y_i - w_2x_2^{(i)} - w_1x_1^{(i)} - w_0\right]}^{e_i}x_1 & : \overbrace{w_2x_2 + w_1x_1 + w_0}^{w\cdot x} > 0 \\ 0 & : w_2x_2 + w_1x_1 + w_0 \le 0 \end{cases}$$

$$w_{1(t+1)} = w_{1(t)} - \eta\frac{\partial L}{\partial w_1}$$

$N \rightarrow$ تعداد داده‌های آموزشی هر batch

برای w_2 و w_0 هم داریم:

$$\frac{\partial L}{\partial w_0} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} -2\left[y_i - w_2x_2^{(i)} - w_1x_1^{(i)} - w_0\right] & : w\cdot x > 0 \\ 0 & : w\cdot x \le 0 \end{cases}$$

$$\frac{\partial L}{\partial w_2} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} -2\left[y_i - w_2x_2^{(i)} - w_1x_1^{(i)} - w_0\right]x_2 & : w\cdot x > 0 \\ 0 & : w\cdot x \le 0 \end{cases}$$

دقت کنید که آپدیت w_2 و w_0 هم کاملا مانند w_1 و از رابطه زیر صورت میگیرد:

$$w_{i(t+1)} = w_{i(t)} + \eta\frac{\partial L}{\partial w_i}$$

sigmoid تابع [ب] . .

$$L = \frac{1}{N} \sum_{i=1}^{N} \left[y_i - \frac{1}{1+e^{-z}} \right]^2$$

$$z = w_2 x_2 + w_1 x_1 + w_0$$

قاعده مشتق زنجیره‌ای :

$$\frac{\delta L}{\delta w_1} = \frac{1}{N} \sum_{i=1}^{N} -2 \left[y_i - \sigma(z_i) \right] \sigma(z_i) \left(1 - \sigma(z_i)\right) \times x_1$$

مخفی درون z

$$\frac{\delta L}{\delta w_1} = \frac{2}{N} \sum_{i=1}^{N} \left[\sigma(z_i) - y_i \right] \sigma(z_i) \left(1 - \sigma(z_i)\right) x_1^{(i)}$$

$$z_i = w_2 x_2^{(i)} + w_1 x_1^{(i)} + w_0$$

دقت شود $\frac{\delta L}{\delta w_2}$:

$$\frac{\delta L}{\delta w_2} = \frac{2}{N} \sum_{i=1}^{N} \left[\sigma(z_i) - y_i \right] \sigma(z_i) \left(1 - \sigma(z_i)\right) x_2^{(i)}$$

و همچنین

$$\frac{\delta L}{\delta w_0} = \frac{2}{N} \sum_{i=1}^{N} \left[\sigma(z_i) - y_i \right] \sigma(z_i) \left(1 - \sigma(z_i)\right)$$

در مورد آپدیت هم :

$$w_{i(t+1)} = w_{i(t)} + \eta \frac{\delta L}{\delta w_i}$$

سوال ۳:
نتیجه این که همگرایی momentum از GD بیشتر است.