بنام خدا — درس یادگیری ماشین — محمدطاها نجفی — ۸۱۰۱۰۱۵۰۴

سوال اول: قسمت الف)

در این سوال با دو مفهوم کلی بهینه‌سازی الگوریتم های جست و جوی خط را بیان کنیم. حالت کلی به این فرم خواهد بود این الگوریتم ها:

$$\boxed{\alpha_t = \arg\min_{\alpha > 0} f(x^t + \alpha p^t)}$$

◦ فرمول کلی با هم ادغام کردیم.

◦ با این صورت خواهد بود که:

① مقدار دهی اولیه. اول باید یک مقدار اولیه یا یک حدس اولیه x^t (نقطه جاری) و یک جهت نزولی p^t را انتخاب کرده باشیم.

② جست و جوی خطی: اندازه گام α را در تابع $f(x^t + \alpha p^t)$ را جا گذاری کرده و سپس باید بیابیم که به حداقل برسد.

این تعریف خواسته شده $\alpha_t = \arg\min_{\alpha > 0} f(x^t + \alpha p^t)$

③ و به‌روزرسانی: باید $x^t + \alpha p^t$ را به x^{t+1} آپدیت و به‌روز کنیم.

کرمان هدف کلی این الگوریتم هدف از جست و جوی خطی یافتن بهترین مقدار مناسب به یک امتداد جهت p^t است.

• این روند را تکرار می‌کنیم به نقطه بهینه یا نزدیک تا به نقطه بهینه رسیده باشیم.

بخش ب) جهت نزولی در الگوریتم های جست و جوی خطی، جهت نزولی جهتی است که در آن مقدار تابع کاهش یابد.

به طور مثال یک جهت نزولی می‌باشد اگر $\nabla f(x^t)^T p^t < 0$ یعنی منفی باشد.

این شرط نشان می‌دهد حرکت در جهت p^t در نقطه x^t منجر به کاهش تابع در نقطه x^t خواهد شد.

بخش ج) روش‌های تعیین کردن گام‌ها در الگوریتم‌های جست و جوی خطی:

• ② حالت کلی یافتن این گام‌ها: ① جست و جوی دقیق: در این روش ما گام α را با حل یک مسئله بهینه‌سازی حل می‌کنیم.

◦ این مسئله را با حل کنیم.

$$\boxed{\alpha^t = \arg\min_{\alpha > 0} f(x^t + \alpha p^t)}$$

② جست و جوی تقریبی:

در این روش به جای حل دقیق آن باید کمی تقریب از α استفاده کنیم از قانون‌ها زیر مثال بزنیم.

① قاعده آرمیجو یا شرایط ولف

② جست و جوی خطی بازگشتی

که با یک گام بزرگ شروع شده و با یک ثابت به تدریج کاهش یابد.

سوال ⑤ بخش د) جستجوی خطی دقیق تعیین کردن عامل گام آن:

$$f(x_1,x_2)=(x_1+x_2^2)^2$$

$x^T=(0,1)$ در جهت $p^T=(-1,-1)$ که یک جهت نزولی است می‌خواهیم نشان بدهیم در راستای بسته

- ابتدا باید گرادیان تابع را بیابیم:

$$\nabla f(x_1,x_2)=\left(\frac{\partial f}{\partial x_1},\frac{\partial f}{\partial x_2}\right)=\left(2(x_1+x_2^2),\ 4x_2(x_1+x_2^2)\right)$$

$$x^T=(0,1)\rightarrow \nabla f(0,1)=\left(2(1),\ 4(1)(1)\right)=(2,4)$$

$$\nabla f(x^T)^T p=(2,4)\cdot(-1,-1)=-6$$

نتیجه منفی ماند یعنی است جهت نزولی است

یا در این جهت حرکت کنیم تابع کاهش می‌یابد

$p^t=(-1,-1)$, $x^t=(0,1)$ بدین

حال باید تابع $f(x^t+\alpha p^t)$ را نسبت به α به حداقل برسانیم

بیان نقطه‌ی جدید می‌نویسیم:

$$x^t+\alpha p^t=(0,1)+\alpha(-1,-1)=(-\alpha,1-\alpha)\rightarrow$$

- این نقطه‌ی جدید ما خواهد بود.

$$f(-\alpha,1-\alpha)=\left(-\alpha+(1-\alpha)^2\right)^2=(\alpha^2-3\alpha+1)^2$$

- حال باید این عبارت را نسبت به α به کمترین مقدار ممکن برسانیم: مشتق می‌گیریم:

$$\frac{d}{d\alpha}\left((\alpha^2-3\alpha+1)^2\right)=2(\alpha^2-3\alpha+1)(2\alpha-3)(2)=0$$

$$\boxed{\alpha=\frac{3\pm\sqrt{5}}{2}}\qquad \alpha=\frac{3}{2}$$

- پس این ③ تا α جواب کاندید خواهند بود:

$$\alpha=\frac{3}{2}\qquad \boxed{\alpha=\frac{3\pm\sqrt{5}}{2}}$$

سوال دوم: روش نیوتن

• Newtons Method:

$$J(\theta) = J(\theta_0) + (\theta-\theta_0)^T \nabla J(\theta_0) + \frac{1}{2}(\theta-\theta_0)^T \nabla^2 J(\theta_0)(\theta-\theta_0)$$

تقریب مرتبه ② بسط تیلور آن حول θ_0 خواهد شد

حال اگر فرض شود که $J(\theta)$ در این نیز مساوی است داریم که

$$\nabla J(\theta) = 0$$

$$\nabla J(\theta_0) + \nabla^2 J(\theta_0)(\theta-\theta_0) = 0 \Rightarrow \theta_1 = \theta_0 - H_J^{-1}(\theta_0)\nabla J(\theta_0)$$

بردار گرادیان است که مشتق مرتبه اول آن خواهد بود.

که ماتریس هسین یا همان مشتق دوم آن می‌باشد

حال به این جست داریم که، معایب روش نیوتن:

① محاسبه هسین‌ها بسیار سخت می‌باشد. درکل از $O(n^3)$ می‌باشد که محاسبه ماتریس هسین اگر ابعاد ماتریس زیاد باشد این محاسبه میزان زیادی از حافظه را از ما می‌گیرد و محاسبه آن زمان

② غیرقابل معکوس بودن هسین ماتریس: در روش نیوتن بیان شده که باید هسین معکوس‌پذیر باشد (برخواهد بود) که تقریب Posetive Definit هسین باشد تا تقریب خوبی از آن داشته باشیم. نبوده و نمی‌توان آن را معکوس کرد. و معکوس آن غیرقابل محاسبه خواهد بود.

و معایب دیگر این است که روش نیوتن به minimum های محلی همگرا می‌شود و لزوماً Global minimum را پیدا نمی‌کند.

و این که ذخیره‌سازی معکوس هسین هم حافظه زیادی را از ما می‌گیرد.

حال چند روش هم برای بهبود روش نیوتن ارائه شده است که در اینجا به آن‌ها اشاره خواهیم کرد:

① اولین روش روش‌های Quasi_Newton بود که تقریبی از روش نیوتن اصلی را ارائه می‌دادند

② دومین روش گرادیان مزدوج بود که در این روش‌ها تقریبی از روش نیوتن را بدست می‌آوریم.

③ و کلاس این دو روش اصلی به عنوان روش‌های جایگزین برای روش نیوتن اصلی بیان شده بود.

ادامه سوال ② : حال این دو روش را بررسی میکنیم (روش شبه نیوتن Quasy_Newton) طبق قسمت (ب) هم باشند که BFGS و DFP را الان توضیح میدهیم:

روش های شبه نیوتن:

- Quasi-Newton Method: $\theta_{i+1} = \theta_i - \alpha_i S_i g_i$

$\nabla J(\theta_i)$ گرادیان آن بود.

Step-size

مانند نیوتن ما ضرب دایرانی ماتریس هسین بود

- با یک تقریب اولیه آن شروع میکنیم. $S_0 = I$

BFGS $\xrightarrow{\text{روش}}$ $S_{i+1} = S_i + \alpha_i$

$$S_{i+1} = \left(I - \frac{\Delta\theta_i \Delta\theta_i^T}{\Delta\theta_i^T \Delta g_i}\right) S_i \left(I - \frac{\Delta\theta_i \Delta g_i^T}{\Delta\theta_i^T \Delta g_i}\right)^T + \frac{\Delta\theta_i \Delta\theta_i^T}{\Delta\theta_i \Delta\theta_i}$$

برای روش اثبات شده که اگر $S_i > 0$ یا همان Positive Definite باشد آنگاه S_{i+1} هم خواهد بود. و دیگر این روش همگرایی آن ها تضمین شده بود. مشکلات بالایی در تابع ما تقریب Quadratic خوبی دارند. همان ماتریس هسین تضمین شده باشد $S_\infty = H_f^{-1}(\theta^*)$ خواهد شد (در نزدیکی بالایی).

- یا به گونه ای که بیان تر باشد میتوان تنها نوشت:

$$H_{k+1} = H_k + \frac{g_k g_k^T}{g_k^T S_k} - \frac{H_k S_k S_k^T H_k}{S_k^T H_k S_k} \qquad x_{k+1} = x_k - H_k \nabla f(x_k)$$

حال بررسی کنیم که کدام از ترم ها در این فرمول چه معنی دارند:

$\left(\frac{g_k g_k^T}{g_k^T S_k}\right)$ ← این عبارت یک تصویر از ماتریس هسین خواهد بود که تغییرات گرادیان را در ترکیب با موقعیت ها مدل میکند. به نوعی ماتریس همبستگی میان تغییرات گرادیان و تغییرات X میسازیم.

$\left(\frac{H_k S_k S_k^{-1} H_k}{S_k^T H_k S_k}\right)$ ← این بخش هم یک اصلاحیه خواهد بود بر روی ماتریس هسین H_k و تغییرات موقعیت ها که S_k را باشد. لغت ها خواهد پذیرفت که این جا به نوعی داریم به روز رسانی ها را اصلاح میکنیم.

حال به روش DFP بپردازیم:

$$B_{k+1} = B_k + \frac{S_k S_k^T}{S_k^T y_k} - \frac{B_k y_k y_k^T B_k}{y_k^T B_k y_k} \Longrightarrow \begin{array}{l} S_k = x_{k+1} - x_k \\ y_k = \nabla f(x_{k+1}) - \nabla f(x_k) \end{array}$$

تغییرات مشتق‌ها

حال به بررسی هر یک از اینها بپردازیم.

این بخش مانند روش BFGS یک اصلاحیه‌ای از B_k ها است که به نوعی ماتریس را تنظیم می‌کند تا دقت تخمین ماتریس افزایش داده باشد. اگر این باعث می‌شود دقت ماتریس در تخمین بالاتر رفته باشد.

این بخش به روزرسانی اولیه‌ای از B_k با شکل تغییرات موقت مشتق‌ها و گرادیان‌ها را برای بهبود B_k استفاده می‌کند. این بخش نسبت بهبود تخمین اصلی را دارد.

به‌طور کلی تست اول یک به روزرسانی B_k را انجام می‌دهد که پایه‌ای برای محاسبه مقدار B_{k+1} است و دیگری ادامه منفی کل آن تخمین را دقیق‌تر می‌کند و تنظیم ماتریس‌ها دقیق‌تر خواهد شد. (این روش در $O(n^2)$ موجب پیچیدگی محاسباتی می‌گردد)

حال به روش گرادیان مزدوج می‌پردازیم که یکی از روش‌های تخمین است. اگر از روش نیوتن و شبه نیوتن خواهیم که $O(n^2)$ باشد و دقت آن بیشتر از روش‌های نیوتن و شبه نیوتن است به جای محاسبه ماتریس iteration ها بلکه تنها به این نیاز خواهد داشت که

$$\theta_{i+1} = \theta_i - \alpha_i P_i \Rightarrow P_i = g_i - \beta_i P_{i-1}$$

در به‌روزرسانی β داریم

$$g_i = \nabla J(\theta)|_{\theta=\theta_i}$$

$$\beta_i = \frac{g_i^T g_i}{g_{i-1}^T g_{i-1}} = \frac{\|g_i\|_2^2}{\|\nabla g_{i-1}\|_2^2}$$

در روش‌های دیگری داریم که برای بهینه nonlinear least square کار می‌کند در روش‌های گاوس نیوتن و روش levenberg-Marquardt معروف:

ادامه سوال ② : روش گاوس نیوتن این گونه است : ابتدا تابع را که به صورت جمع مربعات خطاها باشد :

$$J(\theta) = \sum_{i=1}^{N} f^2(i,\theta) \Rightarrow \frac{\partial J}{\partial \theta} = \nabla J(\theta) = g = \left(\frac{\partial J}{\partial \theta_1}, \dots, \frac{\partial J}{\partial \theta_n}\right)^T$$

گرادیان

$$= 2\sum_{i=1}^{N} f(i,\theta)\frac{\partial f(i,\theta)}{\partial \theta_j} \qquad j = 1, 2, \dots, n$$

$$\Rightarrow \frac{\partial J}{\partial \theta} = 2J^T f$$

هسین

$$H J(\theta) = \frac{\partial}{\partial \theta_i}\left(2\sum_{i=1}^{N} f(i)\frac{\partial f(i,\theta)}{\partial \theta_j}\right) = 2\sum_{i=1}^{N} \frac{\partial f(i)}{\partial \theta_i}\frac{\partial f(i)}{\partial \theta_j}$$

$$\cancel{+2\sum_{i=1}^{N} f(i)\frac{\partial^2 f(i)}{\partial \theta_i \partial \theta_j}}$$

از این ترم صرف نظر کرده بودیم

در اطراف نقاط بهینه : ~~و نزدیک~~

$$\Rightarrow H(J(\theta)) = 2J^T J + 2\sum_{i=1}^{N} f(i)\nabla^2_\theta f(i) \simeq 2J^T J$$

$$\Rightarrow \theta_{i+1} = \theta_i - \alpha_i H_J^{-1}(\theta_i)\nabla J(\theta_i) = (2J^T J)^{-1} 2J^T f$$

$$\Rightarrow \theta_{i+1} = \theta_i - \alpha_i (J_i^T J_i)^{-1} J_i^T f_i \rightarrow \text{Gauss - Newton Method}$$

Levenberg-Marquardt Method $\Rightarrow$ Regularized version of Gauss-Newton Method

$$\left(\theta_i = \theta_i - \alpha_i (J_i^T J_i + \gamma_i I) J_i^T f_i\right)$$

که به این صورت بیان شده بود .

° سوال سوم: بیان ماتریس مربعی که میتوان به شکل تجزیه ماتریس به مقادیر ویژه آن وجود دارد.

قسمت الف)

$$\underline{A = V \Lambda V^{-1}} \longrightarrow$$ به این صورت می‌توان آن را تجزیه کرد.

V ماتریس بردارهای ویژه است که هر ستون آن یک بردار ویژه است.

V_i: مربوط به مقادیر ویژه λ_i هاست.

Λ یک ماتریس قطری است که عناصر قطر آن مقادیر ویژه خواهند بود.

$$V = \begin{bmatrix} \begin{pmatrix} v_1 \end{pmatrix} & \begin{pmatrix} v_2 \end{pmatrix} & \begin{pmatrix} v_3 \end{pmatrix} & \cdots & \begin{pmatrix} v_n \end{pmatrix} \end{bmatrix}$$

n بعد

$$\Lambda = \begin{pmatrix} \lambda_1 & & \text{صفر} \\ & \lambda_2 & \\ \text{صفر} & & \ddots \\ & & & \lambda_n \end{pmatrix}$$

° سپس صرفاً با محاسبه مقادیر ویژه را می‌یابیم و سپس از طریق آن معادله‌های ویژه را بدست می‌آوریم: به این صورت که

$$\underline{\det(A - \lambda I) = 0}$$

حال آن یک معادله چند جمله‌ای به نام (Characteristic Equation) است که با حل آن معادله λ_i ها بدست خواهند آمد.

این را باید حل کنیم تا λ ها بدست آیند.

با مثال زیر حال بیشتر آشنا خواهیم شد:

$$A = \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix} \quad \det(A - \lambda I) = \begin{pmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{pmatrix} = (4-\lambda)(3-\lambda) - 2 = 0$$

$$\Rightarrow \lambda^2 - 7\lambda + 10 = 0 \Rightarrow \lambda = \frac{7 \pm \sqrt{49-40}}{2} = \frac{7 \pm 3}{2} \rightarrow \boxed{\begin{array}{l}\lambda_1 = 5 \\ \lambda_2 = 2\end{array}}$$

پس مقادیر ماتریس Λ برابر با 5 و 2 خواهند بود.

° حال بردارهای ویژه را بدست می‌آوریم.

$$\Rightarrow \lambda_1 = 5$$

$$\Rightarrow A - \lambda_1 I = \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix} - 5\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix}$$

$$\Rightarrow \underline{(A - \lambda_1 I)V = 0} \Rightarrow \begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0 \quad -v_1 + v_2 + 2v_1 - 2v_2 = 0$$

$$v_1 - v_2 = 0 \Rightarrow v_1 = v_2 \qquad V = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

حال این دستگاه را حل کنیم.

که مشترک یک بردار ویژه را انتخاب کنیم.

$$\Rightarrow \lambda_2 = 2$$

$$\Rightarrow A - \lambda_2 I = \begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix} - 2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}$$

° حل معادله: $(A - \lambda_2 I)V = 0$

$$\Rightarrow \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0 \Rightarrow \begin{array}{l} 2v_1 + v_2 \\ 2v_1 + v_2 = 0 \end{array}$$

$$\Rightarrow 4v_1 + 2v_2 = 0 \Rightarrow 2v_1 + v_2 = 0 \quad (2v_1 = -v_2)$$

$$v_1 = 1, \quad v_2 = -2 \quad \rightarrow \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

$$\Rightarrow \underline{A} = \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}}_{V} \begin{pmatrix} 5 & \\ & 2 \end{pmatrix} \underbrace{\frac{1}{3}\begin{pmatrix} +2 & -1 \\ -1 & +1 \end{pmatrix}}_{V^{-1}}$$

° به این صورت به بردارهای ویژه آن رسیدیم.

Scanned with CamScanner

قسمت ب) اینجا که برخی از روش‌های نیوتن تصحیح‌شده از تجزیه مقادیر ویژه‌ی ماتریس هسین استفاده می‌کنند.

ادامه سوال ③: بعد از تجزیه مقادیر ویژه ماتریس هسین؟ ابتدا تقریب زدن ماتریس اصلی استفاده می‌کنیم:

ماتریس هسین

حال چون باید به علامت تجزیه مقادیر ویژه H برای تقریب زدن آن استفاده می‌کنیم:

$$H(x) = V\Lambda V^{-1} = \underline{V\Lambda V^{-1}}$$

اگر این کوچک یک مقدار کوچک (ϵ) را به Λ اضافه کرده باشیم →

$$\hat{\Lambda} = \epsilon I + \Lambda$$

حالا این ماتریس جدید را در ماتریس هسین اعمال کنیم: $\hat{H}(x) = V\hat{\Lambda}V^{-1}$ مثبت معین بودن باشد.

بعد این روش این را ثابت خواهد کرد که ماتریس $\hat{\Lambda}^{-1}$ مثبت معین بود باشد.

برای حرکت با $\hat{H}(x_k)^{-1}$ →

$$x_{k+1} = x_k - \hat{H}(x_k)^{-1}\nabla f(x_k)$$

به روز رسانی کرده‌ایم که فقط $\hat{\Lambda}$ را اصلاح کرده‌ایم.

حال قسمت ج)

$$f(x_1,x_2) = 4x_1^3 + 3x_1x_2 + 5x_2^2 + 2x_1^2x_2$$

و نقطه اولیه را $x^{(0)} = \begin{pmatrix}1\\0\end{pmatrix}$ خواهیم بود:

گرادیان f در x:

$$\nabla f(x) = \begin{pmatrix}\frac{\partial f}{\partial x_1}\\ \frac{\partial f}{\partial x_2}\end{pmatrix}$$

$$\frac{\partial f}{\partial x_1} = 12x_1^2 + 3x_2 + 4x_1x_2$$

$$\frac{\partial f}{\partial x_2} = 3x_1 + 10x_2 + 2x_1^2$$

$$x^{0} = (1,0) \Rightarrow \nabla f(x^{(0)}) = \begin{pmatrix}12\\5\end{pmatrix}$$

ماتریس هسین که حاصل مشتق دوم است

$$\Rightarrow H(x) = \begin{pmatrix}\frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2}\\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2}\end{pmatrix}$$

$$\frac{\partial^2 f}{\partial x_1^2} = 24x_1 + 4x_2 \quad / \quad \frac{\partial^2 f}{\partial x_1 \partial x_2} = 3 + 4x_1$$

$$\frac{\partial^2 f}{\partial x_2 \partial x_1} = 3 + 4x_1 \quad / \quad \frac{\partial^2 f}{\partial x_2^2} = 10$$

$$\Rightarrow H(x) = \begin{pmatrix}24x_1 + 4x_2 & 3+4x_1\\ 3 & 10\end{pmatrix} = \begin{pmatrix}24 & 7\\ 7 & 10\end{pmatrix} = \begin{pmatrix}24 & 7\\ 7 & 10\end{pmatrix}$$

حال قبل از اطمینان از این که مثبت معین است یا نه از تجزیه مقادیر ویژه آن استفاده خواهیم کرد:

$$\det(H-\lambda I) = \begin{pmatrix} 24-\lambda & 7 \\ 7 & 10-\lambda \end{pmatrix} = (24-\lambda)(10-\lambda) - 49 = 240 - 34\lambda + \lambda^2 - 49$$

$$\to \lambda^2 - 34\lambda + 191 = 0$$

$$\Rightarrow \lambda = \frac{34 \pm \sqrt{392}}{2} = \frac{34 \pm 19{,}798}{2} \to \lambda_1 = 26{,}8995, \quad \lambda_2 = 7{,}1005$$

حال چون که هر دو مقدار ویژه مثبت هستند نیازی به اصلاح ندارند

حال گام نیوتن را به روز رسانی کنیم

$$x_{k+1} = x_k - \underbrace{H(x_k)^{-1}}\ \underbrace{\nabla f(x_k)}_{\begin{pmatrix} 12 \\ 5 \end{pmatrix}}$$

$$H = \begin{pmatrix} 24 & 7 \\ 7 & 10 \end{pmatrix} \to H^{-1} = \frac{1}{191}\begin{pmatrix} 10 & -7 \\ 7 & 24 \end{pmatrix}$$

$$\to H^{-1}\nabla f(x^{(0)}) = \frac{1}{191}\begin{pmatrix} 10 & -7 \\ -7 & 24 \end{pmatrix}\begin{pmatrix} 12 \\ 5 \end{pmatrix} = \frac{1}{191}\begin{pmatrix} 85 \\ 36 \end{pmatrix} = \begin{pmatrix} 0{,}4453 \\ 0{,}1885 \end{pmatrix}$$

گام نیوتن خواهد بود

$$\Rightarrow x_1 = x_0 - H^{-1}\nabla f(x_0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0{,}4455 \\ 0{,}1885 \end{pmatrix} = \begin{pmatrix} 0{,}5545 \\ -0{,}1885 \end{pmatrix} \to x_1$$

پس از گام نیوتن خواهیم داشت

بعد از جابجایی گام را دوباره با مقادیر جدید تکرار کنیم.