هوش محاسباتی - تمرین یک

1.

- $T(x_1, x_2) = (1 + x_1, x_2)$

$T((x_1,y_1),(x_2,y_2)) = T(x_1+x_2, y_1+y_2) = (1+x_1+x_2, y_1+y_2)$

$T(x_1,y_1)+T(x_2,y_2) = (1+x_1, y_1)+(1+x_2, y_2) = (2+x_1+x_2, y_1+y_2)$

$\Rightarrow$ جمع پذیر نیست $\rightarrow$ تغییر خطی

- $T(x_1, x_2) = (x_2, x_1)$

$T((x_1,y_1)+(x_2,y_2)) = T(x_1+x_2, y_1+y_2) = (y_1+y_2, x_1+x_2)$

$T(x_1,y_1)+T(x_2,y_2) = (y_1,x_1)+(y_2,x_2) = (y_1+y_2, x_1+x_2)$

جمع پذیر

$T(c\cdot(x,y)) = T(cx, cy) = (cy, cx)$

$c\cdot T(x,y) = c\cdot(y,x) = (cy, cx)$

تعمیم

در نتیجه خطی است

- $T(x_1, x_2) = (x_1^2, x_2)$

$T((x_1,y_1)+(x_2,y_2)) = T(x_1+x_2, y_1+y_2) = ((x_1+x_2)^2, y_1+y_2)$

$T(x_1,y_1)+T(x_2,y_2) = (x_1^2, y_1)+(x_2^2+y_2) = (x_1^2+x_2^2, y_1+y_2)$

جمع ناپذیر

غیرخطی

- $T(x_1, x_2) = (\sin x_1, x_2)$

$T((x_1,y_1)+(x_2,y_2)) = T(x_1+x_2, y_1+y_2) = (\sin(x_1+x_2), y_1+y_2)$

$T(x_1,y_1)+T(x_2,y_2) = (\sin x_1, y_1)+(\sin x_2, y_2) = (\sin x_1+\sin x_2, y_1+y_2)$

جمع ناپذیر

غیرخطی

- $T(x_1, x_2) = (x_1 - x_2, 0)$ مشابه قسمت دوم

$T((x_1,y_1)+(x_2,y_2)) = (x_1+x_2-y_1-y_2, 0) = T(x_1,y_1)+T(x_2,y_2)$

$T(cx_1, cy_1) = (cx_1 - cy_1, 0) = c(x_1-y_1, 0) = cT(x,y)$

2. ابتدا دو ماتریس را به فرم RREF تبدیل کنیم. اگر RREF آن‌ها یکسان باشد هم‌ارز سطری‌اند

RREF(A):

$$A=\begin{bmatrix}2&0&0\\a&-1&0\\b&c&3\end{bmatrix}\xrightarrow{r_1\div 2}\begin{bmatrix}1&0&0\\a&-1&0\\b&c&3\end{bmatrix}\xrightarrow[r_3-br_1]{r_2-ar_1}\begin{bmatrix}1&0&0\\0&-1&0\\0&c&3\end{bmatrix}\xrightarrow[r_3-3]{r_2\times -1\\ r_3-cr_2}\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$$

RREF(B):

$$B=\begin{bmatrix}1&1&2\\0&2&3\\0&2&3\end{bmatrix}\xrightarrow{r_2\div 2}\begin{bmatrix}1&1&2\\0&1&1.5\\0&2&3\end{bmatrix}\xrightarrow{r_3-2r_2}\begin{bmatrix}1&1&2\\0&1&1.5\\0&0&0\end{bmatrix}\xrightarrow{r_1-r_2}\begin{bmatrix}1&0&0.5\\0&1&1.5\\0&0&0\end{bmatrix}$$

$\Rightarrow RREF(A)\neq RREF(B)\Rightarrow$ هم‌ارز سطری نیستند

3.

$$A=PDP^{-1},\quad P=\begin{bmatrix}1&0&\frac{1}{2}\\0&-1&1\\1&1&0\end{bmatrix},\quad D=\begin{bmatrix}1&0&0\\0&2&0\\0&0&3\end{bmatrix}$$

$$P^{-1}=\frac{1}{\det(P)}\cdot adj(P)\Rightarrow \det(P)=1\times-1\times0+0\times1\times1+\frac{1}{2}\times0\times1-(\tfrac{1}{2}\times-1\times1)-(1\times1\times1)-(0)=\frac{1}{2}$$

$$,adj(P)=\begin{bmatrix}-1&\frac{1}{2}&\frac{1}{2}\\1&-\frac{1}{2}&\frac{1}{2}\\1&1&-1\end{bmatrix}\Rightarrow P^{-1}=\frac{1}{\frac{1}{2}}\times adj(P)=\begin{bmatrix}-2&1&1\\2&-1&-1\\2&2&-2\end{bmatrix}$$

$$\Rightarrow A=PDP^{-1}=\begin{bmatrix}1&0&\frac{1}{2}\\0&-1&1\\1&1&0\end{bmatrix}\begin{bmatrix}1&0&0\\0&2&0\\0&0&3\end{bmatrix}\begin{bmatrix}-2&1&1\\2&-1&-1\\2&2&-2\end{bmatrix}=\begin{bmatrix}1&4&-2\\2&8&-4\\2&-1&-1\end{bmatrix}$$

$$\Rightarrow A\begin{bmatrix}3\\-1\\2\end{bmatrix}=\begin{bmatrix}1&4&-2\\2&8&-4\\2&-1&-1\end{bmatrix}\begin{bmatrix}3\\-1\\2\end{bmatrix}=\begin{bmatrix}-5\\-10\\5\end{bmatrix}$$

4. Gradient Decent:

مزایا	معایب
+ پایداری بیشتر به علت استفاده از همه داده‌ها.	- چون روی همه داده‌ها کار می‌کند هزینه بالاتری دارد
+ در توابع محدب مینیمم گلوبال را پیدا می‌کند	- حافظه بیشتر نیاز دارد و در مجموعه های بزرگ کند است
+ نوسان کمتری دارد	- ممکن است در مینیمم های محلی گیر کند

Stochastic Gradient Decent (batch-size = 1)

مزایا	معایب
+ هر تکرار سریع انجام می‌شود	- نوسان زیاد در مسیر
+ مصرف حافظه کم	- همگرایی غیر قطعی: ممکن است حول نقطه بچرخد
+ می‌تواند از مینیمم های محلی فرار کند	- نیاز به تنظیم نرخ یادگیری
+ مناسب برای داده‌های حجیم و یادگیری آنلاین	- حساسیت به نویز

در روش GD ساده پارامترهای مدل بر روی تمامی داده های آموزشی بروز می‌شوند ولی در SGD به اندازه batch-size که اینجا یک در نظر گرفتیم

5. ابتدا رویکرد MSE را بررسی می‌کنیم. اولین و اصلی ترین مشکل این روش این است که ترکیب آن با سیگموید یک تابع غیر محدب است. این به این معناست که الگوریتم بهینه سازی قوی تری نیاز داریم تا در مینیمم های محلی گیر نکند. مشکل دیگری که می‌توان متصور شد این است که

MSE در رگرسیون لاجستیک ناسازگاری آماری دارد . یعنی MSE فرض می‌کند خطاها توزیع نرمال دارند درصورتی که خروجی ها 1,0 هستند و توزیع برنولی دارند.

از طرفی Maximum Likelihood ترکیبش با سیگموید مجذب است وهمچنین ازلحاظ آماری سازگاری دارد.

6. باید نقطه‌ای راپیدا کنیم که اگر حذفش کنیم، شیب خط بیشترین تغییر را بکند.

این به این معناست که هرچه نقطه ای از شیب فاصله بیشتری داشته باشد تاثیر بیشتری هم دارد. باتوجه به تجمع نقاط درنمودار، بدون محاسبه می توان فهمید که دو نقطه‌ی بالا سمت راست نمودار بیشترین فاصله را دارند. پس باهمین فرض می توانیم حدس بزنیم که نقطه سمت راست نمودار همان نقطه مدنظراست (حدودا (9.6,4.5)). برای اثبات آن کافیست شیب را درحالت عادی ودرحالت بدون این نقطه حساب کنیم وببینیم که بیشترین تغییر را دارد.

7. سوال 1: اگر نرخ یادگیری ثابت و بسیار بزرگ باشد ممکن است حول نقطه بهینه نوسان کند و اگر کوچک باشد همگرایی به کندی انجام می‌شود یا در مینیمم محلی گیر می‌کند. برای حل این مشکل می‌توانیم نرخ یادگیری را با الگوریتم‌های مختلف تغییر دهیم مانند Adam

سوال 2: این دو بردار برهم عمود هستند. مثال:

$A=[1,0]$, $B=[0,1]$ $\Rightarrow$ $A.B = 1\times0 + 0\times1 = 0$ می‌توان دید که عمود هستند

سوال 3:

نادرست در رگرسیون خطی اگر واریانس خطاها ناهسان باشد تخمین های مدل نااریب باقی می‌ماند ولی واریانس تخمین‌ها بهینه نیست و پیش بینی ها قابل اعتماد نیستند.

برای حل این مشکل می‌توان از روش هایی مانند وزن دهی یا تبدیل متغیر استفاده کرد

سوال 4: این تابع برای مسئله طبقه‌بندی بهتر است چون تفاوت بین توزیع واقعی و پیش بینی شده را اندازه می‌گیرد و به مدل کمک می‌کند تا احتمال‌های طبقات را بهتر تخمین بزند.

دلیل هاهنگی این تابع این است که برای خروجی های بین صفر و یک طراحی شده و اگر پیش بینی مدل با مقدار واقعی تفاوت داشته باشد جریمه سنگینی اعمال می‌کند.

سوال 5: کاهش بیش از حد نرخ یادگیری ممکن است باعث کند شدن همگرایی شود یا حتی در مینیمم های محلی گیر کند. همگرایی به چیزهای دیگری مانند ساختار تابع هزینه و مقدار اولیه پارامترها هم وابسته است.

8. الف) در رگرسیون خطی هدف یافتن بردار وزن های w است که خطای پیشبینی را کمینه کند.

با توجه به فرض استفاده از j داریم: $y = w_j x_j$ در نتیجه برای خطای مربعات داریم: $J = \|y - w_j x_j\|^2$

برای کمینه کردن این تابع مشتق آن نسبت به w_j را صفر می کنیم.

$$\frac{\partial J}{\partial w_j} = -2 x_j (y - w_j x_j)^T = 0 \Rightarrow x_j y^T = w_j x_j x_j^T \Rightarrow \frac{x_j y^T}{x_j x_j^T} \rightarrow \frac{x_j y}{x_j x_j^T}$$

ب) اگر ویژگی ها مستقل باشد یعنی سطرهای ماتریس X مستقل خطی هستند آنگاه ماتریس $X^T X$ یک ماتریس قطری خواهد بود زیرا ضرب داخلی سطرهای مستقل خطی صفر است. در این حالت بردار وزن بهینه w برای رگرسیون با همه ویژگی ها به صورت

$$w = (X^T X)^{-1} X^T y$$

زیراست:

که اگر $X^T X$ قطری باشد داریم: $w_j = \frac{x_j y}{x_j x_j^T}$ که همین نتیجه ای است که در الف گرفتیم.

ج) در این مدل W یک بردار وزن و w_0 بایاس است.

$Y = W^T x + w_0$

این تابع هزینه ها را کمینه می کنیم:

$J(W, w_0) = \| y - (X^T W + w_0) \|^2$

$$\frac{\partial J}{\partial W} = -2X(y - X^T W - w_0)^T = 0 \quad , \quad \frac{\partial J}{\partial w_0} = -2 \sum_{i=1}^{N} (y_i - W^T x_i - w_0) = 0$$

$$\Rightarrow w_0 = \bar{y} - W^T \bar{x} \quad , \quad \Rightarrow W = (XX^T)^{-1} X(y - w_0)$$

9. با فرض $y = a + bx$ داریم: ابتدا میانگین ها را حساب می کنیم:

$$\bar{x} = \frac{521}{8} = 65.125 \qquad \bar{y} = \frac{232}{8} = 29$$

شیب b:

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$\sum xy = 14971 \quad , \quad \sum x^2 = 34203$$

$$\Rightarrow b = \frac{8 \times 14971 - 521 \times 232}{8 \times 34203 - 501^2} = \frac{-1104}{2183} \simeq 0.506$$

عرض از مبدا a:

$$a = \bar{y} - b\bar{x} = 29 - (-0.506 \times 65.125) \simeq 61.95$$

$$\Rightarrow y = a + bx = 61.95 - 0.506x$$

10. در روش SGD با فرض اینکه تابع هزینه کل به صورت زیر تعریف شود.

$$\frac{1}{n} \sum_{i=1}^{n} J_i(\theta) = J(\theta)$$

در روش SGD به جای $\nabla J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla J_i(\theta)$ که گرادیان کل است، ما از یک ست نمونه

استفاده می‌کنیم ($\nabla J_i(\theta)$) . حالا اگر داده‌ها به صورت تصادفی انتخاب شوند داریم:

$$\nabla J(\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla J_i(\theta) = E_i[\nabla J_i(\theta)]$$

درنتیجه با اینکه در SGD در هر تکرار از بخشی از داده‌ها استفاده می‌شود و گرادیان تقریبی است ولی در بلندمدت و در میانگین، مسیر به سمت مینیمم تابع هزینه است.

11. ابتدا فرمول کلی را می‌نویسیم:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t; x_i; y_i)$$

به صورتی که θ_t پارامترهای مدل در گام t، η نرخ یادگیری و L تابع هزینه است

در این صورت تحت این شرایط SGD می‌تواند به مینیمم سراسری یا محلی همگرا باشد:

الف) اگر تابع هزینه محدب باشد: در این حالت یک مینیمم وجود دارد. اگر شرایط زیر محیا باشند:

- $L(\theta)$ محدب و مشتق‌پذیر. • گرادیان‌ها دارای واریانس محدود باشند
- نرخ یادگیری با شرایط روبرو کاهش یابد:

$$\sum_{t=1}^{\infty}\eta_t = \infty \quad , \quad \sum_{t=1}^{\infty}\eta_t^2 < \infty$$

با این شرایط $\lim_{t\to\infty}\theta_t = \theta^*$

ب) اگر تابع هزینه محدب نباشد: همگرایی لزوما به مینیمم سراسری نیست اما با همان شروط می‌توان به مینیمم محلی رسید.

NADERI

.12

$$f_y(y;a) = a y^{a-1}, \quad y \in (0,1), \quad a>0$$

$$\Rightarrow L(a; y_1, y_2, \dots, y_N) = \prod_{i=1}^{N} a y_i^{a-1} = a^N \prod_{i=1}^{N} y_i^{a-1}$$ تابع likelihood

$$\Rightarrow \ln L(a) = N \ln a + (a-1) \sum_{i=1}^{\infty} \ln y_i$$ لگاریتم آن:

$$\frac{d}{da} \ln L(a) = \frac{N}{a} + \sum_{i=1}^{N} \ln y_i = 0 \Rightarrow \frac{N}{a} = -\sum_{i=1}^{N} \ln y_i$$ مشتق نسبت به a:

$$\Rightarrow a = -\frac{N}{\sum_{i=1}^{N} \ln y_i} \Rightarrow \hat{a}_{ML} = -\left(\frac{1}{N} \sum_{i=1}^{N} \ln y_i\right)^{-1}$$