# تکلیف سری دوم

مریم سعیدمهر
شماره دانشجویی : ۹۶۲۹۳۷۳

فهرست مطالب

۱ سوال اول ۲
۱.۱ الف . . . ۲
۲.۱ ب . . . ۲
۲ سوال دوم ۳
۱.۲ الف . . . ۳
۲.۲ ب . . . ۳
۳.۲ ج . . . ۳
۳ سوال سوم ۴
۱.۳ الف . . . ۴
۲.۳ ب . . . ۴
۳.۳ ج . . . ۴
۴ سوال چهارم ۵
۵ سوال پنجم ۵
۶ سوال ششم ۶
۱.۶ الف . . . ۶
۲.۶ ب . . . ۶

این تکلیف شامل ۶ سوال بوده است که به اشتباه شماره ۳ برای شماره گذاری سوالات جا افتاده است و نتیجتاً شماره آخرین سوال در صورت تکالیف ۷ است. در این پاسخنامه این مورد را اصلاح کرده ام

# ۱ سوال اول

## ۱.۱ الف

$$
\begin{aligned}
L(p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{(1-x_i)} \\
\ell(p) &= \log p \sum_{i=1}^{n} x_i + \log(1-p) \sum_{i=1}^{n} (1-x_i) \\
\frac{\partial \ell_p(X)}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n}(1-x_i)}{1-p} \overset{\text{set}}{=} 0 \\
\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i &= p\sum_{i=1}^{n}(1-x_i) \\
p &= \frac{1}{n}\sum_{i=1}^{n} x_i \quad \blacksquare
\end{aligned}
$$

## ۲.۱ ب

در آمار، برآوردگر بیشینه‌گر احتمال پسین (Maximum a posteriori estimation) یک پارامتر، مد توزیع احتمال پسین آن پارامتر است. به بیان ریاضی، اگر داده X بر اساس توزیع $f(X|p)$با پارامتر p توزیع شده‌باشند و $g(p)$ و $f(X|p)$ به ترتیب احتمال پیشین پارامتر و درستنمایی داده را نشان دهند، برآوردگر بیشینه‌گر احتمال پسین برابر خواهد بود با:

$$\hat{p}_{MAP}(x) = \arg_p \max f(p|x) = \arg_p \max \frac{f(x|p)g(p)}{\int_v f(x|v)g(v)dv}$$

مخرج کسر بالا همواره مثبت است و نیز وابسته به p نیست، در نتیجه نقشی در بهینه سازی تابع ندارد بنابراین :

$$\hat{p}_{MAP}(x) = \arg_p \max \ f(x|p)g(p)$$

همچنین از آنجایی که log یک تابع صعودی است پس در $\arg_p \max$ گرفتن خللی ایجاد نمیکند لذا میتوان گفت :

$$\hat{p}_{MAP}(x) = \arg_p \max \ \log f(x|p) + \log g(p) = \arg_p \max \ \ell(p) + \log g(p)$$

و واضح است که در معادله فوق منظور از $g(p)$ همان تابع چگالی احتمال توزیع بتا (توزیع پیشین (p است. پس در ادامه خواهیم داشت :

$$
\begin{aligned}
\arg_p \max \ \ell(p) + \log f(p|\alpha,\beta) &= \arg_p \max \ \log p \sum_{i=1}^{n} x_i + \log(1-p)\sum_{i=1}^{n}(1-x_i) + \log\frac{1}{B(\alpha,\beta)} \\
&\quad + \log p \times (\alpha-1) + \log(1-p)\times(\beta-1) \\
&= \arg_p \max \ \log\frac{1}{B(\alpha,\beta)} + \log p \times (\sum_{i=1}^{n} x_i + \alpha - 1) \\
&\quad + \log(1-p)\times(\sum_{i=1}^{n}(1-x_i) + \beta - 1) \\
\frac{\partial(\ell(p) + \log f(p|\alpha,\beta))}{\partial p} &= \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{p} - \frac{\sum_{i=1}^{n}(1-x_i) + \beta - 1}{1-p} \overset{\text{set}}{=} 0 \\
\sum_{i=1}^{n} x_i + \alpha - 1 - p(\sum_{i=1}^{n} x_i + \alpha - 1) &= p(\sum_{i=1}^{n}(1-x_i) + \beta - 1) \quad \Longrightarrow \ p = \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{n + \alpha + \beta - 2} \quad \blacksquare
\end{aligned}
$$

## ۲ سوال دوم

| ID | Age | Heartbeat | Oxygen |
|---|---|---|---|
| 1 | 41 | 138 | 37.99 |
| 2 | 42 | 153 | 47.34 |
| 3 | 37 | 151 | 44.38 |
| 4 | 46 | 133 | 28.17 |

### ۱.۲ الف

- MSE

$$h_w(x) = -59.5 - 0.15 \times Age + 0.6 \times HeartBeat$$

$$MSE = \frac{1}{2m}\sum_{i=1}^{m}(h_w(x^i) - y^i)^2$$

$$MSE = \frac{1}{2\times 4}\Big((17.1499 - 37.99)^2 + (26 - 47.34)^2 + (25.5499 - 44.38)^2 + (13.3999 - 28.17)^2\Big)$$

$$MSE = 182.80287500000014$$

- MAE

$$h_w(x) = -59.5 - 0.15 \times Age + 0.6 \times HeartBeat$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|h_w(x^i) - y^i|$$

$$MAE = \frac{1}{4}\Big(|17.1499 - 37.99| + |26 - 47.34| + |25.5499 - 44.38| + |13.3999 - 28.17|\Big)$$

$$MAE = 18.945000000000007$$

### ۲.۲ ب

به روزرسانی وزن ها در روش SGD به صورت زیر خواهد بود (منظور از i یک داده به صورت رندوم از بین داده های موجود انتخاب شده است میباشد.):

$$w_j = w_j - \alpha \times (\hat{y^i} - y^i)x_j^i \quad j = 0,1,2, \quad x_0^i = 1$$

در ادامه وزن ها در هر مرحله به شکل زیر به دست آمد :

$epoch\ 1\ \ for[41,138]:\quad w_0 = -53.617 \qquad w_1 = 241.053 \qquad w_2 = 812.454$

$epoch\ 2\ \ for[46,133]:\quad w_0 = -11957.103300000004 \quad w_1 = -547319.3168 \quad w_2 = -1582351.2239$

به دلیل اینکه اعداد به شدت بزرگ میشوند ، به نوشتن دو اپیاک بسنده کردم! (با اجازه TA در گروه تلگرام)

$$h_w(x) = -11957.103300000004 - 547319.3168 \times Age - 1582351.2239 \times HeartBeat$$

$$MSE = \frac{1}{2m}\sum_{i=1}^{m}(h_w(x^i) - y^i)^2$$

$$MSE = \frac{1}{2\times 4}\Big((-2.40816518e+08 - 37.99)^2 + (-2.65099106e+08 - 47.34)^2$$
$$+ (-2.59197807e+08 - 44.38)^2 + (-2.35641358e+08 - 28.17)^2\Big)$$

$$MSE = 3.1372570452391256e+16$$

### ۳.۲ ج

به روزرسانی وزن ها در روش GD به صورت زیر خواهد بود :

$$w_j = w_j - \alpha \times \frac{1}{4}\sum_{i=1}^{4}(\hat{y^i} - y^i)x_j^i \quad j = 0,1,2, \quad x_0^i = 1$$

در ادامه وزن ها در یک مرحله به روزرسانی به شکل زیر به دست آمد :

$$epoch\ 1: \qquad w_0 = -57.6055 \qquad w_1 = 78.02125 \qquad w_2 = 274.317$$

$$h_w(x) = -57.6055 + 78.02125 \times Age + 274.317 \times HeartBeat$$

$$MSE = \frac{1}{2m}\sum_{i=1}^{m}(h_w(x^i) - y^i)^2$$

$$MSE = \frac{1}{2\times 4}\Big((40997.01175 - 37.99)^2 + (45189.788 - 47.34)^2 + (44251.04775 - 44.38)^2 + (40015.533 - 28.17)^2\Big)$$

$$MSE = 908587593.4252931$$

با توجه به اینکه نرخ یادگیری عدد بزرگی هست ، در این مثال اصلا همگرایی رخ نمی دهد و به همین دلیل است که مقدار MSE افزایش پیدا کرده است.

## ۳ سوال سوم

### ۱.۳ الف

تابع در نقاط $x = \sqrt{2}$ و $x = -\sqrt{2}$ مشتق پذیر نیست لذا در روش subgradient میتوان برای مشتق این نقاط به ترتیب مقدار مشتق چپ و مشتق راست را قرار دهیم. به این ترتیب داریم :

$$f'(\sqrt{2}) = \sqrt{2}$$
$$f'(-\sqrt{2}) = -\sqrt{2}$$

### ۲.۳ ب

در مورد تابع فوق subgradient به شکل زیر است :

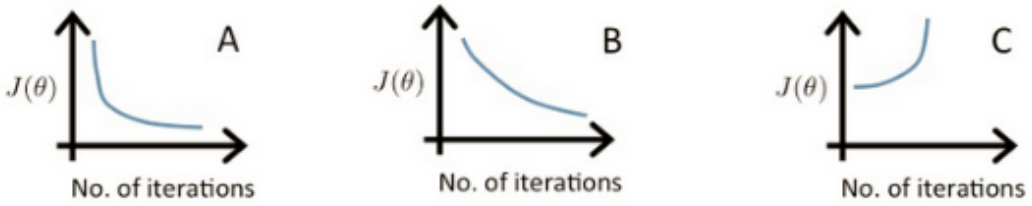
$$f\prime(x) = \begin{cases} \sqrt{2} & \text{if} \quad x \geq \sqrt{2} \\ x & \text{if} \quad -\sqrt{2} < x < \sqrt{2} \\ -\sqrt{2} & \text{if} \quad x \leq -\sqrt{2} \end{cases}$$

به این ترتیب مقدار subgradient در نقاط $(\sqrt{2}, +\infty)$ برابر مقدار ثابت $\sqrt{2}$ است و همچنین در نقاط $(-\infty, -\sqrt{2})$ برابر مقدار ثابت $-\sqrt{2}$ است.به صورت مثال در نقاط $[\sqrt{2}, +\infty[$ مقدار subgradient را از بین مقادیر $[\sqrt{2}, +\infty[$ انتخاب کرده ام و مقدار یکتایی نیست! در مورد نقاط $-\infty, -\sqrt{2}$ نیز مقدار subgradient را بین مقادیر $[-\infty, -\sqrt{2}]$ انتخاب کرده ام و مقدار یکتایی نیست!(به طور کلی در نقاط مشتق ناپذیر ، مقدار subgradient را از بازه ای بین مشتق چپ و مشتق راست انتخاب میکنند.) ولی در نقاطی که تابع مشتق پذیر بوده ، مقدار subgradient یکتا بوده است.

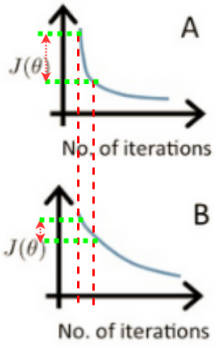
### ۳.۳ ج

در روش MAE در نقاطی که مقدار پیش بینی شده توسط مدل با مقدار واقعی داده در دیتاست مساوی باشد تابع هزینه دچار مشتق ناپذیری است زیرا مقدار مشتق چپ و مشتق راست برابر نیستند و لذا باید از روش subgradient استفاده شود.
اما به طور کلی هر روشی که در برخی نقاط مشتق پذیر نباشد موجب اشکال در روند اجرای الگوریتم های GD میشود و لذا در مورد آنها باید از subgradient استفاده کرد.

## ۴ سوال چهارم

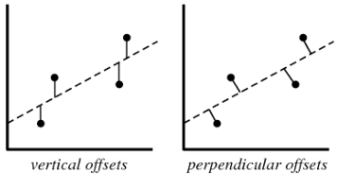



در شکل C ، تابع هزینه در هر ایپاک ، به جای نزول ، صعود کرده است و اصلا به سمت نقطه ی کمینه مطلوب ، همگرا نخواهد شد بلکه روند واگرا شدن را طی کرده و این امر به دلیل بزرگ بودن نرخ یادگیری است.
در شکل های B و A از آنجایی که روند تابع هزینه در هر ایپاک نزولی است میتوان نتیجه گرفت ، نرخ یادگیری شان ، به حد کافی کوچک است. اما اینکه کدام یک نرخ یادگیری کوچک تری دارد را در ادامه بررسی خواهم کرد.




در شکل B ، نرخ یادگیری بسیار بسیار کوچک است زیرا میزان کاهش تابع هزینه در دو ایپاک متوالی ، بسیار ناچیز است و این موضوع به دلیل بسیار کوچک بودن نرخ یادگیری است که در هر مرحله به روزرسانی ضرایب ، ضرایب به مقدار کمی تغییر (بهبود) پیدا کرده اند.
در شکل A ، در ایپاک های اولیه ، میزان کاهش تابع هزینه در دو ایپاک متوالی زیاد است و این یعنی به نسبت شکل B نرخ یادگیری بزرگتر بوده است.
جمع بندی : $\alpha_B < \alpha_A < \alpha_C$

## ۵ سوال پنجم




در مسئله برازش خطی ، در محاسبه تابع هزینه مقدار $h_\theta(x) - y$ به ازای هر نقطه محاسبه میشود. در این عبارت منظور از $h_\theta(x)$ همان تابع Hypothesis است که با توجه به وزن ها و مقدار بایاس و مقادیر x (فیچرها) برازشی انجام میدهد و مقدار y نظیر هر داده را پیش بینی میکند. در نهایت در مسئله برازش خطی فرض شده است که مقدار x نویز یا ارور ندارند و این نکته به منزله استفاده از فاصله vertical offsets است که مقدار x ثابت بوده و تنها ارور مقدار پیش بینی شده و مقدار واقعی محاسبه و استفاده میشود.

## ۶ سوال ششم

### ۱.۶ الف

در مورد دیتاستی که داده های پرت یا outlier دارد استفاده از تابع MAE به عنوان تابع هزینه ، مناسب تر است زیرا داده های outlier از عمومیت داده های دیتاست خیلی متفاوت هستند و عمدتا به دلیل اشتباهات کوچک این داده ها تولید شده اند(مثلا در فیچر قیمت خانه در یک دیتاست تمام داده ها به تومان هستند ولی قیمت خانه برای یک دیتا برحسب ریال ذخیره شده است) و آنچه اهمیت دارد این است که این داده های پرت ، خطای زیادی به مدل تزریق میکنند و نتیجه مخربی روی مدل خواهند داشت باید سعی کنیم تا حد ممکن تاثیر خطای این داده ها روی مدل را به حداقل برسانیم. در تابع MSE از آنجایی که میزان ارور به توان دو میرسد ، اصلا برای این مدل دیتاست ها مناسب نیست زیرا این خطای زیاد را به توان دو میرساند و میزان MSE به یکباره خیلی خیلی زیاد میشود. به همین دلیل توصیه میشود از MAE استفاده کنیم که حداقل همان میزان خطای واقعی این داده های پرت را به تابع هزینه تزریق کند نه مجذور شده آن را !!!

### ۲.۶ ب

تابع $MAE = \frac{1}{m}\sum_{i=1}^{m}|h_\theta(x^i) - y^i|$ تمام داده هایی که مقدار پیش بینی شده توسط مدل با مقدار واقعیشان یکسان باشد ، دچار مشتق ناپذیری میشود و این مدل را با مشکل روبه رو میکند(میدانیم هنگام به روزرسانی وزن ها نیازمند استفاده از گرادیان تابع هزینه هستیم.)(البته میتوان از روش های subgradient استفاده کرد برای حل مشکل، ولی این به هرحال به پردازش های ما سربار اضافه خواهد کرد) نتیجتاً بی خطرتر است اگر از تابع MSE به عنوان تابع هزینه استفاده کنیم.