ML تمرین ۱ 99109393 علیرضا حسیب زاده

1.1.

فرض کنید تا به حال k بار بردار w ما اصلاح شده. آن را w^k می‌نامیم.

فرض کنید اصلاح بعدی روی نقطه‌ی i انجام شود:

$$w^{k+1} = w^k + y^i x^i$$

نشان می‌دهیم که این بردار حتماً به w^* "نزدیک‌تر" شده:

$$w^{k+1} \cdot w^* = w^k \cdot w^* + \underbrace{y^i x^i \cdot w^*}_{\geq \gamma \text{ طبق فرض سوال}}$$

$$\geq w^k \cdot w^* + \gamma$$

$$f_k := w^k \cdot w^*$$

$$f_{k+1} \geq f_k + \gamma$$

از $w^0 = 0$ شروع کردیم پس:

$$f_0 = 0 \cdot w^* = 0$$

$$\rightarrow f_k \geq k\gamma$$

اما این ضرب نمی‌تواند همین‌طور بزرگ و بزرگتر شود. چرا که:

$$f_k = w^k \cdot w^* \leq |w^k| \, . \, \cancelto{1}{|w^*|} = |w^k|$$

از آنجایی که بردارهای x^i محدود هستند (داخل کره‌ی r) اندازه‌ی W^K محدود است:

$$w^{k+1} = w^k + y^i x^i$$

$$|w^{k+1}|^2 = |w^k|^2 + \underbrace{{y^i}^2}_{=1} |x^i|^2 + 2 y^i \times w^k . x^i$$

$$y \in \{0, 1\}$$

با فرض این که از نادرست بودن حدس استفاده می‌کنیم. اینجا آنجاست!
حدس $sign(w^k x^i)$ حدس مدل است و می‌دانیم این حدس اشتباه بوده که نیاز به آپدیت رخ داده پس با y^i علامت مخالف دارد و

$$\left\{\begin{array}{l} 2 y^i w^k x^i \leq 0 \\ \text{و} \; |x^i|^2 \leq r^2 \end{array}\right. \Longrightarrow |w^{k+1}|^2 \leq |w^k|^2 + r^2$$

$$g_k := |w^k|^2$$

$$\left\{\begin{array}{l} g_{k+1} \leq g_k + r^2 \\ g_0 = 0 \end{array}\right. \Longrightarrow g_k \leq k r^2$$

$$k\gamma \leq f_k = w^k \cdot w^* \leq |w^k|$$

$$g_k = |w^k|^2 \leq k r^2$$

$$\Longrightarrow \quad k\gamma \leq \sqrt{k}\, r \quad \Longrightarrow \quad k \leq \left(\frac{r}{\gamma}\right)^2 \quad \square$$

پس با شروع از $W = 0$ تعداد اصلاحات از $\left(\frac{r}{\gamma}\right)^2$ کمتر است چون K

نامحدود نمی‌شود.

1.2.

1) $$J(w) = \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

در حد $W_i \to \pm\infty$ ، $J \to +\infty$ میل می‌کند پس کمینه نقطه‌ی اپتیمم ایجاد تقعر موضعی است.

$$|y\rangle_i := y_i$$

$$X_i := \langle x_i| \qquad X = \begin{bmatrix} \langle x_1| \\ \langle x_2| \\ \vdots \\ \langle x_n| \end{bmatrix}$$

$$J(w) = \| \, |y\rangle - X|w\rangle \|^2$$

هر مولفه‌ی بردار یک جمله‌ی مجموع است.

$$J(w) = (\langle y| - \langle w|X^T)(|y\rangle - X|w\rangle)$$

$$= \langle y|y\rangle + \langle w|X^TX|w\rangle - \underline{\langle w|X^T|y\rangle} - \underline{\langle y|X|w\rangle}$$

$$= y^Ty + w^TX^TXw - 2w^TX^Ty \qquad (\alpha^T = \alpha \,,\ \alpha \in \mathbb{R})$$

$$\frac{dJ}{dw} = \frac{d}{dw}\left(w^TX^TXw\right) - 2\frac{d}{dw}\left(w^TX^Ty\right)$$

مقادیر را با توجه به تمرین 0 جاگذاری می‌کنیم:

$$\frac{dJ}{dw} = \left[X^TX + (X^TX)^T\right]w - 2X^Ty$$

$$= 2\left[X^TXw - X^Ty\right] = 0$$

$$X^TXw = X^Ty \implies w = (X^TX)^{-1}X^Ty$$

اگر X^TX وارون پذیر باشد.

2) اولین مشکل این است که ممکن است X^TX وارون پذیر نباشد. اگر در

$X_{N\times(d+1)}$ ، $d+1 < N$ باشد حتما X^TX که $(d+1)\times(d+1)$ است

وارون پذیر است و مشکلی نداریم (پس باید به تعداد کافی داده برای learning داشته

باشیم.)

درخصوص استفاده‌ی مستقیم از رابطه $w = (x^T x)^{-1} x^T y$

تعدادی ضرب ماتریسا و وارون گرفتن از $x^T x$ که $(d+1)(d+1)$ است

داریم که هزینه‌ی محاسباتی زیادی دارد.

$$x^T x \xrightarrow{+} O((d+1) \times N \times (d+1)) = O(N(d+1)^2)$$

$$(x^T x)^{-1} \xrightarrow{+} O((d+1)^3)$$

$$x^T y \xrightarrow{+} O((d+1) \times N)$$

$$(x^T x)^{-1} x^T y \xrightarrow{+} O((d+1) \times (d+1))$$

$$O(N(d+1)^2 + (d+1)^3) = O(N(d+1)^2)$$

یک راه‌حل برای این مسئله این است که ما نهایتاً می‌خواهیم $E = \| xw - y \|^2$ را کمینه کنیم.

با حل تحلیلی می‌دانیم این cost function یک و فقط یک نقطه‌ی کمینه دارد. پس gradient descent حتماً جواب خواهد داد.

(3 $$J(w) = \sum F_i (w^T x_i - y_i)^2$$

اگر می‌دانستیم $F_i > 0$ آن را داخل $(\;)^2$ می‌بریم و جواب x_i و y_i

میکنیم . اما لزوماً این طور نیست .

$$J(w) = \langle y - Xw \,|\, J \,|\, y - Xw \rangle$$

$$J(w) = (y - Xw)^T F \, (y - Xw)$$ یا

که F یک ماتریس قطری است :

$$F := \begin{bmatrix} F_1 & & & \\ & F_2 & & \\ & & \ddots & \\ & & & F_N \end{bmatrix}$$

$$J(w) = y^T F y - \underbrace{y^T F X w}_{} - w^T X^T F y + w^T X^T F X w$$

$$= (w^T X^T F y)^T = w^T X^T F y$$

$$= y^T F y - 2 w^T X^T F y + w^T (X^T F X) w$$

$$\frac{dJ}{dw} = -2 X^T F y + \Big(X^T F X + \underbrace{(X^T F X)^T}_{X^T F X} \Big) w$$

$$= 2 \left[X^T F X w - X^T F y \right] = 0$$

$$\Longrightarrow \quad \boxed{w = [X^T F X]^{-1} X^T F y} \qquad F = \begin{bmatrix} F_1 & & \\ & F_2 & \\ & & \ddots \; F_N \end{bmatrix}$$

$$\hat{w} = \arg\min_w E_{xy}\left[(y - w^T x)^2\right]$$ 4)

$$= \arg\min_w \frac{1}{N}\sum_i (y_i - w^T x_i)^2 = \arg\min_w \sum_i (y_i - w^T x_i)^2$$

پس معادل ما

$$X^T X w = X^T y$$

$$X^T = \underbrace{\begin{bmatrix} | & & | \\ x_1 & \cdots & x_d \\ | & & | \end{bmatrix}}_{n}$$

x_i = داده‌های ویژگی i

$$X^T X w = \begin{bmatrix} x_1^T x_1 & \cdots & x_1^T x_d \\ \vdots & & \\ x_d^T x_1 & \cdots & x_d^T x_d \end{bmatrix} w = \begin{bmatrix} x_1^T y \\ \vdots \\ x_d^T y \end{bmatrix}$$

$$x_i^T x_j = \sum_{1}^{n} x_i^k x_j^k = n E_x[x x^T]_{ij} = n R_{ij}$$

$$x_i^T y = \sum_{1}^{n} x_i^k y^k = n E_{xy}[xy]_i = n C_i$$

$$\Longrightarrow \cancel{n} R w = \cancel{n} C \Longrightarrow \boxed{w = R^{-1} C}$$

$$E \;=\; \mathbb{E}_{xy}\left[\,\|X\hat{w} - y\|^2\,\right]$$

خطا

$\hat{w} :=$ بردار w به دست آمده

$\hat{w}^* :=$ بردار w ایده‌آل که از رگرسیون خطی روی کل داده‌ها به دست می‌آید

$$E = \mathbb{E}_{xy}\left[\|X(\hat{w} - \hat{w}^*) + X\hat{w}^* - y\|^2\right]$$

$$= \mathbb{E}_{xy}\|X(\hat{w} - \hat{w}^*)\|^2 + \mathbb{E}_{xy}\|X\hat{w}^* - y\|^2$$

$$+\, 2\,\mathbb{E}_{xy}\left[(\hat{w} - \hat{w}^*)^T X^T (X\hat{w}^* - y)\right]$$

$$= X^T X \hat{w}^* - X^T y$$

$\mathbb{E}$ این عبارت صفر است چون تعریف $\hat{w}^*$ بردار ایده‌آلی است که از رگرسیون روی تمامی داده‌ها (X) به دست آمده. $\left(\hat{w}^* = (X^T X)^{-1} X^T y\right)$

$$\Rightarrow\; \boxed{E = \underbrace{\mathbb{E}_{xy}\|X(\hat{w} - \hat{w}^*)\|^2} + \underbrace{\mathbb{E}_{xy}\|X\hat{w}^* - y\|^2}}$$

ناشی از اختلاف $\hat{w}$ با $\hat{w}^*$ که از نداشتن بیتای کل ناشی می‌آید

approximation error

اگر کل داده $X \to$ ، $\hat{w} \to \hat{w}^*$ و این بخش صفر می‌شود.

structural error

ناشی از کافی نبودن مدل خطی

حتی بهترین خط که روی کل دیتاست

ترین شود هم خطا دارد. (خطای آن)

1.3

۱) کافی است در پاسخ سوال قبل X را یک ماتریس $1 \times N$ قرار دهیم.
(بقیه را صفر بگیریم.)

بردار افقی — بردار عمودی

$$w_i = \overbrace{(X^T X)^{-1}}^{\text{اسکالر}} X^T y$$

$(x_i^1, x_i^2, \dots, x_i^N)$

$$(X^T X)^{-1}_{1\times 1} = \frac{1}{X^T X} = \frac{1}{\langle x_i, x_i \rangle} \quad , \quad X^T y = \langle x_i, y \rangle$$

$$\Longrightarrow \quad w_i = \frac{\langle x_i, y \rangle}{\langle x_i, x_i \rangle}$$

۲) ماتریس X را به صورت بردار ویژگی‌ها جدا می‌کنیم:

$$X = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_d \\ | & & | \end{bmatrix} \qquad x_i = \begin{bmatrix} x_i^1 \\ \vdots \\ x_i^N \end{bmatrix}$$

که هر بردار N داده‌ی هر ویژگی را در خود دارد.

$$X^T = \begin{bmatrix} — & x_1 & — \\ & \vdots & \\ — & x_d & — \end{bmatrix}$$

$$X^T X = \begin{bmatrix} — & x_1 & — \\ & \vdots & \\ — & x_d & — \end{bmatrix} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_d \\ | & & | \end{bmatrix}$$

$$= \begin{bmatrix} x_1^T x_1 & x_1^T x_2 & \cdots & x_1^T x_d \\ \vdots & & & \\ x_d^T x_1 & \cdots & & x_d^T x_d \end{bmatrix}$$

اما ستون ها بر هم عمود بودند پس : $\quad x_i^T x_j = x_i^T x_j \, \delta_{ij}$

پس این ماتریس قطری است و وارون آن به سادگی :

$$(X^T X)^{-1} = \begin{bmatrix} x_1^T x_1 & & \\ & \ddots & \\ & & x_d^T x_d \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{x_1^T x_1} & & \\ & \ddots & \\ & & \frac{1}{x_d^T x_d} \end{bmatrix}$$

$$\Longrightarrow \quad W = (X^T X)^{-1} X^T y = \begin{bmatrix} \frac{x_1^T y}{x_1^T x_1} & \frac{x_2^T y}{x_2^T x_2} & \cdots & \frac{x_d^T y}{x_d^T x_d} \end{bmatrix}^T$$

$$\Longrightarrow \quad w_i = \frac{x_i^T y}{x_i^T x_i} = \frac{\langle x_i , y \rangle}{\langle x_i , x_i \rangle} \quad \square$$

همان شکل (۱) بدست آوردیم.

3) از روابط قبلی :

$$(X^T X) W = X^T y$$

$$X = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \Rightarrow X^T = \begin{bmatrix} x_1 & \cdots & x_n \\ 1 & \cdots & 1 \end{bmatrix}$$

$$X^T X\, W = \begin{bmatrix} \sum x_i^2 & \sum x \\ \sum x & \underbrace{\sum 1}_{=n} \end{bmatrix} \begin{bmatrix} w_1 \\ w_0 \end{bmatrix} = X^T y$$

$$= \begin{bmatrix} \sum x_i y_i \\ \sum y_i \end{bmatrix}$$

$$\times \tfrac{1}{n} \Rightarrow \begin{bmatrix} E(X^2) & E(X) \\ E(X) & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_0 \end{bmatrix} = \begin{bmatrix} E(XY) \\ E(Y) \end{bmatrix}$$

$$\times \begin{bmatrix} 1 & -E(x) \\ 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} \underbrace{E(X^2) - E^2(X)}_{Var(X)} & 0 \\ E(X) & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_0 \end{bmatrix} = \begin{bmatrix} \underbrace{E(XY) - E(X)E(Y)}_{Cov(X,Y)} \\ E(Y) \end{bmatrix}$$

$$E(X^2) - E^2(X) = Var(X) \quad , \quad E(XY) - E(X)E(Y) = Cov(X,Y)$$

سطر اول : $Var(X)\, w_1 = Cov(X,Y) \implies w_1 = \frac{Cov(X,Y)}{Var(X)}$

سطر دوم : $E(X)\, W_1 + W_0 = E(Y) \implies W_0 = E(Y) - W_1 E(X)$

2.1. $L = P(X|\theta) = \prod P(x_i|\theta)$

$$= \begin{cases} \left(\frac{1}{2w}\right)^n & -w \leq x_1, x_2, \ldots, x_n \leq w \\ 0 & \text{o.w.} \end{cases}$$

باید با تنظیم w ، L را بیشینه کنیم. اگر حالت دوم رخ دهد $L=0$ میشود ولی چون داریم حالات $L>0$ هم داریم پس حالت اول باید رخ دهد :

$$-w \leq x_1, \ldots, x_n \leq w \implies w \geq |x_1|, \ldots, |x_n|$$

در این صورت $L = \left(\frac{1}{2w}\right)^n$ است و یک تابع نزولی اکید در $w>0$ است پس w باید کمترین مقدار ممکن باشد تا L بیشینه شود.

$$w = \operatorname{argmax} L = \operatorname{argmax} \left(\frac{1}{2w}\right)^n \mid w \geq |x_1|, \ldots, |x_n|$$

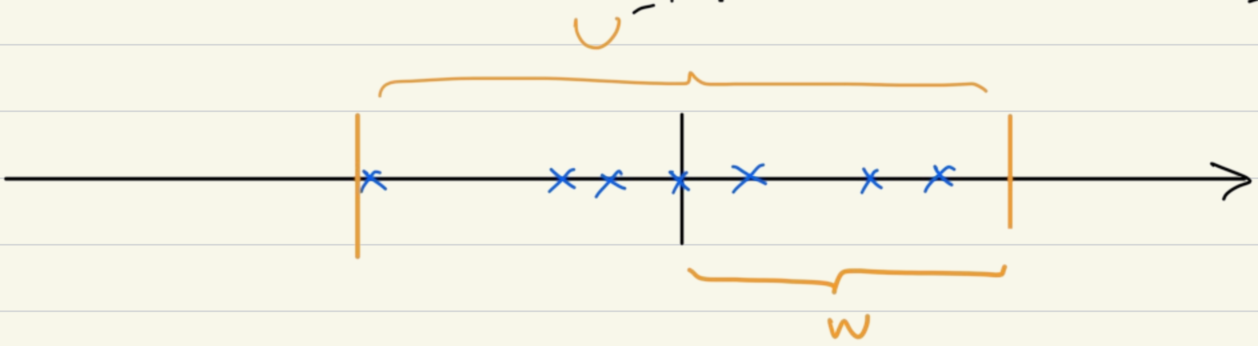
$$= \min w \mid w \geq |x_1|, \ldots, |x_n|$$

$$w = \boxed{\max\left[|x_1|, |x_2|, \ldots, |x_n|\right]}$$

لازم به ذکر است این عبارت برقرار باشد تا w از همه ی $|x_i|$ ها بزرگتر باشد و کوچکترین مقدار هم باشد که احتمال صفر نشود.

پاسخ به دست آمده شهودی هم هست. w را از صفر شروع به زیاد کردن کنید.

اولین جایی که همه‌ی X_iها را در بر بگیرد.



2.2

$$L = \prod P(x_i|\theta) = \prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[\frac{-1}{2\sigma^2} \sum_1^n (x_i-\mu)^2\right]$$

در $\mu \to \pm\infty$ ، $L \to 0$ میل می‌کند و L مثبت است پس بیشینه‌ی از

مقدار نقطه (نقاط) مشتق صفر است:

$$\frac{\partial L}{\partial \mu} = \overset{>0}{L} \times \frac{-1}{2\sigma^2} \sum_1^n 2(\mu - x_i) = 0$$

$$\Rightarrow \sum \mu = N\mu = \sum_1^N x_i \Rightarrow \hat{\mu} = \frac{\sum_1^N x_i}{N}$$

رابطه‌ی معروف منسوب به "میانگین" داده‌ها.

$$P(\mu, x) = P(x|\mu)\,P(\mu) = P(\mu|x)\,P(x) \qquad (2$$

$$\Longrightarrow P(\mu|x) = P(x|\mu)\,\frac{P(\mu)}{P(x)} \;\propto\; P(x|\mu)\,P(\mu)$$

$$\mu \sim \mathrm{Norm}(\mu_0, \beta^2) \quad , \quad x_i \sim \mathrm{Norm}(\mu, \sigma^2)$$

$$\Longrightarrow P(\mu|x) \;\propto\; \frac{1}{\sqrt{2\pi\beta^2}}\, e^{-\frac{(\mu-\mu_0)^2}{2\beta^2}} \times \prod_{1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$P(\mu|x) \;\propto\; \frac{1}{\sqrt{2\pi\beta^2}} \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{(\mu-\mu_0)^2}{2\beta^2} + \frac{-1}{2\sigma^2}\sum (x_i-\mu)^2\right]$$

به استدلال مشابه چون $\mu \to \pm\infty$ ، $P \to 0$ میل کند پس بیشینه‌ها موجود است.

$$\frac{\partial P}{\partial \mu} \;\propto\; -\underbrace{P(\mu)}_{>0}\left[\frac{\hat{\mu}-\mu_0}{\beta^2} + \frac{1}{\sigma^2}\sum(\mu - x_i)\right] = 0$$

$$\Longrightarrow \hat{\mu} = \frac{\frac{1}{\beta^2}\cdot\mu_0 + \frac{1}{\sigma^2}\sum x_i}{\frac{1}{\beta^2} + \frac{N}{\sigma^2}}$$

انگار داده‌ها با ضریب $\frac{1}{\sigma^2}$ و μ_0 را با ضریب $\frac{1}{\beta^2}$ میانگین حسابی گرفته باشیم.

در صورتی که prior توزیع هرچه دقیق‌تر μ را بدهد (β کوچکتر) سهم بیشتری

دارد و هرچه بی‌دقت‌تر باشد یا یکنواخت باشد ($\beta \to \infty$) بی سهم‌تر است و

داده‌ها سهیم می‌شوند.

3) پاسخ MLE :
$$\hat{\mu} = \frac{\sum x_i}{N}$$

باید واریانس این تخمین‌گر را حساب کنیم:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}\left[\frac{\sum x_i}{N}\right] = \frac{1}{N^2}\mathrm{Var}\left[\sum x_i\right]$$

$$\overset{iid}{\Longrightarrow} \quad = \frac{1}{N^2}\sum \mathrm{Var}(x_i) = \frac{N}{N^2}\sigma^2 = \frac{\sigma^2}{N}$$

و حال امید ریاضی تخمین‌گر:
$$\mathbb{E}(\hat{\mu}) = \mathbb{E}\left(\frac{\sum x_i}{N}\right) = \frac{N}{N}\mu = \underset{\text{واقعی}}{\mu}$$

در $N \to \infty$ ، $\mathrm{Var}(\hat{\mu}) \longrightarrow 0$ میل می‌کند و $\mathbb{E}(\hat{\mu}) = \mu$ است.

پس در این حالت تخمین‌گر حتی مقدار واقعی μ را گزارش می‌کند. (با مدرکزی هم می‌شد گفت)

پاسخ MAP :

$$\hat{\mu} = \lim_{N \to \infty} \frac{\frac{1}{\beta^2}\mu_0 + \frac{1}{\sigma^2}\sum_{1}^{N} x_i}{\frac{1}{\beta^2} + \frac{N}{\sigma^2}}$$

$$\hat{\mu} = \frac{\frac{1}{\sigma^2}\sum_{1}^{N} x_i}{\frac{N}{\sigma^2}} = \frac{\sum x_i}{N} = \hat{\mu}_{MLE} = \mu$$

درصورت جمع $\sum x_i$ به سمت بی‌نهایت میل می‌کند. (چون امید ریاضی آن
جمع $N\mu$ است و چون طبق حد مرکزی به گاوسی میل می‌کند) و در
مقابل آن می‌توان از $\frac{\beta}{\sigma^2}$ که محدود است در حد چشم پوشید.

پس در این حد $\hat{\mu}$ MAP هم به $\hat{\mu}$ MLE میل کرد که آن هم با
واریانس صفر μ واقعی را می‌دهد.

2.3

$$f(x|\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

توزیع x

$$f(\beta) = \frac{\beta'^{\alpha'}}{\Gamma(\alpha')} \beta^{\alpha'-1} e^{-\beta'\beta}$$

توزیع β، گامای x

$$f(\beta|x) \propto f(\beta) f(x|\beta)$$

$$= \frac{\beta'^{\alpha'}}{\Gamma(\alpha')} \beta^{\alpha'-1} e^{-\beta'\beta} \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

جزء α

$$\propto \beta^{\alpha+\alpha'-1} \cdot e^{\overbrace{-(\beta'+x)}^{\text{جزء } \beta}\beta}$$

$$\beta | x \sim \text{Gamma}(\alpha+\alpha', \beta'+x)$$

توزیع گاما $x|\beta$

توزیع گاما

prior خود β