علی قربانپور - ۹۶۱۰۵۷۹۴

تمرین سری سوم یادگیری ماشین.

۱) a

$$L_{MSE}(\theta) = \frac{1}{n}\sum_i (x_i-\theta)^2 \quad , \quad \arg\min L_{MSE}(\theta) = ?$$

$$L'_{MSE}(\theta) = \frac{1}{n}\times 2\times(-1)\times\sum_i (x_i-\theta) = 0 \rightarrow \left(\sum_{i=1} x_i\right) - n\theta = 0$$

$$\rightarrow \theta = \frac{1}{n}\sum_{i=1} x_i \rightarrow$$ بازای $\theta = mean(x_i)$ مقدار مشتق تابع صفر می‌شود

$$\rightarrow \arg\min L_{MSE}(\theta) = \underset{i=1}{\overset{n}{mean}}(x_i)$$

$$L_{MAE}(\theta) = \frac{1}{n}\sum_i |x_i-\theta| \quad , \quad \arg\min L_{MAE}(\theta) = ?$$

لم ۱: مقادیر a و b فرضی را در نظر بگیرید. می‌خواهیم حداقل مقدار θ را در عبارت $f=|a-\theta|+|b-\theta|$ بدست آوریم. بدون از دست رفتن کلیت مسئله فرض کنید $a \geq b$ است. حال برای θ، ۳ حالت داریم

$$b \leq a \leq \theta \rightarrow f = \theta - a + \theta - b = 2\theta - (a+b) \quad \text{I}$$

$$b \leq \theta \leq a \rightarrow f = a - \theta + \theta - b = (a-b) \quad \text{II}$$

$$\theta \leq b \leq a \rightarrow f = a - \theta + b - \theta = (a+b) - 2\theta \quad \text{III}$$

$$\text{I}, \text{II} \rightarrow a-b \leq 2\theta - a - b \rightarrow 2a \leq 2\theta \rightarrow a \leq \theta \checkmark$$

$$\text{II} \text{ و } \text{III} \rightarrow a-b \leq a+b-2\theta \rightarrow 2\theta \leq 2b \rightarrow \theta \leq b \checkmark$$

پس حالت کمینه زمانی است که θ بین a و b باشد. $\mathrm{II} \le \mathrm{III}$ و $\mathrm{II} \le \mathrm{I}$ ←

حال به مسئله‌ی اول باز می‌گردیم. مقادیر x_iها را به ترتیب صعودی مرتب می‌کنیم.

$$x_1 \le x_2 \le \dots \le x_{n-1} \le x_n$$

در ادامه جمله‌های مجموع را دو به دو از ابتدا و انتها کنار هم قرار می‌دهیم. بدین صورت:

$$\sum_i |x_i-\theta| = \overbrace{\left[|x_1-\theta|+|x_n-\theta|\right]}^{\alpha_1} + \overbrace{\left[|x_2-\theta|+|x_{n-1}-\theta|\right]}^{\alpha_2} + \dots$$
$$+ \left[|x_{\frac{n}{2}}-\theta|+|x_{\frac{n}{2}+1}-\theta|\right]$$

حال برای هر کدام از ترم‌های $\alpha_1+\alpha_2+\dots$ را هر کدام در حالت کمینه قرار دهیم. بدین ترتیب چون تمام جملات مثبت و مستقل از یکدیگر هستند پس به ایده‌آل مسئله می‌رسیم. طبق ترم اول (α_1) مقدار θ مورد نظر باید عضو بازه‌ی $[x_1, x_n]$ باشد. طبق ترم دوم (α_2) این مقدار باید عضو بازه‌ی $[x_2, x_{n-1}]$ باشد. به همین ترتیب الی آخر. پس اگر مقدار θ عددی درون آخرین بازه‌ی میانی باشد تمامی شروط قبلی برآورده شده و تمام زوج قدرمطلق‌ها در حالت کمینه قرار خواهند گرفت. در نظر داشته باشید اگر تعداد اعداد زوج باشد θ عددی در بازه‌ی آخرین است و اگر فرد باشد برابر عضو میانه خواهد بود تا بتواند جمله‌ی آن حذف شده را صفر کند. بدین ترتیب در هر ۲ حالت مقدار عددی میانه مقدار قابل قبول خواهد بود.

$$\underset{\theta}{\arg\min}\, L_{MAE}(\theta) = \underset{x_i}{\mathrm{median}}$$

ط) در مورد توابع خطای نرم یک و دو موارد قابل بحث است. نخست آن که تابع نرم دو خطاهای با اندازه بزرگ را به خوبی آشکار می‌کند زیرا با به توان ۲ رساندن آن می‌تواند آشکارساز خوبی برای این خطاها باشد. از طرفی تابع نرم یک به تمام خطاها یکسان چشم می‌نگرد و با جمع کردن قدر مطلق خطاها را محاسبه می‌کند. بدین ترتیب بین خطاهای کوچک و بزرگ تفاوت مرتبه‌ها قائل نمی‌شود.

تابع نرم ۲ یا MSE می‌تواند مقادیر داده‌های پرت را به خوبی شناسایی کند. همچنین تابع نرم یک مشتق‌پذیر نیست و برای تمامی حالات نمی‌توان از گرادیان بهره برد در حالی که تابع نرم ۲ این مشکل را ندارد.

با توجه به بخش قبل متوجه شدیم MSE مقدار میانگین و MAE مقدار میانه را به عنوان argmin عبارت loss تعریف می‌کند. حال فرض کنید ۹۰ درصد اعداد برابر ۲۰۰ و ۱۰ درصد باقی مانده اعداد بین ۵ — ۵۰ باشند. بدین ترتیب MAE مقدار ۲۰۰ را به عنوان پارامتر در نظر می‌گیرد و هیچ دیدی از اعداد دیگر نخواهد داشت. همچنین همین مورد برای داده‌های پرت در MSE نیز وجود خواهد داشت و وزن بسیار زیادی به آن‌ها می‌دهد در حالی که ممکن است تخمین خوبی زده باشیم و تنها یک داده‌ی بسیار پرت مقدار loss را خراب در نظر بگیرد.

$$\text{Huber } L_{(\theta, y)} = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} \frac{1}{2}(y_i-\theta)^2 & |y_i-\theta| \leq \alpha \\ \alpha\left(|y_i-\theta|-\frac{1}{2}\alpha\right) & \text{o.w} \end{cases}$$

همان طور که از تعریف تابع Huber پیداست ترکیبی از دو تابع MSE و MAE ارائه کرده است که برای مقادیر بزرگ خطا از MAE بهره برده تا میزان حساسیت کمتری داشته باشد و برای مقادیر اندک خطا از MSE تا از مزایای MSE بهره گیرد و هم مشکل مشتق آن برای 0 (مشتق‌ناپذیری) را نداشته باشد.

$$\text{logCosh } L(y, y^p) = \sum_{i=1}^{n} \log\left(\cosh(y_i^p - y_i)\right)$$

با توجه به تعریف تابع log Cosh می‌توان تخمین زد برای مقادیر کوچک خطا این تابع تقریباً برابر با $\frac{x^2}{2}$ و برای مقادیر بزرگ‌تر برابر با $|x| - C$ خواهد بود. بدین ترتیب مانند تابع پیشین این تابع نیز به نوعی از مزایای دو تابع MSE و MAE بهره می‌برد. همچنین این تابع با توجه به تعریف آن تا ۲ مرتبه مشتق‌پذیر است که این ویژگی مفیدی در محاسبات و بهینه‌سازی برخی پارامترها خواهد بود. ویژگی‌ای که توابع MAE و Huber ندارند.

۲) زمانی که دو valid set از دو توزیع مختلف بوده باشند. با توجه به نمودار می‌توان حدس زد که توزیع داده‌های train با داده‌های validate نزدیک بوده ولی مدل‌ها به نوعی بر روی این توزیع overfit شده‌اند. برای بهبود این مشکل می‌توان از افزودن regularization term استفاده کرد تا از overfit جلوگیری کنیم.

در مورد حالتی که داده‌های در test validation از یک نوع توزیع بوده باشند می‌توان این تصور را داشت که هر دوی توزیع‌ها با توزیع train تفاوت دارند به صورت اتفاقی یا بر اساس برخی پارامترهای توزیع باشد. بنابراین، این مدل توانسته برای یک مجموعه خوب عمل کند و برای دیگری خیر. راه حل این کار بهتر می‌توان Cross Validate کرد تا به فضای میانگین بهتری برای سنجش برسیم.

(۳)

$$E[L_q] = \iint |f(x) - y|^q \cdot p(x,y) \cdot dx \cdot dy$$

$$\xrightarrow{\frac{\partial}{\partial f}} \operatorname{sign}(f-y) \int \underbrace{q \cdot |f-y|^{q-1} p(x,y)\, dy}_{\alpha} = 0 \longrightarrow$$

$$\int_{-\infty}^{f(x)} \alpha - \int_{f(x)}^{+\infty} \alpha = 0 \xrightarrow{q \neq 0} \int_{-\infty}^{f(x)} |f-y|^{q-1} \cdot p(x,y) \cdot dy =$$

$$\int_{f}^{+\infty} |f-y|^{q-1} p(x,y) \cdot dy \longrightarrow \underline{q=1} \quad , \int_{-\infty}^{f} p(x,y) \cdot dy = \int_{f}^{+\infty} p(x,y) \cdot dy$$

$\underline{q \to 0} \longrightarrow |f-y|^q \simeq 1$ به جز در حالتی که $|f(x)-y| = 0$ باشد، حاصل

در حالی که $f(x)$ برابر است max $p(x,y)$ یعنی توزیع احتمال را داریم که بیشترین ⟵

فراوانی دارد را پیدا داریم. یعنی مد.

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} = 1 - \frac{2}{e^{2x} + 1} \; *$$

$$\text{sigmoid}(x) = \frac{e^{x}}{e^{x} + 1} = 1 - \frac{1}{e^{x} + 1} \; * \longrightarrow \underline{\tanh(x) = 2\,\text{sigmoid}(2x) - 1}$$

$$f(2x \mid w^{(s)}) = w_0^{(s)} + \sum_{i=1}^{n} w_i^{(s)} \cdot s(2x) \,,\; f(x \mid w^{(t)}) =$$

$$w_0^{(t)} + \sum_{i=1}^{n} w_i^{(t)} \cdot \underbrace{t(2x)}_{2S(2x)-1}$$

$$\Rightarrow w_0^{(s)} + \sum_{i=1}^{n} w_i^{(s)} S(2x) = w_0^{s} + \frac{1}{2}\left[\sum_{i=1}^{n} w_i^{(s)} \left(2S(2x) - 1\right) + \sum_{i=1}^{n} w_i^{(s)} \right]$$

$$\longrightarrow w_0^{(t)} = w_0^{(s)} + \frac{\sum_{i=1}^{n} w_i^{(s)}}{2} \quad \Big/ \quad w_i^{(t)} = \frac{1}{2} w_i^{(s)} \;:\; 1 \leq i$$

$\textcircled{5}$ در حالت بدون نویز: $L = \frac{1}{2}\sum_n \left(y(x_i, w) - t_n\right)^2$

noise added $\longrightarrow$ $L = \frac{1}{2}\sum_n \left(y(x_i+\varepsilon, w) - t_n\right)^2 = \frac{1}{2}\sum_n \left(\left[w_0 + \sum_i w_i x_i\right] + \sum w_i \varepsilon_i - t_n\right)^2$

$$= \frac{1}{2}\sum \Big(\underbrace{y(x_i, w) - t_n}_{a} + \underbrace{\sum w_i \varepsilon_i}_{b} \Big)^2 \quad = a^2 + b^2 + 2ab$$

$$= \underbrace{\frac{1}{2}\sum_n \left(y(x_i,w) - t_n\right)^2}_{\text{حالت بدون نویز } \checkmark} + \frac{1}{2}\Big(\underbrace{\sum w_i \varepsilon_i}\Big)^2 + \sum \underbrace{(w_i \varepsilon_i)\left(y(x_i,w) - t_n\right)}$$

$$\underbrace{E\left[\sum (w_i \varepsilon_i)\right]^2 = \frac{\sigma^2}{2} \cdot \sum_i w_i^2}_{\text{regularization term } \checkmark}$$

$$E\left[\left(y(x_i,w) - t_n\right) \cdot \sum_i (w_i \cdot \varepsilon_i)\right] = 0$$

(ی)

$$L(w) = \sum_{i=1}^{n} F_i \left(y^{(i)} - w^T . x\right)^2 \longrightarrow \min L(w) = ?$$

$$\frac{\partial L}{\partial w} = 0 \longrightarrow \sum_i F_i (y_i - w^T . x_i) x_i^T = 0 \longrightarrow \sum_i F_i y_i x_i^T = \sum w^T F_i x_i x_i^T$$

$$\longrightarrow w^T = \frac{\sum_i F_i y_i x_i^T}{\left(\sum F_i x_i x_i^T\right)} \longrightarrow w^T = \left(\sum_i F_i x_i x_i^T\right)^{-1} \left(\sum_i F_i y_i x_i^T\right)$$

$$\longrightarrow w^T = \left(\sum_i F_i \Phi_{x_i} \Phi_{x_i}^T\right)^{-1} \left(\sum_i F_i y_i \Phi^T_{(x_i)}\right)$$

$$\longrightarrow \boxed{w = \left(\Phi^T_{(x)} F \Phi_{(x)}\right)^{-1} \left(\Phi^T_{(x)} F y\right)}$$

(ک)

$\varepsilon \sim N(0, \sigma^2)$.

$$E(w) = \frac{1}{2} \sum_{i=1}^{n} (y_i - w^T . x_i)^2 \xrightarrow{x_i + \varepsilon} E(w) = \frac{1}{2} \sum_{i=1}^{n} (\underbrace{y_i - w^T x_i}_{a} - \underbrace{w^T \varepsilon_i}_{b})^2$$

$$\text{loss} = \frac{1}{2} \sum_i (y_i - w^T x_i)^2 + \frac{1}{2} \sum_i (w^T \varepsilon_i)^2 - \frac{1}{2} \sum_i \underbrace{2 (y_i - w^T x_i) w^T \varepsilon_i}_{k}$$

$$E(\text{loss})_{\text{noisy}} \longrightarrow E(\text{loss})_{\text{noiseless}} + \frac{1}{2} \sum_{i=1}^{n} w_i^2 \sigma^2 - k$$

a

$$X = \begin{bmatrix} \cdots & X_0 & \cdots \\ \cdots & X_1 & \cdots \\ & \vdots & \\ \cdots & X_n & \cdots \end{bmatrix} \quad y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$L_{(x,y)} = \sum_{i=1}^{n} (y_i - w_i x_{ji})^2 \longrightarrow \frac{\partial L}{\partial w_i} = -2 \sum x_{ji} (y_i - w_i x_{ji}) = 0$$

$$\longrightarrow \sum_{i=1}^{n} y_i \cdot x_{ji} = \sum_{i=1}^{n} w_i \, x_{ji}^2 \longrightarrow X^T \cdot y = w \, X^T \cdot X$$

$$\longrightarrow \boxed{w = \frac{X^T \cdot y}{X^T \cdot X}}$$

b) $\theta_{optimum} = (X^T \cdot X)^{-1} \cdot X^T \cdot y \xrightarrow{X:\text{orthogonal}} X^T \cdot X = ||X||^2$

$\rightarrow \theta = \dfrac{X^T \cdot y}{||X||^2} \rightarrow \theta = \left[\dfrac{x_1^T \cdot y}{||x_1||^2} \,\middle|\, \cdots \,\middle|\, \dfrac{x_n^T y}{||x_n||^2} \right]$

همچنین : $\theta_j = \dfrac{x_j^T y}{x_j^T \cdot x_j}$, $x_j^T \cdot x_j \rightarrow 1$ $\rightarrow \theta_j = y_j$

که ماتریس قطری با درایه‌های X_{ii}^2

در نتیجه انگار که هر کدام از ستون‌ها را جداگانه بررسی داریم.

ج $$L(w) = \sum_{i=1}^{n} (y_i - w_j x_{ij} - w_0)^2 \longrightarrow \frac{\partial L}{\partial w_0} = 0 \quad *$$

$$\sum y_i = \sum w_0 + w_j x_{ij} \longrightarrow w_0 = \frac{\sum_i y_i - w_j \sum_i x_{ij}}{n}$$

$$\longrightarrow w_0 = E[y] - w_j E[x_j]$$

$$* \; \frac{\partial L}{\partial w_j} = 0 \longrightarrow \sum_i x_{ij} \cdot y_i = \sum_i x_{ij} w_0 + x_{ij} w_j x_{ij} = 0$$

$$\longrightarrow w_j \sum_i x_{ij}^2 = \sum x_{ij} y_i - \sum x_{ij} w_0 \longrightarrow w_j = \frac{Cov(x_i, y)}{Var(x_i)}$$