دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

درس یادگیری ماشین

استاد سید عباس حسینی

تمرین سوم | مدل‌سازی رگرسیون | مصطفی قدیمی

سؤال ۱. خطاها

۱. • MSE:

$$L_{MSE}(\theta) = \frac{1}{n}\Sigma_i(x_i - \theta)^2$$

$$\rightarrow \theta^* = argmax_\theta \frac{1}{n}\Sigma_i(x_i - \theta)^2 = \frac{\partial}{\partial\theta}\frac{1}{n}\Sigma_i(x_i - \theta)^2 = 0$$

$$\rightarrow \theta^* = \frac{2}{n}\Sigma_i(x_i - \theta) = 0 \rightarrow \theta^* = (\frac{2}{n}\Sigma_i x_i) - 2\theta = 0 \rightarrow \theta^* = \frac{\Sigma_i}{n}$$

عبارت به‌دست آمده برای کم‌ترین مقدار میانگین مربعات خطا، میانگین است.

• MAE:

$$L_{MAE}(\theta) = \frac{1}{n}\Sigma_i|x_i - \theta|$$

با توجه به قواعدی که برای مشتق‌گیری از قدرمطلق می‌دانیم اگر $x_i > \theta$ باشد، داریم $x_i - \theta$ و در غیر این‌صورت ($x_i < \theta$)، برابر با $x_i - \theta$ است. حال برای به دست‌آوردن کم‌ترین مقدار خطا، مجموعه‌ای از عبارات داریم که ۱ـ یا ۱ هستند. بنابراین برای آن که مقدار آن صفر شود باید از نیمی بزرگ‌تر و از نیمی دیگر کوچک‌تر باشد که همان تعریف میانه است.

۲. مقایسه MSE و MAE

• MSE:

– مزایا: یک راه عالی برای مطمئن شدن از این است که مدل آموزش‌دیده‌ی ما پیش‌بینی‌های outlier با مقدار خطاهای بسیار بزرگ ندارد؛ زیرا روش میانگین مربعات خطاها وزن بزرگ‌تری نسبت به مقدار واقعی خطا (به خاطر توان دو در فرمول) اختصاص می‌دهد.

– معایب: اگر مدل ما یک پیش‌بینی بسیار بد داشته باشد، به دلیل بزرگ‌کردن و اختصاص وزن بیش‌تر از مقدار واقعی خطا به آن باعث وجود عیب در آن می‌شود. زیرا در بسیاری از موردهای عملی فقط دنبال این هستیم که روی اکثریت داده‌ها، مدل خوبی داشته باشیم و توجه زیادی به outlierها ندارند.

• MAE: مزیت این روش، برطرف کردن مشکل MSE و مشکل آن، نداشتن مزیت MSE است؛ یعنی:

– مزایا: از آن‌جایی که در این مدل در حال محاسبه‌ی قدرمطلق هستیم، مقیاس خطاها تغییری نمی‌کند و خطی باقی می‌ماند.

– عیب: اگر پیش‌بینی‌های outlier برای مدل ما مهم باشد، مدل خوبی نیست؛ زیرا وزن outlierها مشابه با وزن خطاهای کوچک‌تر است که ممکن است در برخی موارد منجر به پیش‌بینی‌های بسیار ضعیفی شود.

۳. توابع اندازه‌گیری خطا Huber و Log-Cosh

- Huber ترکیبی از MAE و MSE است و فرمول آن در ادامه آورده شده است:

$$L_\delta(y,f(x)) = \begin{cases} \frac{1}{2}(y-f(x))^2 & for\ |y-f(x)| \le \delta, \\ \delta|y-f(x) - \frac{1}{2}\delta^2 & o.w. \end{cases}$$

برای توضیح این فرمول می‌توان گفت که برای مقادیری که دلتا کوچک‌تر هستند، از MSE و در غیر این صورت، از MAE استفاده می‌کند. به بیان دیگر برای خطاهای با اندازه بزرگ از MAE و برای خطاهای کوچک از MSE بهره می‌برد. در واقع با این کار معایب روش اندازه‌گیری خطا برای هر دو روش قبل را برطرف می‌کند.

- Log-Cosh:

$$L_{log-Cosh}(x,\theta) = \Sigma_i log(cosh(\theta - x))$$

استفاده از تابع logarithm و cosh در فرمول این روش اندازه‌گیری خطا، باعث می‌شود که این روش، بسیار مشابه با روش MSE باشد با این تفاوت که وقتی پیش‌بینی بسیار بدی داریم، در مقایسه با MSE، خیلی روی خطا تاثیرگذار نباشد. (مشکل MSE را برطرف می‌کند.)
علاوه‌بر مزیت‌های روش Huber، در هر نقطه‌ای دو بار مشتق‌پذیر است.

سؤال ۲.

- اگر validationها از توزیع‌های متفاوتی باشند، در نمودار پس از چندین مرحله validation 1 مقدار خطایش از حالت نزولی به حالت صعودی تغییر پیدا می‌کند و آن را می‌توان به‌خاطر overfit شدن بعد از بهبود پارامترها توجیه کرد و نمودار validation 2 هم به توزیع داده‌ی train نزدیک است.
- اگر توزیع validationهایکسان باشند، تنها روشی که می‌توان این نمودار را توجیه کرد این است که داده‌ها بسیار کم باشند، زیرا در حالت کلی، هنگامی که داده‌ها زیاد باشد، به‌طور میانگین نمودار validationها باید یکسان باشد.

سؤال ۳.

چون مقدار $f(x)$ را به طور مستقل برای هر x می‌توان انتخاب کرد، مینیمم L_q را می‌توان با کمینه کردن انتگرال زیر به دست آورد:

$$\int |f(x) - y|^q p(y|x)dy$$

حال اگر نسبت به $f(x)$ مشتق بگیریم و آن را برابر با صفر قرار دهیم، داریم:

$$\int q|f(x) - y|^{q-1} sign(f(x) - y) p(y|x)dy = q\int_{-\infty}^{f(x)} |f(x) - y|^{q-1} p(y|x)dy - q\int_{f(x)}^{\infty} |f(x) - y|^{q-1} p(y|x)dy = 0$$

$$\rightarrow \int_{-\infty}^{f(x)} |f(x) - y|^{q-1} p(y|x)dy = \int_{f(x)}^{\infty} |f(x) - y|^{q-1} p(y|x)dy$$

اگر $q = 1$:

$$\int_{-\infty}^{f(x)} p(y|x)dy = \int_{f(x)}^{\infty} p(y|x)dy$$

بنابراین همان‌طور که از تعریف می‌دانیم و در سوال ۱ به آن پرداختیم، در میانه (در حالت پیوسته) مساحت قسمت سمت چپ و سمت راست با یک‌دیگر برابر است، یعنی $f(x)$ باید میانه باشد.

اگر $q \rightarrow 0$:

مقدار $|f(x) - y|^q$ بسیار نزدیک ۱ می‌شود (به جز همسایه‌های کوچکی نزدیک $f(x) = y$ که مقدار آن صفر می‌شود.).
بنابراین مقدار $\int |f(x) - y|^q p(y|x)dy$ نزدیک ۱ می‌شود؛ زیرا $p(y)$ نرمال شده است اما مقدار آن کمی در نزدیکی $y = f(x)$ کاهش پیدا می‌کند. بیش‌ترین کاهش مربوط به همین نقطه است که بیش‌ترین $p(y)$ را دارد.

سؤال ۴.

- بازنویسی فرمول $tanh$ بر اساس $sigmoid$

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

دو رابطه‌ای که می‌توان از فرمول‌های بالا به دست آورد، به شکل زیر هستند:

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

با توجه به نتایج بالا داریم:

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

- پیدا کردن رابطه‌ی بین وزن‌های $tanh$ و $sigmoid$

$$f(x; w^{(g)}) = w_0^{(g)} + \Sigma_i^m w_j^{(g)} g(x)$$

حال با رابطه‌ای که در قسمت قبل به‌دست آورده‌ایم و جای‌گذاری آن‌ها داریم:

$$f(x; w^{(sigmoid)}) = f(x; w^{(sigmoid)}) = w_0^{(sigmoid)} + \Sigma_i^m w_j^{(sigmoid)} S(x)$$

$$f(x; w^{(tanh)}) = f(x; w^{(tanh)}) = w_0^{(tanh)} + \Sigma_i^m w_j^{(tanh)} (2S(2x) - 1)$$

حال با تغییر متغیر $X = \frac{x}{2}$ داریم:

$$f(x; w^{(sigmoid)}) = f(x; w^{(sigmoid)}) = w_0^{(sigmoid)} + \Sigma_i^m w_j^{(sigmoid)} S(2X) = w_0^{(sigmoid)} + \Sigma_i^m \frac{w_j^{(sigmoid)}}{2} (2S(2X) - 1 + 1)$$

$$w_0^{(tanh)} + \Sigma_i^m w_j^{(tanh)} (tanh(X))$$

$$w_0^{tanh} = w_0^{sigmoid} + \Sigma_i^M \frac{w_j^{(sigmoid)}}{2},\ w_j^{(tanh)} = \frac{w_j^{(sigmoid)}}{2}\ j \in \{1, ..., M\}$$

سؤال ۵.

اگر $y_n = y(x_n, w)$ آن‌گاه داریم:

$$y_n^{noisy} = w_0 + \Sigma_i^D w_i(x_{ni} + \epsilon_{ni}) = y_n + \Sigma_i^D w_i \epsilon_{ni}$$

حال با استفاده از $E_D(w) = \frac{1}{2}\Sigma_{n=1}^N (y(x_n, w) - t_n)^2$ داریم:

$$E^{noisy} = \frac{1}{2}\Sigma_{n=1}^N (y_n^{noisy} - t_n)^2 = \frac{1}{2}\Sigma_{n=1}^N (y_n^{noisy\,2} - 2y_n^{noisy} t_n + t_n^2)$$

$$\to E^{noisy} = \frac{1}{2}\Sigma_{n=1}^N (y_n^2 + 2y_n \Sigma_{i=1}^D w_i \epsilon_{ni} + (\Sigma_{i=1}^D w_i \epsilon_{ni})^2 - 2t_n y_n - 2t_n \Sigma_{i=1}^D w_i \epsilon_{ni} + t_n^2)$$

حال اگر امید ریاضی E^{noisy}را بگیریم که جملعی $2y_n \Sigma_{i=1}^D w_i \epsilon_{ni}$ و جمله‌ی $-2t_n \Sigma_{i=1}^D w_i \epsilon_{ni} + t_n^2$ به دلیل $\mathbb{E}(\epsilon_{ni}) = 0$ حذف می‌شوند.
به دلیل استقلال ϵ_{ni}ها داریم:

$$\mathbb{E}[(\Sigma_{i=1}^D w_i \epsilon_{ni})^2] = \Sigma_{i=1}^D w_i^2 \sigma^2$$

$$\to \mathbb{E}[E^{noisy}] = E_D + \frac{1}{2}\Sigma_{i=1}^D w_i^2 \sigma^2$$

سؤال ۶.

$$L(w) = \Sigma_{i=1}^{N} F_i(y^{(i)} - w^T x)^2$$

اگر نسبت به w_j مشتق بگیریم، داریم:

$$\frac{\partial L(w)}{\partial w_j} = -2\Sigma_{i=1}^{N} F_i(y^{(i)} - w^T x_i)x_{ij} = 0,\ for\ j \in \{1, ...M\}$$

با جابه‌جا کردن طرفین معادله داریم:

$$\frac{\partial L(w)}{\partial w_j} = \Sigma_{i=1}^{N} x_{ij} F_i w^T x_i = \Sigma_{n=1}^{N} x_{ij} F_i y^{(i)}$$

اگر قسمت سمت چپ را باز کنیم، داریم:

$$\Sigma_{i=1}^{N} \Sigma_{k=1}^{M} x_{ij} F_i x_{ik} w_k = \Sigma_{i=1}^{N} x_{ij} F_i y^{(i)}$$

در صورتی که فرض کنیم که F ماتریس قطری و F_i همان F_{ii} است بنابراین:

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

سؤال ۷.

- الف) می‌دانیم:

$$w_j = (x_j^T x_j)^{-1} x_j^T y$$

چون $(x_j^T x_j)$ یک عدد است بنابراین معادله بالا را می‌توان به شکل زیر نوشت:

$$w_j = \frac{x_j^T y}{x_j^T x_j}$$

- ب)

$$w = (X^T X)^{-1} y = \begin{pmatrix} |x_1|^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & |x_M|^2 \end{pmatrix}^{-1} X^T y = \begin{pmatrix} \frac{1}{|x_1|^2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{|x_M|^2} \end{pmatrix}^{-1} X^T y$$

$$\rightarrow w_j = \frac{1}{|x_j|^2} x_j^T y = \frac{x_j^T y}{x_j^T x_j}$$

- پ)

$$w = (X^T X)^{-1} X^T y = \begin{pmatrix} N+1 & (N+1)\mathbb{E}[X_j] \\ (N+1)\mathbb{E}[X_j] & (N+1)(var(X_j) + \mathbb{E}[X_j]^2) \end{pmatrix}^{-1} X^T y$$

$$\rightarrow w = (X^T X)^{-1} X^T y = \begin{pmatrix} N+1 & (N+1)\mathbb{E}[X_j] \\ (N+1)\mathbb{E}[X_j] & \Sigma X_j^2 \end{pmatrix}^{-1} X^T y$$

$$\rightarrow w = \frac{1}{N+1}(N+1)\frac{1}{var(X_j)} \begin{pmatrix} var(X_j) + \mathbb{E}[X_j]^2 & -\mathbb{E}[X_j] \\ -\mathbb{E}[X_j] & 1 \end{pmatrix} \begin{pmatrix} \mathbb{E}[y] \\ cov(X_J, y) + \mathbb{E}[X_j]\mathbb{E}[y] \end{pmatrix}$$

$$\rightarrow w_j = \frac{cov(X_j, y)}{var(X_j)},\ w_0 = \frac{var(X_j)\mathbb{E}[y] - \mathbb{E}[X_j]cov(X_j, y)}{var(X_j)} = \mathbb{E}[y] - w_j\mathbb{E}[X_j]$$