تمرین سری 2 درس یادگیری عمیق - رادین خیام - 99101579

سوال 1-

در ابتدا برای اثبات محدب بودن ①:

نام گذاری $\hat{y}_n$

$$E_{w,b} = -\sum_n y_n \ln \hat{y}(x_n) + (1-y_n)\ln(1-\hat{y}(x_n))$$

$$\hat{y}_n = \frac{1}{1+e^{-(w^T x_n + b)}} \longrightarrow \frac{\partial \hat{y}_n}{\partial w} = \hat{y}_n(1-\hat{y}_n)x_n$$

$$\longrightarrow \frac{\partial \hat{y}_n}{\partial b} = \hat{y}_n(1-\hat{y}_n)$$

$$\frac{\partial E_{w,b}}{\partial w} = \sum_n \left(\frac{1-y_n}{1-\hat{y}_n} - \frac{y_n}{\hat{y}_n}\right)\frac{\partial \hat{y}_n}{\partial w} = \sum_n (\hat{y}_n - y_n)x_n$$

$$\frac{\partial^2 E_{w,b}}{\partial w\, \partial w^T} = \sum_n \frac{\partial}{\partial w}\left((\hat{y}_n - y_n)x_n\right) = \sum_n \underbrace{\hat{y}_n(1-\hat{y}_n)}_{\text{همیشه مثبت}}\underbrace{x_n x_n^T}_{PSD}$$

⟸ بنابراین ماتریس هسین PSD هست و درنتیجه $E_{w,b}$ محدب هست.

$$\frac{\partial E_{w,b}}{\partial b} = \sum_n \left(\frac{1-y_n}{1-\hat{y}_n} - \frac{y_n}{\hat{y}_n}\right)\frac{\partial \hat{y}_n}{\partial b} = \hat{y}_n - y_n$$

ادامه سؤال ۱-

$$\frac{\partial^2 E_{w,b}}{\partial b^2} = \sum_n \frac{\partial}{\partial b}(\hat{y}_n - y_n) = \sum_n \underbrace{\hat{y}_n(1-\hat{y}_n)}_{\text{همیشه مثبت}}$$

بنابراین ماتریس هسین PSD هست و تابع محدب هست.

حال چون می‌دانیم که $E_{w,b}$ یک مینیمم گلوبال دارد بدین شکل می‌توانیم b و w را در آن نقطه بدست آوریم:

$$w_{n+1} = w_n - \eta_1 \sum_n (\hat{y}_n - y_n) x_n$$

$$b_{n+1} = b_n - \eta_2 \sum_n (\hat{y}_n - y_n)$$

سؤال 2-

الف) Covariate shift به پدیده‌ای می‌گویند که در آن توزیع داده‌های ورودی یک شبکه عصبی در هنگام آموزش تغییر می‌کند و باعث سخت‌تر شدن یادگیری می‌شود. در شبکه‌های عصبی خروجی هر لایه ورودی لایه بعد است و زمانی که وزن‌های هر لایه آپدیت می‌شوند، توزیع ورودی به هر یک از لایه‌ها هم تغییر می‌کند. این پدیده سرعت آموزش را کاهش می‌دهد چون شبکه باید همش خودش را توزیع متغیر داده‌ها مطابق کند.

ادامه سؤال 2 الف)

BN برای هر مینی بچ میانگین و واریانس داده‌های ورودی به لایه را محاسبه می‌کند و آن را به گونه‌ای نرمالایز می‌کند که میانگین صفر و واریانس واحد داشته باشند. به این ترتیب سعی می‌کند که توزیع داده‌ها بهم نریزد.
همچنین بعد از اینکه به میانگین صفر و واریانس واحد رسیدند، BN پارامترهای قابل آموزشی برای شیفت و اسکیل در نظر می‌گیرد تا اگر میانگین صفر و واریانس واحد مناسب نبود مدل بتواند آن را تغییر بدهد.

ب)

همین کاهش Covariate shift می‌تواند به جنرالیزیشن شبکه کمک کند.
BN اثراتی مشابه با بقیه روش‌های رگولاریزیشن مثل L2 و dropout دارد و برای همین می‌توانیم آن‌ها را از شبکه حذف کنیم تا مدل ساده‌تر شود. همچنین می‌توانیم از لرنینگ ریت بالاتری استفاده کنیم.
در نهایت می‌توان گفت که BN مسیر اپتیمیزیشن را نرم‌تر می‌کند که این می‌تواند به جنرالایز شدن مدل کمک کند.

پ)

$$\mu = \frac{1}{n} \sum_{j=1}^{n} x_j$$

$$\hat{x}_i = x_i - \mu = x_i - \frac{1}{n} \sum_{j=1}^{n} x_j$$

$$L = \sum_{j=1}^{n} L_j$$

ادامه سؤال 2 پ)

$$\rightarrow \frac{\partial \hat{x}_i}{\partial x_j} = \begin{cases} 1 - \frac{1}{n}, & i = j \\ -\frac{1}{n}, & i \neq j \end{cases} \quad , \quad y_i = \gamma \hat{x}_i + \beta \;\rightarrow\; \frac{\partial L}{\partial \hat{x}_i} = \gamma \frac{\partial L}{\partial y_i}$$

$$\frac{\partial L}{\partial x_i} = \sum_{j=1}^{n} \frac{\partial L}{\hat{x}_j} \times \frac{\partial \hat{x}_j}{\partial x_i} = -\frac{1}{n}\left(\sum_{j=1}^{n} \frac{\partial L}{\partial \hat{x}_j}\right) + \frac{\partial L}{\partial \hat{x}_j}$$

$$= \gamma \left(\frac{\partial L}{\partial y_i} - \frac{1}{n} \sum_{j=1}^{n} \frac{\partial L}{\partial y_j} \right)$$

ت)

$n = 1$: $\quad \mu = x_1 \rightarrow \hat{x}_1 = x_1 - x_1 = 0 \rightarrow \frac{\partial y_1}{\partial \hat{x}_1} = 0$

$$\rightarrow \frac{\partial L}{\partial x_1} = 0$$

$n \rightarrow \infty$:

در این حالت تأثیر هر x_i بر روی μ بسیار اندک هست ← $\frac{\partial \mu}{\partial x_i} = 0$

$$\rightarrow \frac{\partial \hat{x}_i}{\partial x_j} \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad \rightarrow \quad \frac{\partial L}{\partial x_i} = \gamma \frac{\partial L}{\partial y_i}$$

* نتیجه گرفته می‌شود که اگر n خیلی بزرگ شود اثر نرمالایزیشن از بین می‌رود.

سوال 3 -

الف)

$$\hat{y}_k = \text{softmax}(z^{(2)}) = \frac{e^{z_k^{(2)}}}{\sum_{j=1}^{K} e^{z_j^{(2)}}}$$

$$\frac{\partial \hat{y}_k}{\partial z_i^{(2)}} = \frac{\frac{\partial e^{z_k^{(2)}}}{\partial z_i^{(2)}} \sum_{j=1}^{K} e^{z_j^{(2)}} - \left(\sum_{j=1}^{K} \frac{\partial e^{z_j^{(2)}}}{\partial z_i^{(2)}}\right) \times e^{z_k^{(2)}}}{\left(\sum_{j=1}^{K} e^{z_j^{(2)}}\right)^2}$$

$$i = k: \quad \frac{\partial \hat{y}_k}{\partial z_k^{(2)}} = \frac{e^{z_k^{(2)}} \sum_{j=1}^{K} e^{z_j^{(2)}} - \left(e^{z_k^{(2)}}\right)^2}{\left(\sum_{j=1}^{K} e^{z_j^{(2)}}\right)^2} = \hat{y}_k (1 - \hat{y}_k)$$

$$i \neq k: \quad \frac{\partial \hat{y}_k}{\partial z_i^{(2)}} = \frac{-e^{z_k^{(2)}} e^{z_i^{(2)}}}{\left(\sum_{j=1}^{K} e^{z_j^{(2)}}\right)^2} = -\hat{y}_k \hat{y}_i$$

ب)

$$L = -\sum_{i=1}^{K} y_i \log(\hat{y}_i) = -\log(\hat{y}_k)$$

$$\rightarrow \frac{\partial L}{\partial z_i^{(2)}} = \frac{\partial L}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_i^{(2)}}$$

$$\frac{\partial L}{\partial z_i^{(2)}} = \hat{y}_i - y_i$$

$$\frac{\partial L}{\partial \hat{y}_i} = \begin{cases} \frac{-1}{\hat{y}_k}, & i = k \\ 0, & i \neq k \end{cases} \longrightarrow \frac{\partial L}{\partial z_i^{(2)}} = \begin{cases} \hat{y}_k - 1 & i = k \\ \hat{y}_i & i \neq k \end{cases}$$

پ)

$$\frac{\partial L}{\partial z^{(2)}} = \hat{y} - y$$

$$\frac{\partial z^{(2)}}{\partial a^{(1)}} = (w^{(2)})^T \quad , \quad \frac{\partial a^{(1)}}{\partial \tilde{a}^{(1)}} = \begin{cases} 1, & p = 0.8 \\ 0, & p = 0.2 \end{cases}$$

$$\frac{\partial \tilde{a}^{(1)}}{\partial z_i^{(1)}} = \begin{cases} 1, & z_i^{(1)} > 0 \\ 0.01, & z_i^{(1)} \leq 0 \end{cases} \quad , \quad \frac{\partial z^{(1)}}{\partial w^{(1)}} = x^T$$

$$\longrightarrow \frac{\partial L}{\partial w^{(1)}} = \frac{\partial L}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial \hat{a}^{(1)}} \frac{\partial \hat{a}^{(1)}}{\partial z_i^{(1)}} \frac{\partial z^{(1)}}{\partial w^{(1)}}$$

سؤال ۴-

$$\nabla y = \begin{bmatrix} \frac{\partial y}{\partial z} \\ \frac{\partial y}{\partial v} \\ \frac{\partial y}{\partial u} \end{bmatrix} \longrightarrow J(\nabla y) = \begin{bmatrix} \frac{\partial^2 y}{\partial u^2} & \frac{\partial^2 y}{\partial u \partial v} & \frac{\partial^2 y}{\partial u \partial z} \\ \frac{\partial^2 y}{\partial v \partial u} & \frac{\partial^2 y}{\partial v^2} & \frac{\partial^2 y}{\partial v \partial z} \\ \frac{\partial^2 y}{\partial z \partial u} & \frac{\partial^2 y}{\partial z \partial v} & \frac{\partial^2 y}{\partial z^2} \end{bmatrix}$$

$$H = \begin{bmatrix} \frac{\partial^2 y}{\partial u^2} & \frac{\partial^2 y}{\partial u \partial v} & \frac{\partial^2 y}{\partial u \partial z} \\ \frac{\partial^2 y}{\partial v \partial u} & \frac{\partial^2 y}{\partial v^2} & \frac{\partial^2 y}{\partial v \partial z} \\ \frac{\partial^2 y}{\partial z \partial u} & \frac{\partial^2 y}{\partial z \partial v} & \frac{\partial^2 y}{\partial z^2} \end{bmatrix} \longrightarrow$$ میبینیم که مشابه هستند

سؤال 5 -

$$J_1 = \frac{1}{2}\left(y_d - \sum_{k=1}^{n} \delta_k w_k x_k\right)^2 \quad , \quad \delta_k \sim N(1, \sigma^2)$$

$$\frac{\partial J_1}{\partial w_i} = \frac{1}{2} \frac{\partial}{\partial w_i}\left(y_d - \sum_{k=1}^{n} \delta_k w_k x_k\right)^2$$

$$= \left(y_d - \sum_{k=1}^{n} \delta_k w_k x_k\right) \times -\delta_i x_i$$

$$\rightarrow E\left[\frac{\partial J_1}{\partial w_i}\right] = -x_i y_d \, \overset{1}{\cancel{E[\delta_i]}} + x_i \sum_{k=1}^{n} E[\delta_i \delta_k] \, w_k x_k$$

$$E[\delta_i \delta_k] = \begin{cases} E[\delta_i]^2 + var(\delta_i) = 1 + \sigma^2 & i = k \\ E[\delta_i] \, E[\delta_k] = 1 & i \neq k \end{cases}$$

$$\rightarrow E\left[\frac{\partial J_1}{\partial w_i}\right] = -x_i y_d + x_i\left((1+\sigma^2)\, w_i x_i + \sum_{\substack{k=1 \\ k \neq i}}^{n} w_k x_k\right)$$

$$= -x_i\left(y_d - (1+\sigma^2) w_i x_i - \sum_{\substack{k=1 \\ k \neq i}}^{n} w_k x_k\right)$$

ادامه سؤال 5 -

non-regularized $\longrightarrow \hat{\mathcal{F}}_1 = \frac{1}{2}\left(y_d - \sum_{k=1}^{n} w_k x_k\right)^2$

$\longrightarrow \frac{\partial \hat{\mathcal{F}}_1}{\partial w_i} = -x_i \times \left(y_d - w_i x_i - \sum_{\substack{k=1 \\ k \neq i}}^{n} w_k x\right.$

یک ترم σ^2 کم شده

در واقع ما یک ترم $-\sigma^2 w_i x_i$ به مشتق اضافه کرده ایم و این مانند

این هست که ما به w_i یک نویز اضافه کرده ایم.

سؤال 6 -

$e_k = x_k - x^* \rightarrow x^* = x_k - e_k$

$f(x) = g'(x) \longrightarrow x_{k+1} = x_k - \frac{g'(x)}{g''(x)} = x_k - \frac{f(x)}{f'(x)}$

$f(x^*) = f(x_k - e_k) \simeq f(x_k) - e_k f'(x_k) + \frac{e_k^2}{2} f''(\varepsilon_k)$

$x_k < \varepsilon_k < x^*$

$f(x^*) = 0 \longrightarrow f(x_k) - e_k f'(x_k) + \frac{e_k^2}{2} f''(\varepsilon_k) = 0$

ادامه سؤال 6 -

$$\to \frac{f(x_k)}{f'(x_k)} - e_k + \frac{e_k^2}{2f'(x_k)} f'''(\varepsilon_k)$$

$$\to x_k - x_{k-1} - x_k + x^* + \frac{(x_k - x^*)^2}{2f'(x_k)} f'''(\varepsilon_k)$$

$$\to x_{k-1} - x^* = \frac{(x_k - x^*)^2}{2f'(x_k)} f''(\varepsilon_k) \to |x_{k-1} - x^*| \leqslant \frac{(x_k - x^*)^2}{2|f'(x_k)|} |f''(\varepsilon_k)|$$

اگر $k \to \infty$ با توجه به پیوستگی خواهیم داشت :

$$f'(x_k) \to f'(x^*) \quad , \quad x_k \to x^*$$

سؤال 7 -

الف)

$$\frac{\partial}{\partial z_i} L(z,y) = \frac{\partial}{\partial z_i}\left(-\sum_{k=1}^{K} y_k \log \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}\right) = \frac{\partial}{\partial z_i}\left(-\sum_{k=1}^{K} y_k z_i + \sum_{k=1}^{K} y_k \left(\log \sum_{j=1}^{K} e^{z_j}\right)\right)$$

$$= -\sum_{k=1}^{K} y_k + \sum_{k=1}^{K} y_k \left(\frac{z_i}{\sum_{j=1}^{K} e^{z_j}}\right) = \hat{y}_i - 1 = \hat{y}_i - y_i$$

$$\to \nabla z\, L(z,y) = \hat{y} - y$$

ادامه سؤال ۷ -

ب)

$$\frac{\partial^2}{\partial z_i^2} L(z,y) = \frac{\partial}{\partial z_i}(\hat{y}_i - 1) = \frac{\partial}{\partial z_i}\left(\frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} - 1\right) = \hat{y}_i(1-\hat{y}_i)$$

$$\frac{\partial^2}{\partial z_i \partial z_j} L(z,y) = \frac{\partial}{\partial z_j}\left(\frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} - 1\right) = -\hat{y}_i\hat{y}_j$$

$$\rightarrow H = \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & -\hat{y}_1\hat{y}_2 & \cdots & -\hat{y}_1\hat{y}_k \\ -\hat{y}_2\hat{y}_1 & \hat{y}_2(1-\hat{y}_2) & & -\hat{y}_2\hat{y}_k \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{y}_k\hat{y}_1 & -\hat{y}_k\hat{y}_2 & \cdots & \hat{y}_k(1-\hat{y}_k) \end{bmatrix}$$

$$\rightarrow H = diag(\hat{y}) - \hat{y}\hat{y}^T$$

$$\rightarrow x^T H x = x^T diag(\hat{y})\, x - x^T \hat{y}\hat{y}^T x = \sum_{i=1}^{k} \hat{y}_i x_i^2 - (\hat{y}^T x)^2$$

$$(\hat{y}^T x)^2 = \left(\sum_{i=1}^{k} \hat{y}_i x_i\right)^2$$

$$\sum_{i=1}^{k} \hat{y}_i x_i^2 = \left(\sum_{i=1}^{k} \hat{y}_i\right)\left(\sum_{i=1}^{k} \hat{y}_i x_i^2\right)$$

طبق نامساوی کوشی شوارتز :

$$\left(\sum_{i=1}^{k} \hat{y}_i x_i\right)^2 \leq \left(\sum_{i=1}^{k} \hat{y}_i\right)\left(\sum_{i=1}^{k} \hat{y}_i x_i^2\right) \rightarrow \sum_{i=1}^{k} \hat{y}_i x_i^2 \geq (\hat{y}^T x)^2$$

PSD $\leftarrow x^T H x \geq 0$ پس

چون ماتریس Hessian ، PSD هست پس تابع محدب هست.