به نام خدا

امیرمحمد یعقوبیان زاده ۴۰۱۴۰۰۵۲۵ تمرین ۲ یادگیری عمیق

(۱) (آ)

$$L = f_{(\omega_0)} + \nabla f_{(\omega_0)}^{T}(\omega-\omega_0) + \frac{1}{2}(\omega-\omega_0)^T H (\omega-\omega_0)$$

چون با GD همگرا شده‌ایم پس $\nabla f(\omega_0) \approx 0$. هدف ما این است که کدام وزن‌ها موجود در ω_0 را تغییر دهیم به طوری که L (خطا) کمترین افزایش را داشته باشد.

$$\hat{\omega} = \omega_0 - \omega_i$$

تمام درایه‌های آن صفر است به جز درایه‌ی iام که برابر $\omega_{0(i)}$ است.

$$\omega_i = \begin{bmatrix} \vdots \\ \omega_{0(i)} \\ \vdots \end{bmatrix} \leftarrow \text{درایه‌ی } i\text{ام}$$

$$\Rightarrow L_i = f_{(\omega_0)} + 0 + \frac{1}{2}(-\omega_i)^T H (-\omega_i) = f_{(\omega_0)} + \frac{1}{2}\omega_i^T H \omega_i$$

$$L_i = f_{(\omega_0)} + \frac{1}{2}|\omega_{0(i)}|^2 H_{ii}$$

از بین وزن‌های موجود باید وزنی را صفر کنیم که مقدار $|\omega_{0(i)}|^2 H_{ii}$ برای آن کمینه باشد.

(ب)

$H = I \rightarrow H_{ii} = 1 \rightarrow$ باید وزن‌هایی را حذف کنیم که کمترین نرم ۲ را داشته باشند.

(۴) آ)

$$b^* = \underset{b}{\operatorname{argmin}} \frac{1}{N}\|y - Xb\|^2 \xrightarrow{\frac{\partial}{\partial b}} \frac{\partial}{\partial b}(y^Ty + b^TX^TXb - 2y^TXb) = 0$$

$$\Rightarrow 2X^TXb = 2X^Ty \Rightarrow b^* = (X^TX)^{-1}X^Ty$$

ب)

$A^T = X(X^TX)^{-1}X^T = A \rightarrow$ متقارن است. A

$$A^TA = A^2 = A \Rightarrow Av = \lambda v \Rightarrow \begin{cases} A^2v = \lambda Av = \lambda^2 v \\ A^2v = Av = \lambda v \end{cases} \Rightarrow \lambda^2 = \lambda \Rightarrow \boxed{\lambda = 1 \text{ یا } 0}$$

با نوشتن تجزیه SVD برای $X = U\Sigma V^H$ و بدست آوردن A می‌بینیم که ماتریس A دارای رنک d بوده دیگ ماتریس قطری با d درایه‌ی آ

و $N-d$ درایه‌ی صفر است.

$$\epsilon\left[\frac{1}{N}\|\underbrace{(X(X^TX)^{-1}X^T - I)}_{A}\epsilon\|_2^2\right] = \frac{1}{N}E\left[\sum_{i=1}^{N-d}\epsilon_i^2\right] = \frac{1}{N}\times\sum_{i=1}^{N-d}E[\epsilon_i^2]$$

ماتریس خطی با $N-d$ درایه‌ی ۱ و d درایه‌ی صفر در قطر اصلی

$$= \frac{1}{N}(N-d)\sigma^2$$

(ج) هرچه مقدار d به N نزدیک‌تر شود مقدار $N-d$ به سمت صفری رود و خطای محاسبه شده رفته‌رفته کاهش می‌یابد.

$$\lim_{d\to N}\left(\frac{N-d}{N}\right)\sigma^2 = 0$$

---

(۳) آ)

$$\hat{\beta} = \arg\min_{\beta}\|Y - X\beta\|_F^2 \rightarrow \frac{\partial}{\partial\beta}(Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta) = 0$$

$$-2X^TY + 2X^TX\beta = 0 \Rightarrow \beta = (X^TX)^{-1}X^TY$$

ب)

$$\hat{\beta} = \arg\min_{\beta}\|Y - X\beta\|_F^2 + \lambda\|\beta\|_F^2 \rightarrow \frac{\partial}{\partial\beta}(\|Y - X\beta\|_F^2 + \lambda\|\beta\|_F^2) = 0$$

$$\Rightarrow -2X^TY + 2X^TX\beta + 2\lambda\beta = 0 \Rightarrow \beta = (X^TX + \lambda I)^{-1}X^TY$$

ج) $\Sigma_{N\times N}X_{N\times K} = X_{N\times K}F_{K\times K}$ ← F ماتریس غیرتکین است و بنابراین دارای وارون است.

$$X = \Sigma^{-1}XF$$

(د) اگر ←

$$\hat{\beta} = (X^TX)^{-1}X^TY = (\underbrace{F^TX^T\Sigma^{-1}}_{X^T}X)^{-1}\underbrace{F^TX^T\Sigma^{-1}}_{X^T}Y = (X^T\Sigma^{-1}X)^{-1}F^{-T}F^TX^T\Sigma^{-1}Y$$

$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

(تنها اگر) ←

$$(X^TX)^{-1}X^TY = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y \Rightarrow X^TY = X^TX(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

چون باید بازاء هر Y برقرار باشد.

$$X^T = X^TX(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} \xrightarrow{\times\Sigma,\ \text{transpose}} \Sigma X = X\underbrace{(X^T\Sigma^{-1}X)^{-1}X^TX}_{\triangleq F}$$

← برای آنکه تساوی برقرار باشد باید $\Sigma X = XF$ باشد و طبق تعریف، F یک ماتریس معکوس‌پذیر و غیرتکین است.

(د) اگر $X\in\mathbb{R}^{N\times d}$ باشد و نمونه‌ها دارای d ویژگی باشند با افزودن λ_1 تا نمونه به صورت $(x=e_1$ و $y=0)$، λ_1 تا نمونه به صورت $(x=e_2$ و $y=0)$ و ... و λ_1 تا نمونه به صورت $(x=e_d$ و $y=0)$ به دیتاست و اعمال یک رگولایزر L_1 به تابع

هدف می‌توان تابع خطای نشان داده شده را یافت:

$$L = \| y - X\beta \|^2 + \lambda_2 \|\beta\|_1 = \| y_{old} - X_{old}\beta \|^2 + \lambda_1 \sum_{i=1}^{d} (0 - e_i\beta)^2$$

$$+\lambda_2 \|\beta\|_1 = \| y_{old} - X_{old}\beta \|^2 + \lambda_1 \sum_{i=1}^{d} \beta_i^2 + \lambda_2 \|\beta\|_1$$

$$= \| y_{old} - X_{old}\beta \|^2 + \lambda_1 \|\beta\|_2^2 + \lambda_2 \|\beta\|_1$$

---

(۴) آ)

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

$$\Rightarrow E\left[\frac{\partial J}{\partial w_i}\right] = E\left[-2y_d x_i \delta_i + 2x_i \delta_i \sum_{k=1}^{n} \delta_k w_k x_k\right] = -2y_d x_i + 2x_i \sum_{\substack{k=1 \\ k\neq i}}^{n} w_k x_k$$

$$+ 2x_i^2 w_i (\sigma^2 + 1) = 2x_i^2 w_i \sigma^2 - 2y_d x_i + 2x_i \sum_{k=1}^{n} w_k x_k \quad (I)$$

ب) ابتدا گرادیان را برای حالتی که این نوع رگولاریزاسیون وجود ندارد محاسبه می‌کنیم:

$$\frac{\partial \hat{J}}{\partial w_i} = -2\left(y_d - \sum_{k=1}^{n} w_k x_k\right) \times x_i = -2y_d x_i + 2x_i \sum_{k=1}^{n} w_k x_k \quad (II)$$

با مقایسه روابط (I) و (II) متوجه می‌شویم که عبارت $2x_i^2 w_i \sigma^2$ در گرادیان حالت dropout اضافه شده است.

آنچه تابع خطای $\hat{J}$ مقدار $\sigma^2 \|x_0 w\|^2$ اضافه شود هنگام $\frac{\partial}{\partial w_i}$ گرفتن مقدار $2x_i^2 w_i \sigma^2$ تولید خواهد شد. بنابراین چنین رگولاریزری به تابع خطا اضافه شده است که اندکی با رگولاریزر L_2 متفاوت است و در آن مقادیر بردار وزن‌ها scale ابتدا بابت می‌شود و سپس رگولاریزر L_2 بر روی بردار جدید اعمال می‌گردد.

---

(۵) آ)

$$w_t = w_{t-1} - \epsilon \nabla_w (w^T H w)\Big|_{w = w_{t-1}} = w_{t-1} - 2\epsilon H w_{t-1} = (I - 2\epsilon H) w_{t-1}$$

ب)

$$w_t = (I - 2\epsilon H) w_{t-1} = (I - 2\epsilon H)^2 w_{t-2} = \dots = (I - 2\epsilon H)^t w_0$$

ج)

$$I - 2\epsilon H = QQ^T - 2\epsilon Q \Lambda Q^T = Q(\underbrace{I - 2\epsilon \Lambda}_{\text{ماتریس قطری}})Q^T$$

$$w_t = Q(\underbrace{I - 2\epsilon \Lambda}_{\text{ماتریس قطری}})^t Q^T w_0$$

w ام ان $\rightarrow$ $(1 - 2\epsilon\lambda_i)^t$

اگر $|1-2\epsilon\lambda_i|<1$ باشد این الگوریتم همگرا خواهد شد.

$$\Rightarrow -1<1-2\epsilon\lambda_i<1 \quad \forall\lambda_i \Rightarrow \begin{cases}\epsilon\lambda_i>0\\ \epsilon\lambda_i<1\end{cases}, \forall i$$

با $\lambda_i>0$ داریم: $\epsilon<\frac{1}{\lambda_{max}}$.

(ج)

$$w_{t+1}=w_t-(2H)^{-1}\times 2Hw_t=w_t-H^{-1}Hw_t=0$$

با این روش در یک گام به جواب بهینهٔ $w=0$ می‌رسیم.

(د) تابع خطا در شبکه‌های عصبی پیچیده‌تر است و هم‌سبه ماتریس هسین و معکوس کردن آن در هر گام بسیار هزینه‌بر است.

(ه) در این شبکه $W=0$ را یاد می‌گیرد و اصولاً مفهوم خاصی را آموزش نمی‌بیند. اگر $W=0$ باشد $h_1=h_2=\tanh(b)$ و $J=0+0=0$ خواهد شد که کمترین مقدار تابع هدف است.

(ب)

$$\|h_1-h_2\|_2^2=\sum_{k=1}^{m}\left(\tanh\left[\sum_{i=1}^{n}w_{ki}x_{1i}+b_k\right]-\tanh\left[\sum_{i=1}^{n}w_{ki}x_{2i}+b_k\right]\right)^2$$

$$\frac{\partial\|h_1-h_2\|_2^2}{\partial w_{\ell,t}}=2(\quad)\times\left[\left(1-\tanh^2\left[\sum_{i=1}^{n}w_{\ell i}x_{1i}+b_\ell\right]\right)\times x_{1t}-x_{2t}\times\right.$$

$$\left.\left(1-\tanh^2\left[\sum_{i=1}^{n}w_{\ell i}x_{2i}+b_\ell\right]\right)\right]=2(h_{1\ell}-h_{2\ell})\left[(1-h_{1\ell}^2)x_{1t}-(1-h_{2\ell}^2)x_{2t}\right]$$

$$=2\left[(h_1-h_2)\circ\left[(1-h_1\circ h_1)x_1^T-(1-h_2\circ h_2)x_2^T\right]\right]_{\ell,t}$$

$\circ$ ← ابرالگو ضرب درایه به درایه (ضرب هادامارد)

$$\Rightarrow\begin{cases}\frac{\partial J}{\partial W}=2(h_1-h_2)\circ\left[(1-h_1\circ h_1)x_1^T-(1-h_2\circ h_2)x_2^T\right]+2W\\ W_{t+1}=W_t-\epsilon\left.\frac{\partial E}{\partial W}\right|_{W=W_t}\end{cases}$$

$$\frac{\partial |h_1 - h_2|_2^2}{\partial b_j} = 2(h_{1j} - h_{2j})\left[(1 - h_{1j}^2) - (1 - h_{2j}^2)\right]$$

$$\Rightarrow \begin{cases} \frac{\partial J}{\partial b} = 2(h_1 - h_2) \circ (h_2 \circ h_2 - h_1 \circ h_1) \\ b_{t+1} = b_t - \epsilon \left.\frac{\partial J}{\partial b}\right|_{b = b_t} \end{cases}$$