تمرین سری دوم درس یادگیری عمیق

علیرضا ملکی - ۴۰۱۲۱۰۹۵۶

سوال ۱)

آ-

$$\delta E = \left(\frac{\partial E}{\partial w}\right)^T \cdot \delta w + \frac{1}{2}\,\delta w^T \cdot H \cdot \delta w + O(\|\delta w\|^3)$$

← هدف آن است که جمله دوم کمترین تغییرات را داشته باشد:

$$\operatorname{Min}_q \left\{ \operatorname{Min}_{\delta w} \left\{ \frac{1}{2}\,\delta w^T \cdot H \cdot \delta w \right\} \text{ such that } e_q^T \cdot \delta w + w_q = 0 \right\}$$

Lagrangian:

$$L = \frac{1}{2}\,\delta w^T \cdot H \cdot \delta w + \lambda\left(e_q^T \delta w + w_q\right)$$

$$\delta w = -\frac{w_q}{[H^{-1}]_{qq}} H^{-1} \cdot e_q \quad ; \quad L_q = \frac{1}{2}\frac{w_q^2}{[H^{-1}]_{qq}}$$

مقدار افزایش خطای وزن q با حذف آن

← اگر L_q نسبت به E کوچک باشد، وزن q ام را حذف می‌کنیم

سوال ۲ -

$$E^2 = (y - Xb)^T(y - Xb) = y^Ty - y^TXb - b^TX^Ty + b^TX^TXb$$

آ) $\rightarrow \dfrac{dE^2}{db} = -2y^TX + 2X^TXb \Rightarrow \hat{b} = (X^TX)^{-1}X^Ty$

ب) $Y = Xb + \varepsilon \rightarrow y \sim N(Xb, \sigma^2)$

← خروجی یک متغیر نرمال با میانگین Xb و واریانس σ^2 است، Xb یک ثابت است و خطای y تنها برای ε است.

← MSE برای یک داده متناسب با σ^2 است

← همچنین با زیاد شدن تعداد داده‌ها خطا افزایش می‌یابد (خطای جمع شونده) ← متناسب با N نیز است

← اگر بعد زیاد شود و σ^2 ثابت بماند، نیز کاهش می‌یابد

* همچنین رنک ماتریس X برابر d می‌باشد.

$$\hat{Y} - Y = X(X^TX)^{-1}X^Ty - y = (X(X^TX)^{-1}X^T - I)Y$$

اگر رنک ماتریس d باشد:

$$X(X^TX)^{-1}X^T = \overbrace{\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & \ddots \; 0 \end{pmatrix}}^{d} \Rightarrow A - I = \overbrace{\begin{pmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & & -1 & \\ & & & & \ddots \; -1 \end{pmatrix}}^{d}$$

$$\rightarrow \frac{1}{N}\|(A-I)\varepsilon\|_2^2 = \frac{N-d}{N}\sigma^2$$

ج) همانطور که در قسمت قبل مشاهده کردیم، $\frac{N-d}{N}$ نماینده خطای برازش می باشد که می توان دید با زیاد شدن d ($d \to N$) خطا کاهش می یابد.

می توان به نحوی گفت که خطای ε در d بعد پخش می شود و در نرم اثر خود را کمتر نشان می دهد. ولی این موضوع در داده test خود را بروز می دهد و باعث افزایش خطا می شود.

سوال ۳)

آ- در سوال قبل اثبات شده است.

ب-

$$E = \frac{\alpha}{2}W^TW + (XW-Y)^T(XW-Y)$$

$$= \frac{\alpha}{2}W^TW + W^TX^TXW - 2W^TX^TY$$

$$= W^T(I\frac{\alpha}{2}+X^TX)W - 2W^TX^TY$$

$$\frac{dE}{dW}=0 \Rightarrow W = (\frac{\alpha}{2}I+X^TX)^{-1}X^TY$$

ج)

$$\beta^* = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$$

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

$$= (X^TXF^{-1})^{-1}X^T\Sigma^{-1}y = F(X^TX)^{-1}X^T\Sigma^{-1}y \quad (1)$$

داریم: $\Sigma X = XF \rightarrow (X^TX)^{-1}X^T\Sigma X = (X^TX)^{-1}X^TXF = F$

$$\rightarrow \beta^* = (X^TX)^{-1}X^T\underbrace{\Sigma X(X^TX)^{-1}X^T\Sigma^{-1}}_{=I}y = (X^TX)^{-1}X^TY$$

ز)

$$L = |y - X\beta|^2 + \lambda_1 \|0 - I\beta|^2 + \lambda_2 \|\beta\|_1$$

$$= |y - X\beta|^2 + |0 - \sqrt{\lambda_1} I\beta|^2 + \lambda_2 \|\beta\|_1$$

می‌توان برای هر بعد یک $\hat{e}_i$ داده با برچسب ۰ اضافه کرد

$$\forall i \in \{1, \dots, d\} : \sqrt{\lambda_1}\, \hat{e}_i \,,\, 0$$

سوال ۴ -

الف -

$$D = \begin{pmatrix} \delta_1 & & \emptyset \\ & \delta_2 & \\ \emptyset & & \ddots \;\delta_n \end{pmatrix}$$

$$J = (y - X^T D w)^T (y - X^T D w)$$

$$= y^T y - 2 w^T D X y + w^T D X X^T D w$$

$$\rightarrow \nabla_w J = -2 D X y + 2 D X X^T D w$$

$$E(D) = I \qquad A = \begin{pmatrix} x_1 x_1 \delta_1 \delta_1 & & x_1 x_n \delta_1 \delta_n \\ x_2 x_1 \delta_2 \delta_1 & \ddots & \\ \vdots & & x_n x_n \delta_n \delta_n \end{pmatrix}$$

$$E[A] = \begin{pmatrix} x_1^2 & & \emptyset \\ & x_2^2 & \\ \emptyset & & \ddots\; x_n^2 \end{pmatrix} (\sigma^2 + 1) + \begin{pmatrix} 0 & x_2 x_1 & \dots \\ x_1 x_2 & 0 & x_3 x_2 \; \dots \\ x_1 x_3 & & \ddots \\ \vdots & & \quad 0 \end{pmatrix}$$

$$\rightarrow \nabla_w J = -2Xy + 2(1+\sigma^2) \begin{pmatrix} x_1^2 & & \emptyset \\ & \ddots & \\ \emptyset & & x_n^2 \end{pmatrix} w + 2 \begin{pmatrix} 0 & x_1 x_2 & \\ x_2 x_1 & 0 & \\ & \ddots & x_{n-1} x_n \\ & x_n x_{n-1} & 0 \end{pmatrix}$$

ب - می‌توان با تنظیم تاثیر جمله اول نسبت جمله دوم را تغییر داد.

$$\nabla J = -2Xy + 2\sigma^2 \sum x_i^2 w_i + 2 \sum_i \sum_j w_i x_i x_j$$

$$\Rightarrow J = \Big(y - \sum_i w_i x_i\Big)^2 + \sigma^2 \sum_i x_i^2 w_i^2$$

سوال ۵ -

الف - $J(w) = w^T H w \rightarrow \nabla_w J(w) = 2Hw$

$$\rightarrow w^{t+1} = w^t - 2\epsilon H w^t = (I - 2\epsilon H) w^t$$

ب - $w^t = (I - 2\epsilon H)^t w_0$

ج - برای همگرایی

$$\det(I - 2\epsilon H) \le 1$$

$\lim_{t \to \infty} w^t = 0$ ؛ در این صورت

$$\rightarrow \prod_{i=1}^{n} (1 - 2\epsilon \lambda_i) \le 1$$

د - در گام اول صفر می‌شود $\rightarrow$

$$w_i^{t+1} = w_i^t - \frac{f(a)}{f'(a)_i} = w_i^t - \frac{w^{t^T} H w^t}{(H w^t)_i} = 0$$

۵ - محاسبات زیادی می‌خواهد و زمان و توان نیاز دارد، همچنین با کم شدن نرخ ممکن است گام‌های زیادی را داشته باشیم که مطلوب نیست و یا در saddle point ها متوقف شویم

سوال ۶-

الف - در این مدل می‌خواهیم خروجی دو ورودی را همزمان به یکدیگر میل بدهیم. هم چنین معیاری برای شباهت بین ورودی های با شد.

ب)

$$J = \|h_1 - h_2\|^2 + \underbrace{\|W\|_F^2}_{tr(W^TW)}$$

$$\rightarrow \nabla J_i = 2(1-h_1^2)(h_1-h_2) - 2(1-h_2^2)(h_1-h_2) + 2W$$

$$= 2(h_2^2 - h_1^2)(h_1 - h_2) + 2W$$

$$= -2(\underbrace{h_1^3}_{\tanh^3(Wx_1+b)} + h_2^3 + h_1h_2^2 - h_1^2h_2) + 2W$$

$$E = \sum_{i=1}^{b} \nabla J_i \rightarrow W^{t+1} \leftarrow W^t - \varepsilon E$$