سینا نمازی شماره دانشجویی: 400110154

۱.

الف) طبق تعریف سؤال ϵ محدوده‌ای است که error در آن در نظر گرفته نمی‌شود. طبق راهنمایی سؤال ξ_i و ξ_i^* نشان دهنده تخطی از این محدوده هستند، به شکلی که ξ_i میزان تخطی بیشتر بودن و ξ_i^* تخطی مربوط به کمتر بودن را نشان می‌دهد.

$$\Rightarrow \quad y_i - w^T x_i \leq \epsilon + \xi_i \quad \Rightarrow \quad y_i - w^T x_i - \epsilon \leq \xi_i$$

$$\Rightarrow \quad w^T x_i - y_i \leq \epsilon + \xi_i^* \quad \Rightarrow \quad w^T x_i - y_i - \epsilon \leq \xi_i^*$$

$$\xi_i \geq 0 \,,\quad \xi_i^* \geq 0 \,,\quad (i = 1, \dots, n)$$

بنابراین با توجه به روابط بالا می‌توان صورت primal را به شکل زیر نوشت:

Primal form:
$$\min_{\substack{w \in R^m \\ \xi_i \in R^n \\ \xi_i^* \in R^n}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{s.t.} \quad y_i - w^T x_i - \epsilon \leq \xi_i$$

$$w^T x_i - y_i - \epsilon \leq \xi_i^*$$

$$\xi_i \geq 0 \,,\quad \xi_i^* \geq 0 \quad (i = 1, \dots, n)$$

ب) با تعریف محدودیت‌ها و تابع هدف به شکل قبل، می‌توان تابع Lagrangian را به شکل زیر نوشت:

$$L(w,\xi,\xi^*,a,a^*,b,b^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n} a_i(\epsilon + \xi_i - y_i + w^T x_i)$$

$$- \sum_{i=1}^{n} a_i^*(\epsilon + \xi_i^* + y_i - w^T x_i) - \sum_{i=1}^{n}(b_i \xi_i + b_i^* \xi_i^*)$$

و $a_i, a_i^*, b_i, b_i^* \geqslant 0 \qquad (i=1,\dots,n)$

حال باید این عبارت را حل کنیم:

$$(w,\xi,\xi^*,a,a^*,b,b^*) = \underset{w,\xi,\xi^*}{\arg\min}\ \underset{a,a^*,b,b^*}{\arg\max}\ L(w,\xi,\xi^*,a,a^*,b,b^*)$$

$$\xRightarrow{\text{با عوض کردن } \min \text{ و } \max} = \underset{a,a^*,b,b^*}{\arg\max}\ \underset{w,\xi,\xi^*}{\arg\min}\ L(w,\xi,\xi^*,a,a^*,b,b^*)$$

در ادامه $\frac{\partial L}{\partial w}$ و $\frac{\partial L}{\partial \xi_i}$، $\frac{\partial L}{\partial \xi_i^*}$ را بدست می‌آوریم و برابر با صفر قرار می‌دهیم:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}(a_i - a_i^*)x_i = 0 \ , \ \underbrace{\frac{\partial L}{\partial \xi_i} = C - a_i - b_i = 0 \ , \ \frac{\partial L}{\partial \xi_i} = C - a_i^* - b_i^* = 0}$$

$$b_i \geqslant 0 \ , \ b_i = C - a_i \Rightarrow C - a_i \geqslant 0 \Rightarrow C \geqslant a_i$$

$$b_i^* \geqslant 0 \ , \ b_i^* = C - a_i^* \Rightarrow C - a_i^* \geqslant 0 \Rightarrow C \geqslant a_i^*$$

حال این موارد را در Lagrangian جایگزین می‌کنیم:

$$L(w,\xi,\xi^*,a,a^*,b,b^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i+\xi_i^*) - \sum_{i=1}^{n} a_i(\epsilon+\xi_i-y_i+w^T x_i)$$

$$- \sum_{i=1}^{n} a_i^*(\epsilon+\xi_i^*+y_i-w^T x_i) - \sum_{i=1}^{n}(b_i\xi_i + b_i^*\xi_i^*)$$

$$= \frac{1}{2}\left\|\sum_{i=1}^{n}(a_i-a_i^*)x_i\right\|^2 + \sum_{i=1}^{n}\xi_i\overbrace{(C-b_i-a_i)}^{0} + \sum_{i=1}^{n}\xi_i^*\overbrace{(C-b_i^*-a_i^*)}^{0}$$

$$-\epsilon\sum_{i=1}^{n}(a_i+a_i^*) + \sum_{i=1}^{n} y_i(a_i-a_i^*) + \sum_{i=1}^{n}(a_i^*-a_i)\underbrace{w^T x_i}_{(\sum_{k=1}^{n}(a_k-a_k^*)x_k)^T x_i}$$

$$= -\frac{1}{2}\sum_{i,j=1}^{n}(a_i-a_i^*)(a_j-a_j^*)x_i^T x_j - \epsilon\sum_{i=1}^{n}(a_i+a_i^*) + \sum_{i=1}^{n} y_i(a_i-a_i^*)$$

بنابراین مسئله dual به شکل روبه‌رو است:

$$\max_{a,a^*} -\frac{1}{2}\sum_{i,j=1}^{n}(a_i-a_i^*)(a_j-a_j^*)x_i^T x_j - \epsilon\sum_{i=1}^{n}(a_i+a_i^*) + \sum_{i=1}^{n} y_i(a_i-a_i^*)$$

$$\text{s.t.}\quad C \geq a_i, a_i^* \geq 0$$

پ) بله، چون تابع هدف به شکل quadratic است و محدودیت ها خطی هستند، می‌توان آن را با استفاده از یک QP Solver حل کرد.

ت) با توجه به شرط های KKT complemetry slackness داریم:

$$a_i(\epsilon + \xi_i - y_i + w^T x_i) = 0$$

$$a_i^*(\epsilon + \xi_i^* - y_i + w^T x_i) = 0$$

$$b_i \xi_i = 0$$

$$b_i^* \xi_i^* = 0$$

همانطور که مشخص است، اگر $a_i > 0$ باشد، $(\epsilon + \xi_i - y_i + w^T x_i) = 0$ است، بنابراین x_i یک support vector است. حال اگر $\xi_i = 0$ باشد روی مرز محدوده خطا قرار می‌گیرد و x_i یک margin support vector است ولی اگر $\xi_i > 0$ باشد، x_i یک non-margin support vector است. همین موارد به شکل مشابه برای a_i^* و ξ_i^* برقرار است.

ث) اگر $f(x)$ تابع پیشبینی ما باشد، آنگاه:

$$f(x) = w^T x \quad , \quad w = \sum_{i=1}^{n} (a_i - a_i^*) x_i$$

$$\Rightarrow f(x) = \sum_{i=1}^{n} (a_i - a_i^*) x_i^T x$$

بله می‌شود از تکنیک Kernel استفاده کرد و می‌توان عبارت بالا را به شکل زیر نوشت:

$$\Rightarrow f(x) = \sum_{i=1}^{n} (a_i - a_i^*) k(x_i, x)$$

ج) نقش ϵ برخلاف C است، هر ϵ کمتر باشد مدل ما تلاش بیشتری می‌کند خطاهای کوچک را fit کند بنابراین مدل ما پیچیده‌تر خواهد شد و Bias آن کم و Variance آن زیاد خواهد بود، ولی اگر ϵ بزرگ باشد Bias بیشتر و Variance کمتر خواهد بود.

-2

طبق Haussler's Theorem داریم:

$$m \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right) \Rightarrow m \geq \frac{1}{0{,}05}\left(\ln(1000) + \ln\left(\frac{1}{0{,}05}\right)\right)$$

$$\Rightarrow m \geq 198.06 \Rightarrow$$ حداقل 199 نمونه نیاز داریم

حال Haussler's Theorem را اثبات می‌کنیم.

فرض کنید در Hypothesis space ما k تا Hyposhesis وجود دارند که مقدار true error مربوط به آن ها از ϵ بیشتر است.

$$\text{error}(h_1, \dots, h_k) > \epsilon$$

Concept (true)

$$\Pr_{x \sim D}(h_i(x) = C(x)) \leq 1 - \epsilon$$

↳ taking sample

samples are independent

$$\Rightarrow \Pr(h_i \text{ is Consistent with Concept on } \underline{m} \text{ examples}) \leq (1-\epsilon)^m$$

$$\Rightarrow \Pr(\text{at least one of } h_i \text{ is consistent with concept on } m \text{ examples}) \leq k(1-\epsilon)^m$$

$$\leq |H|(1-\epsilon)^m$$

$$\leq |H|e^{-\epsilon m}$$

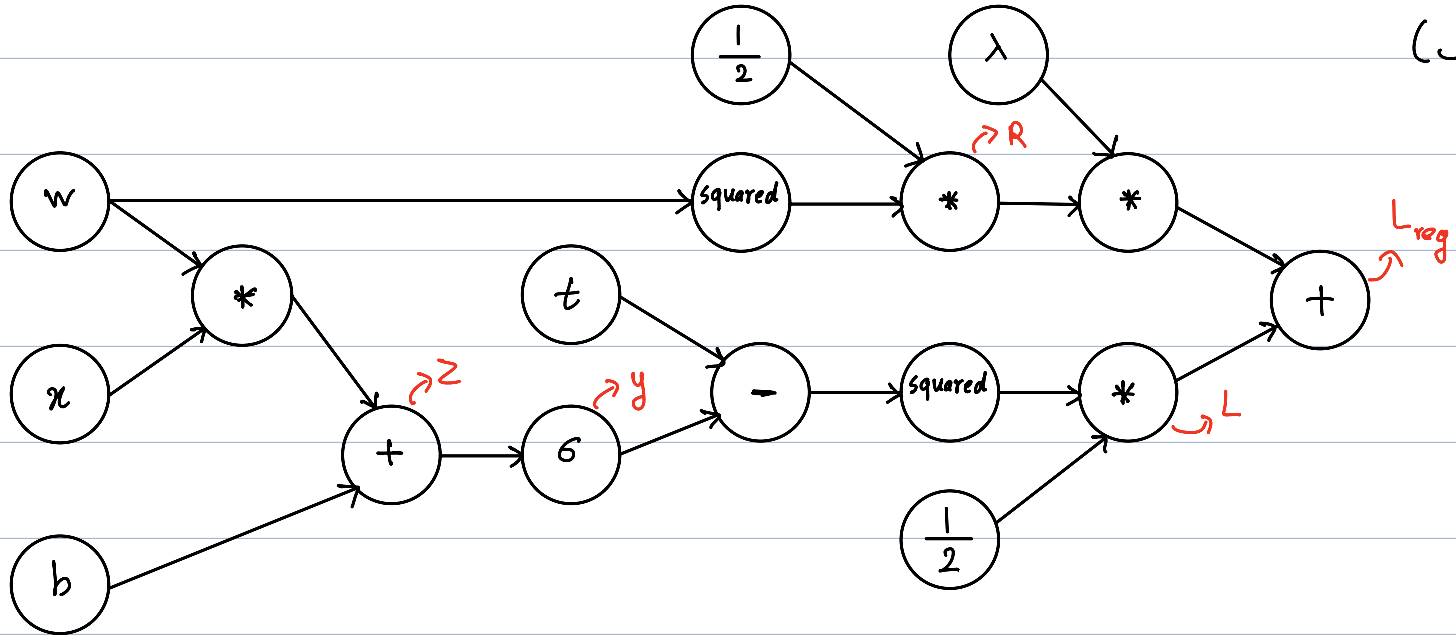
$$\Rightarrow |H|e^{-\epsilon m} \leq \delta$$

$$\Rightarrow \ln|H| - \epsilon m \leq \ln\delta$$

$$\Rightarrow m \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

3-

الف)



$$\frac{\partial L_{reg}}{\partial L} = 1 \ , \quad \frac{\partial L_{reg}}{\partial R} = \lambda \ , \quad \frac{\partial L}{\partial y} = y - t \ , \quad \frac{\partial L}{\partial t} = t - y \ , \quad \frac{\partial R}{\partial w} = w \ ,$$

$$\frac{\partial y}{\partial z} = \sigma(z)(1-\sigma(z)) \ , \quad \frac{\partial z}{\partial b} = 1 \ , \quad \frac{\partial z}{\partial w} = x \ , \quad \frac{\partial z}{x} = w$$

نسبت به پارامترها:

$$\frac{\partial L_{reg}}{\partial b} = \frac{\partial L_{reg}}{\partial L} \times \frac{\partial L}{\partial y} \times \frac{\partial y}{\partial z} \times \frac{\partial z}{\partial b} = 1 \times (y-t) \times \sigma(z)(1-\sigma(z)) \times 1$$

$$= (\sigma(wx+b)-t) \ \sigma(wx+b) \ (1-\sigma(wx+b))$$

$$\frac{\partial L_{reg}}{\partial w} = \frac{\partial L_{reg}}{\partial L} \times \frac{\partial L}{\partial y} \times \frac{\partial y}{\partial z} \times \frac{\partial z}{\partial w} + \frac{\partial L_{reg}}{\partial R} \times \frac{\partial R}{\partial w} = 1 \times (y-t) \times \sigma(z)(1-\sigma(z)) \times x + \lambda w$$

$$= (\sigma(wx+b)-t) \ \sigma(wx+b) \ (1-\sigma(wx+b))x + \lambda w$$

ب) اگر پارامترهای شبکه در ابتدا با مقادیر بزرگ مقداردهی شوند، ممکن است گرادیان های محاسبه شده بسیار بزرگ شوند و اصطلاحاً به مشکل exploding gradients برخورد کردیم. در این مشکل، پارامترهای شبکه به ویژه برای لایه‌های اول، با هر بار update کردن زیاد تغییر می‌کنند و converge کردن آن‌ها با مشکل مواجه شود.

برعکس اگر پارامترهای شبکه در ابتدا با مقادیر بسیار کوچک مقداردهی شوند، ممکن است با مشکل vanishing gradients مواجه شویم که در آن گرادیان های محاسبه‌شده برای لایه‌های اولیه به ویژه در شبکه‌های خیلی عمیق بسیار کوچک خواهد بود و این لایه‌ها در طول آموزش، تغییر زیادی نمی‌کنند و یاد نمی‌گیرند.

اگر مقداردهی به پارامترهای شبکه به صورت رندوم نباشد، ممکن است بین بعضی نورون‌های شبکه تقارن بوجود بیاید و این نورون ها همه‌شان ویژگی‌های یکسانی را یاد بگیرند. برای مثال فرض کنید همه نورون های یک لایه در ابتدا وزن یکسانی داشته باشند، در ادامه همه این نورون ها ویژگی‌های یکسانی را یاد می‌گیرند که به دلیل redundancy بوجود آمده، شبکه ما efficiency کمتری خواهد داشت.

ج)

مقادیر درنظر گرفته شده: $x=2$, $w=0.1$, $b=0.2$, $t=1$, $\lambda=0.1$, $\text{Learning Rate}=0.1$

$$\frac{\partial L_{reg}}{\partial b} = (\sigma(0.4)-1)\ \sigma(0.4)\ (1-\sigma(0.4)) \simeq -0.096419$$

$$\frac{\partial L_{reg}}{\partial w} = (\sigma(0.4)-1)\ \sigma(0.4)\ (1-\sigma(0.4))\times 2 + 0.1\times 0.1 \simeq -0.182839$$

$$b_{new} = b - \text{Learning Rate} \times \frac{\partial L_{reg}}{\partial b} \simeq 0.2096419$$

$$w_{new} = w - \text{Learning Rate} \times \frac{\partial L_{reg}}{\partial w} \simeq 0.1182839$$