Subject
Date

۱ـ الف) مفهوم این تابع loss این است که به مدل اجازه می‌دهد، تا مقدار ε از خطا چشم‌پوشی کند،

یعنی فاصله ε از لیبل واقعی را به عنوان جواب صحیح درنظر می‌گیرد که با توجه به راهنمایی‌های صورت سوال، loss

را بصورت زیر تعریف می‌کنیم:

$$\begin{cases} y > f(x) \rightarrow y - f(x) - \varepsilon \equiv y - f(x) - \varepsilon \leq \xi_i \\ y < f(x) \rightarrow f(x) - y - \varepsilon \equiv -y + f(x) - \varepsilon \leq \xi_i^* \end{cases}$$

و ξ_i و ξ_i^* نامنفی هستند، پس می‌توان عبارت min را بصورت زیر بازنویسی کرد:

$$\min_{w \in R^n, \xi \in R^n, \xi^* \in R^n} \frac{1}{2} \|w\| + C \sum_1^n (\xi_i + \xi_i^*)$$

درواقع صورت مسئله را به مسئله soft margin با متغیرهای slack جدید تبدیل کردیم.

ب) فرم اصلی تکنیک ضرایب لاگرانژ بصورت زیر است:

$$\begin{aligned} &\min f(x) \\ &\text{s.t } g_i(x) \leq 0 \quad i = 1, \dots, m \end{aligned} \equiv \begin{aligned} &\min_x \max_{\alpha_i \geq 0} \mathcal{L}(x, \alpha) \\ &\max_{\alpha_i \geq 0} \min_x \mathcal{L}(x, \alpha) \end{aligned}, \quad \mathcal{L}(x, \alpha) = f(x) + \sum_1^m \alpha_i g_i(x)$$

$$\begin{aligned} &\min \frac{1}{2}\|w\| + C\sum_1^n (\xi_i + \xi_i^*) \\ &\text{s.t } y - f(x) - \varepsilon \leq \xi_i \\ &\quad -y + f(x) - \varepsilon \leq \xi_i^* \\ &\quad \xi_i, \xi_i^* \geq 0 \end{aligned} \longrightarrow \begin{aligned} &\min \frac{1}{2}\|w\| + C\sum_1^n (\xi_i + \xi_i^*) \\ &\text{s.t } y - f(x) - \varepsilon - \xi_i \leq 0 \rightarrow g^1, \alpha \\ &\quad -y + f(x) - \varepsilon - \xi_i^* \leq 0 \rightarrow g^2, \alpha^* \\ &\quad -\xi_i \leq 0 \rightarrow g^3, \beta \\ &\quad -\xi_i^* \leq 0 \rightarrow g^4, \beta^* \end{aligned}$$

$$\Rightarrow \mathcal{L}(w, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*) = \frac{1}{2}\|w\| + C\sum_1^n (\xi_i + \xi_i^*) + \sum_1^n (\alpha_i g_i^1 + \alpha_i^* g_i^2 + \beta_i g_i^3 + \beta_i^* g_i^4)$$

PAPCO

$$= \frac{1}{2}\|w\| + C\sum_{1}^{n}(\xi_i+\xi_i^*) + \sum_{1}^{n}\Big(\alpha_i(y_i - w^T x_i - \varepsilon - \xi_i) + \alpha_i^*(y_i + w^T x - \varepsilon - \xi_i^*) - \beta_i\xi_i - \beta_i^*\xi_i^*\Big) = L$$

حال باید از این عبارت برحسب w گرادیان و برحسب متغیرهای slack مشتق بگیریم: (L تابع لاگرانژ است نه Loss)

$$\nabla_w L = w + \nabla_w \sum_{1}^{n}(\alpha_i^* - \alpha_i)w^T x_i = w + \sum_{1}^{n}(\alpha_i^* - \alpha_i)x_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \ , \ \frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \beta_i^* = 0 \rightarrow \alpha_i, \alpha_i^* \leq C$$

به شروط افزوده خواهد شد

$$\Rightarrow w = \sum_{1}^{n}(\alpha_i - \alpha_i^*)x_i \ , \ \beta_i = C - \alpha_i \ , \ \beta_i^* = C - \alpha_i^* \ :$$

قرینه ضرایب فعلی ξ_i, ξ_i^*، در نتیجه ضرایب صفر می‌شوند

$$= \frac{1}{2}\Big\|\sum_{1}^{n}(\alpha_i - \alpha_i^*)x_i\Big\| - \sum_{1}^{n}(\alpha_i - \alpha_i^*)w^T x_i + \sum_{1}^{n}(\alpha_i - \alpha_i^*)y_i$$

$$- \varepsilon\sum_{1}^{n}(\alpha_i + \alpha_i^*)$$

جای w عبارتی که از $\nabla_w L$ بدست آمد را اگر جایگذاری کنیم، به فرمی شبیه به $\frac{1}{2}\|\sum(\alpha_i - \alpha_i^*)x_i\|$ خواهیم رسید که

این دو عبارت نتیجتاً ساده می‌شوند:

$$\underbrace{\frac{1}{2}\|\sum(\alpha_i - \alpha_i^*)x_i\|}_{} - \sum(\alpha_i - \alpha_i^*)\Big(\sum(\alpha_j - \alpha_j^*)x_j\Big)^T x_i$$

$$\frac{1}{2}\sum\sum(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^T x_j$$

$$= -\frac{1}{2}\sum\sum(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^T x_j$$

$$\Rightarrow L = -\frac{1}{2}\sum\sum(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i^T x_j + \sum_{1}^{n}(\alpha_i - \alpha_i^*)y_i - \varepsilon\sum_{1}^{n}(\alpha_i - \alpha_i^*)$$

$$\Rightarrow \max_{\alpha, \alpha^*} L \quad s.t: \ 0 \leq \alpha_i, \alpha_i^* \leq C$$

Subject
Date

پ) در حالت اصلی مجددا شروط خطی هستند و تابع هدف درجه دو است؛ پس مسئله به شکل QP معتبر و قابل حل است.

ت) حالت support vector برای یک نقطه وقتی رخ می‌دهد که روی مرز باشد که اینجا در حالتی رخ می‌دهد که $|y-f(x)|=\varepsilon$، طبق KKT می‌دانیم برای هر شرط، $\alpha_i g_i$ صفر است؛ یعنی یا $\left.\begin{matrix} g_i>0,\ \alpha_i=0 \\ g_i=0,\ \alpha_i>0 \end{matrix}\right\}$

$\alpha_i g_i = 0 \longrightarrow \alpha_i(y_i - w^T x_i - \varepsilon - \xi_i) = 0 \ , \ \beta_i \xi_i = 0$

$\alpha_i^* g_i^* = 0 \longrightarrow \alpha_i^*(-y_i + w^T x_i - \varepsilon - \xi_i^*) = 0 , \beta_i^* \xi_i^* = 0$

چون برای support vector ها ξ_i و ξ_i^* برابر صفر است، بدین گونه g_i صفر خواهد بود و $\alpha_i > 0$ پس:

$\xi_i = 0 \ , \ \alpha_i = 0$ یا $\xi_i^* = 0 \ , \ \alpha_i^* = 0 \longrightarrow$ support vector

$\xi_i > 0 \ , \ \alpha_i > 0$ یا $\xi_i^* > 0 \ , \ \alpha_i^* > 0 \longrightarrow$ non-support vector

ث) در بخش ب) نشان دادیم با جایگذاری مقدار w که در لاگرانژین فرم اصلی بدست آمد، $x^T w$ به صورت $w^T x = \sum(\alpha_i - \alpha_i^*) x_i^T x$ درآمد که برابر است با $(\sum(\alpha_i - \alpha_i^*) x_i)^T x$ که به صورت ضرب داخلی نمونه‌ها ظاهر شده، پس می‌توان آن را به صورت کرنلی نوشت و از این تکنیک استفاده کرد.

ج) پیشتر توضیح دادیم ε اندازه margin و فاصله هر نمونه را مشخص می‌کند پس با افزایش آن، به مدل آزادی عمل می‌دهیم و بایاس افزایش و واریانس کاهش می‌یابد. با افزایش C بنابراین مدل در صورت افزایش متغیرهای slack زیان می‌شود و تأثیری معکوس افزایش ε دارد پس واریانس افزایش یافته و بایاس کاهش می‌یابد.

۲- ابتدا خواسته‌ی مسئله را به صورت عبارتی شبیه به hoeffding می‌نویسیم،

$$P[\exists h \in H : \bar{\mathcal{E}}(h) = 0 \wedge \mathcal{E}(h) > \varepsilon] \leq 1 - \frac{95}{100} = \frac{5}{100}$$

این عبارت یعنی می‌خواهیم اگر $h \in H$ وجود داشت که تطابق با train دارد، احتمال اینکه با داده‌ی دیده نشده خطای بزرگتر از $\varepsilon = \frac{5}{100}$ داشته باشد از $\delta = 1 - \frac{95}{100} = \frac{5}{100}$ کمتر باشد؛ حال با این معادله تلاش می‌کنیم برای m، حد پایینی بیابیم:

اگر $h_i \in H$ را در نظر بگیریم می‌خواهیم احتمال نرخ خطا برای $\mathcal{E}(h_i)$ را حساب کنیم،

$$P[error_{true}(h_i) > \varepsilon] \leq 1 - \varepsilon \xrightarrow{m \text{ نمونه مستقل}} P[error_{true}(h_i) > \varepsilon] \leq (1-\varepsilon)^m$$

که اگر این احتمال را برای $\forall h \in H$ جمع کنیم،

$$P[\exists h \in H : \bar{\mathcal{E}}(h) = 0 , \mathcal{E}(h) > \varepsilon] \leq \sum_{1}^{1000} P[error_{true}(h_i) > \varepsilon] \leq 1000(1-\varepsilon)^m$$

و می‌خواهیم این عبارت از δ کوچکتر باشد،

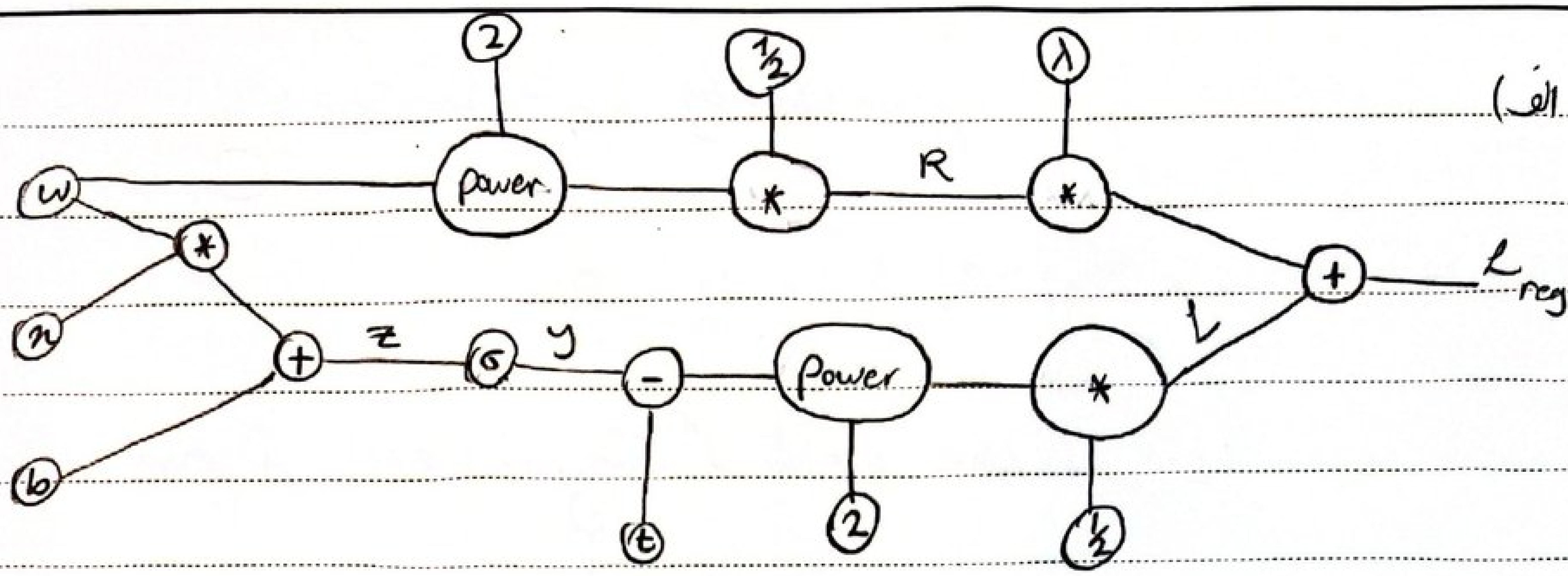
$$1000(1-\varepsilon)^m \leq \frac{5}{100} \xRightarrow{\ln} \ln 1000 + m \ln(1-\varepsilon) \leq \ln\frac{5}{100} \rightarrow m \geq \frac{\ln 1000 - \ln\frac{5}{100}}{-\ln\frac{95}{100}} \simeq 193.07$$

$$\Rightarrow m \geq 194$$

Subject

Date

۳ـ الف)



$$\frac{\partial L_{reg}}{\partial w} = \frac{\partial L}{\partial w} + \frac{\partial \lambda R}{\partial w} = \frac{\partial \frac{1}{2}(y-t)^2}{\partial w} + \lambda \frac{\partial \frac{1}{2} w^2}{\partial w}$$

$$\frac{1}{2}\frac{\partial}{\partial w}(y-t)^2 = \frac{1}{2}\frac{\partial}{\partial w}(y^2+t^2-2yt) = \frac{1}{2}\frac{\partial}{\partial w}y^2 - \frac{\partial}{\partial w}yt = A$$

$$\frac{\partial yt}{\partial w} = t\frac{\partial y}{\partial w} = t\frac{\partial z}{\partial w}\frac{\partial y}{\partial z} = t x y(1-y)$$

$$\frac{\partial}{\partial w}y^2 = 2y\frac{\partial y}{\partial w} = 2y\Big(\underbrace{\frac{\partial y}{\partial z}}_{y(1-y)}\underbrace{\frac{\partial z}{\partial w}}_{x}\Big) = 2y^2(1-y)x \xRightarrow{\times\frac{1}{2}} y^2(1-y)x$$

$$\Rightarrow A = y^2(1-y)x - ty(1-y)x = y(1-y)x(y-t)$$

$$\frac{\partial L_{reg}}{\partial w} = \boxed{y(1-y)x(y-t) + \lambda w}$$

$$\frac{\partial L_{reg}}{\partial b} = \frac{\partial L}{\partial b} + \underbrace{\frac{\partial \lambda R}{\partial b}}_{0} = \frac{1}{2}\frac{\partial}{\partial b}y^2 - \frac{\partial}{\partial b}yt = B$$

$$\frac{\partial yt}{\partial b} = t\frac{\partial y}{\partial b} = t\underbrace{\frac{\partial z}{\partial b}}_{1}\underbrace{\frac{\partial y}{\partial z}}_{y(1-y)} = ty(1-y)$$

$$\frac{\partial}{\partial b}y^2 = 2y\frac{\partial y}{\partial b} = 2y\,y(1-y)$$

$$\Rightarrow B = y^2(1-y) - ty(1-y) = y(1-y)(y-t) \Rightarrow \frac{\partial}{\partial b}L_{reg} = \boxed{y(1-y)(y-t)}$$

Subject

Date

ب) اگر پارامترها تصادفی نباشند، بین نورون ها تقارن رخ می دهد و نورون های مختلف نمی‌توانند فیچرهای مختلفی یاد بگیرند

و در update شدن وزن ها نیز این تقارن مشهود بوده و مدل در پیش‌بینی pattern های پیچیده ضعیف عمل خواهد کرد

اگر پارامترها کوچک نباشند، امکان دارد در الگوریتم gradiant desc ، convergance رخ ندهد و گام های گرادیان بزرگ باشند

(مانند اتفاقی که هنگام بزرگ بودن learning rate رخ می‌دهد) نمی‌توان به جواب بهینه دست پیدا کرد.

ج) صورت update rule الگوریتم را می‌نویسیم:

$$w^{i+1} = w^{i} - 0.1 \frac{\partial L_{reg}}{\partial w}$$

$$b^{i+1} = b^{i} - 0.1 \frac{\partial L_{reg}}{\partial b}$$

$$x = 2,\ w = 1.5,\ b = 1 \text{ و } t = 2 \Rightarrow \begin{cases} z = 4 \\ y = 0.982 \end{cases} \text{ و } \lambda = 0.3$$

$$\Rightarrow w^{i+1} = 1.5 - \frac{1}{10} \times \left(\underbrace{(0.982)(0.018)(2)}_{-0.036}(-1.018) + \underbrace{0.3 \times 1.5}_{0.45} \right) \simeq 1.458$$

$$0.414$$

$$b^{i+1} = 1 - \frac{1}{10} \times \left(0.982 \times 0.018 \times -1.018 \right) \simeq 1.001$$