به نام خدا

پاسخ تمرین پنجم یادگیری ماشین

امین اسدی

۸۱۰۱۹۶۴۱۰

پاسخ سوالات ۱ تا ۶

سوال اول)

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}_{3\times1} \quad X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix}_{3\times2} \quad Y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}_{3\times3}$$

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\ \alpha^T 1_n - \frac{1}{2}\alpha^T YGY\alpha$$

$$G = X.X^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

$$YGY = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \rightarrow \hat{\alpha} = \underset{\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}}{\operatorname{argmax}} \sum_{1}^{3}\alpha_i - \frac{1}{2}\left((\alpha_1 \alpha_2 \alpha_3)\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}\right)$$

$$\rightarrow \hat{\alpha} = \underset{\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}}{\operatorname{argmax}} \left[\alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2}\left(\alpha_1(\alpha_1 + \alpha_3) + \alpha_2^2 + \alpha_3(\alpha_1 + \alpha_3)\right)\right]$$

$$\text{subject to } \sum_{i=1}^{3}\alpha_i y_i = 0 \rightarrow \boxed{\alpha_1 + \alpha_2 = \alpha_3}$$

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \left[2(\alpha_1 + \alpha_2) - \frac{1}{2}\left(\alpha_1(2\alpha_1 + \alpha_2) + \alpha_2^2 + (\alpha_1 + \alpha_2)(2\alpha_1 + \alpha_2)\right)\right]$$

$$= \underset{\alpha}{\operatorname{argmax}} \left[2\alpha_1 + 2\alpha_2 - \frac{1}{2}\left(\underbrace{2\alpha_1^2 + \alpha_1\alpha_2 + \alpha_2^2 + 2\alpha_1^2 + \alpha_1\alpha_2 + 2\alpha_1\alpha_2 + \alpha_2^2}_{4\alpha_1^2 + 2\alpha_2^2 + 4\alpha_1\alpha_2}\right)\right]$$

$$= \underset{\alpha}{\operatorname{argmax}} \left[\underbrace{2\alpha_1 + 2\alpha_2 - 2\alpha_1^2 - \alpha_2^2 - 2\alpha_1\alpha_2}_{L}\right]$$

بررسی نقطه بحرانی

$$\begin{cases} \dfrac{\partial L}{\partial \alpha_1} = 0 \rightarrow 2 - 4\alpha_1 - 2\alpha_2 = 0 \\[2ex] \dfrac{\partial L}{\partial \alpha_2} = 0 \rightarrow 2 - 2\alpha_1 - 2\alpha_2 = 0 \end{cases} \longrightarrow \begin{matrix} \alpha_1 = 0 \\ \alpha_2 = 1 \\ \alpha_3 = 0 + 1 = 1 \end{matrix}$$

$$\frac{\partial^2 L}{\partial\alpha_1 \partial\alpha_1} \times \frac{\partial^2 L}{\partial\alpha_2 \partial\alpha_2} - \left(\frac{\partial^2 L}{\partial\alpha_1 \partial\alpha_2}\right)^2 > 0 \ , \ \frac{\partial^2 L}{\partial\alpha_1 \partial\alpha_1} < 0 \rightarrow \text{maximum}$$

$$\rightarrow \alpha = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

$$w = \sum_{i=1}^{3} \alpha_i y_i x_i = 0(1,0) + 1(0,1) + 1(1,0) = (1,1)$$

$$\downarrow$$

$$w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

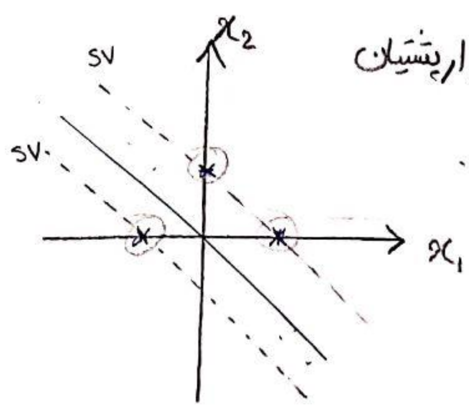
$$b = -\frac{\min_{i: y^{(i)}=1} w^T x^{(i)} + \max_{i: y^{(i)}=-1} w^T x^{(i)}}{2}$$

$$= -\frac{\min\left([1,1]\begin{bmatrix} 1 \\ 0 \end{bmatrix}, [1,1]\begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) + [1,1]\begin{bmatrix} -1 \\ 0 \end{bmatrix}}{2}$$

$$= -\frac{1-1}{2} = 0$$

$\longrightarrow$ معادله خط جداساز : $\overset{1}{\cancel{w_1}} x_1 + \overset{1}{\cancel{w_2}} x_2 + \overset{0}{\cancel{b}} = 0 \longrightarrow \boxed{x_2 = -x_1}$

خط های حاشیه : $\begin{cases} \overset{1}{\cancel{w_1}} x_1 + \overset{1}{\cancel{w_2}} x_2 + \overset{0}{\cancel{b}} = 1 \rightarrow \boxed{x_2 = -x_1 + 1} \\ \overset{1}{\cancel{w_1}} x_1 + \overset{1}{\cancel{w_2}} x_2 + \overset{0}{\cancel{b}} = -1 \rightarrow \boxed{x_2 = -x_1 - 1} \end{cases}$



در نتیجه هر سه نقطه بردار پشتیبان (SV) می باشند.

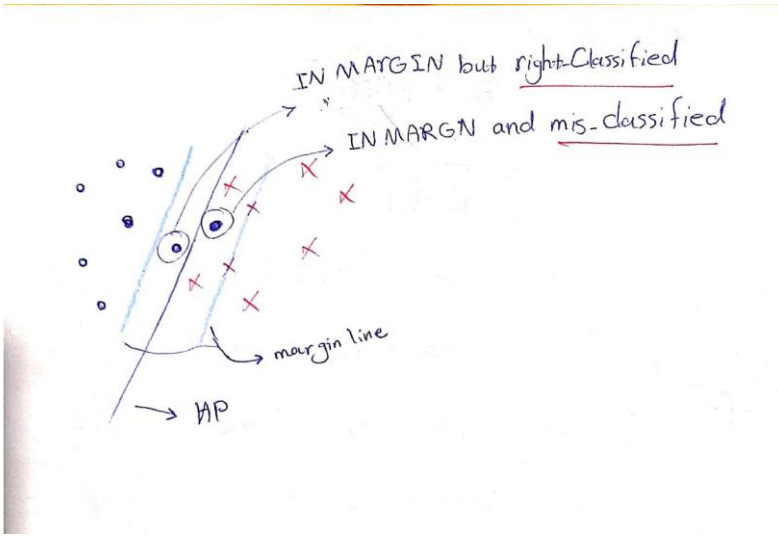
سوال دوم)

۱)

مساله‌ی ثانویه به ما این امکان را می‌دهد که مفهوم kernel را وارد مساله کنیم تا برای حل مسائلی که به طور خطی جداپذیر نیستند(یا حتی در مسائل جداپذیر خطی برای رفع مشکل واریانس زیاد) از kernel استفاده کنیم. همچنین مساله‌ی ثانویه به ما امکان بهینه انجام دادن محاسبات لازم را می‌دهد زیرا ابتدا آلفا ها را می‌یابیم که به جز برای SV ها، برای بقیه نقاط برابر 0 هستند و ترم‌های مربوط به آن‌ها از محاسبات حذف می‌شوند و کافیست برای SV ها محاسبات انجام شوند که به دلیل اینکه تعداد SV ها معمولا می‌تواند بسیار محدود باشد، محاسبات بسیار ساده تر خواهند شد.

۲)

از hard margin برای حل مسائلی به کار می‌رود که به صورت خطی جداپذیر هستند که منطقا در آن ها هیچ نقطه‌ای mis-classify نمی‌شود. در دو حالت از soft-margin ها استفاده می‌شود: ۱- در مسائلی که مساله به صورت خطی جداپذیر نیست و بنابراین انعطاف بیشتری به margin می‌دهیم و در نتیجه تعدادی از نقاط به اشتباه طبقه‌بندی خواهند شد.۲- در حالتی که مساله به صورت خطی جداپذیر است اما حاضر هستیم به خاطر bias-variance tradeoff تعدادی کمی از نقاط به اشتباه طبقه‌بندی شوند اما در عوض generalization طبقه‌بند را بالا ببریم.



۳)

این ضریب کنترل‌کننده‌ی trade-off بین خطا و حاشیه است. وقتی این ضریب افزایش می‌یابد در واقع به خطا بهای بیشتری می‌دهیم و در واقع می‌خواهیم خطا را کمتر کنیم بنابراین در این حالت خطا را کمینه می‌کنیم اما حاشیه ممکن است کوچک باشد. اگر این ضریب را کم کنیم به بهای mis-classify شدن برخی نقاط، حاشیه را بیشینه می‌کنیم.

سوال سوم)

کرنل‌ها توابعی هستند که برای تبدیل داده‌ها از یک فضای ویژگی به یک فضای ویژگی با بُعد بالاتر استفاده می‌شوند. علت بردن داده‌ها به فضای بُعد بالاتر این است که ممکن است در بُعد پایین داده‌ها به صورت خطی جداپذیر نباشند و در نتیجه بُعد فضای را افزایش می‌دهیم تا داده‌ها unfold شوند و تا جای ممکن به صورت خطی جداپذیر شوند. یک روش این است که از هر نقطه‌را به فضای جدید منتقل کنیم و سپس محاسبات را در فضای جدید انجام دهیم، امّا این روش خصوصا وقتی ابعاد فضای جدید خیلی زیاد است، بسیار هزینه‌بر یا حتی غیر ممکن خواهد بود. اینجاست که از کرنل‌ها استفاده می‌کنیم. با استفاده از کرنل‌ها می‌توانیم به جای انجام محاسبات در بُعد بالا، محاسبات را در بُعد پایین(اولیه) با هزینه کم انجام دهیم و به همان نتیجه دست پیدا کنیم.

اثبات رابطه کاشی–شوارتز:

می‌دانیم کرنل-مارتیس(Kernel Matrix) متناظر با هر زیر مجموعه ناتهی از نقاط باید PSD(Positive Semi-Difinite) باشد. در نتیجه کرنل-مارتیس مربوط به دو نقطه x و y که به صورت زیر است باید PSD باشد:

$$\begin{bmatrix} K(x,x) & K(x,y) \\ K(x,y) & K(y,y) \end{bmatrix}$$

طبق تعریف PSD، باید همه مقادیر ویژه این ماتریس نامنفی باشند و در نتیجه دترمینان آن نیز باید نامنفی باشد پس داریم:

$$K(x,x)\,K(x,x) - K(x,x)^2 \geq 0 \Rightarrow K(x,x)^2 \leq K(x,x)K(x,x)$$

فرض کنید که $x_{m,i}$ نشان‌دهنده مولفه i ام داده m ام باشد. در این صورت طبق تعریف کرنل داریم:

$$\sum_{m=1}^{Q}\sum_{n=1}^{Q} K_{\emptyset}(x_m, x_n) = \sum_{m=1}^{Q}\sum_{n=1}^{Q} \langle \emptyset(x_m), \emptyset(x_n) \rangle$$

$$= \sum_{m=1}^{Q}\sum_{n=1}^{Q} \langle \emptyset\begin{pmatrix} x_{m,0} \\ \vdots \\ x_{m,d_1} \end{pmatrix}, \emptyset\begin{pmatrix} x_{n,0} \\ \vdots \\ x_{n,d_1} \end{pmatrix} \rangle$$

$$= \sum_{m=1}^{Q}\sum_{n=1}^{Q} \langle \begin{pmatrix} x'_{m,0} \\ \vdots \\ x'_{m,d_2} \end{pmatrix}, \begin{pmatrix} x'_{n,0} \\ \vdots \\ x_{n,d'_2} \end{pmatrix} \rangle$$

$$= [{x'_{1,0}}^2 + {x'_{2,0}}^2 + \cdots + {x'_{Q,0}}^2 + 2(x'_{1,0}x'_{2,0} + \cdots + x'_{Q-1,0}x'_{Q,0})] + \cdots + [{x'_{1,d_2}}^2 + {x'_{2,d_2}}^2 + \cdots + {x'_{Q,d_2}}^2 + 2(x'_{1,d_2}x'_{2,d_2} + \cdots + x'_{Q-1,d_2}x'_{Q,d_2})]$$

$$= \left(x'_{1,0} + \cdots + x'_{Q,0}\right)^2 + \cdots + \left(x'_{1,d_2} + \cdots + x'_{Q,d_2}\right)^2$$

$$\rightarrow \frac{1}{Q}\sqrt{\sum_{m=1}^{Q}\sum_{n=1}^{Q} K_{\emptyset}(x_m, x_n)} = \sqrt{\frac{1}{Q^2}\sum_{m=1}^{Q}\sum_{n=1}^{Q} K_{\emptyset}(x_m, x_n)}$$

$$= \sqrt{\left(\frac{x'_{1,0} + \cdots + x'_{Q,0}}{Q}\right)^2 + \cdots + \left(\frac{x'_{1,d_2} + \cdots + x'_{Q,d_2}}{Q}\right)^2} = \|\mu_{\emptyset}\|$$

و اثبات به پایان می‌رسد.

سوال چهارم)

1) $K(x,y) = f(x)\, K_1(x,y)\, f(y)$

$\forall c \in R^n$:

$$c^T K(x,y) c = \sum_{i,j} c_i\, K(x_i,y_j) c_j = \sum_{i,j} \underbrace{c_i\, f(x_i)}_{a_i} K_1(x_i,y_j) \underbrace{f(y_j) c_j}_{a_j}$$

$$= \sum_{i,j} a_i K_1(x_i,y_j) a_j = A^T \underbrace{K_1}_{\geq 0} A$$

$$\rightarrow K_1(x,y) \geq 0 \rightarrow \text{Valid}$$

3) $K(x,y) = K_1(x,y) + K_2(x,y)$

$\forall c \in R^n$:

$$c^T K(x,y) c = \sum_{i,j} c_i \left(K_1(x_i,y_j) + K_2(x_i,y_j)\right) c_j$$

$$= \sum_{i,j} c_i K_1(x_i,y_j) c_j + \sum_{i,j} c_i K_2(x_i,x_j) c_j$$

$$= \underbrace{c^T K_1 c}_{\geq 0} + \underbrace{c^T K_2 c}_{\geq 0} \geq 0 \rightarrow \text{Valid}$$

4) $K(x,y) = K_1(x,y)\, K_2(x,y)$

$\forall c \in R^n$:

$$c^T K(x,y) c = \sum_{i,j} c_i\, K_1(x_i,y_j) K_2(x_i,y_j) c_j$$

$$= \sum_{i,j} K_1(x_i,y_j)\, c_i K_2(x_i,y_j) c_j \longrightarrow$$ طبق بخش قبل ≥ 0 است

$$= \sum_{i,j} K_1(x_i,y_j)\, K_2'(x_i,y_j)$$

$$= trace(K_1 K_2'^T) \geq 0 \rightarrow \text{Valid}$$

$\hookleftarrow$ حاصل جمع مقادیر ویژه ≥ 0 است

سوال پنجم)

می‌دانیم فاصله دو نقطه $\underline{u}$ و $\underline{v}$ را می‌توان بر حسب ضرب داخلی به صورت زیر بیان کرد:

$$d^2(\underline{u},\underline{v}) = \|\underline{u}-\underline{v}\| = \langle \underline{u}-\underline{v}, \underline{u}-\underline{v} \rangle = \langle \underline{u},\underline{u} \rangle + \langle \underline{v},\underline{v} \rangle - 2\langle \underline{u},\underline{v} \rangle$$

و نیز طبق تعریف کرنل داریم:

$$K_{\emptyset}(x_1,x_2) = \langle \emptyset(x_1), \emptyset(x_2) \rangle$$

$$\langle \emptyset(x_1), \emptyset(x_2) \rangle = K(x_1,x_2) = K([1,1]^T,[3,4]^T) = \exp\left(-\frac{(3-1)^2+(4-1)^2}{2}\right) = \exp(-6.5)$$

$$\langle \emptyset(x_1), \emptyset(x_1) \rangle = K(x_1,x_1) = K([1,1]^T,[1,1]^T) = \exp\left(-\frac{(1-1)^2+(1-1)^2}{2}\right)$$
$$= \exp(0) = 1$$

$$\langle \emptyset(x_2), \emptyset(x_2) \rangle = K(x_2,x_2) = K([3,4]^T,[3,4]^T) = \exp\left(-\frac{(3-3)^2+(4-4)^2}{2}\right) = \exp(0) = 1$$

$$\rightarrow\ d^2(\emptyset(x_1),\emptyset(x_2)) = 1 + 1 - 2\exp(-6.5) \cong 2 - 2\times\ 0.001 = 1.998 \rightarrow$$
$$d = \sqrt{1.998}$$

قسمت دوم سوال ۵:

کافیست تعداد تک‌جمله‌ای‌های حاصل کِرنل را بدست آوریم. با توجه به اینکه تعداد ابعاد اولیه برابر d است می‌توان نوشت:

$$x = [x_1, x_2, \dots, x_d]^T$$
$$y = [y_1, y_2, \dots, y_d]^T$$
$$K(x,y) = (x^Ty+1)^2 = (x_1y_1 + x_2y_2 + \cdots + x_dy_d + 1)^2$$

به وضوح حاصل کِرنل برابر مجموع تک‌جمله‌ای‌هایی به فرم $x_1^{c_1}y_1^{c_1}\dots x_d^{c_d}y_d^{c_d}1^{c_{d+1}}$ خواهد بود که:

$$c_1 + c_2 + \cdots + c_d + c_{d+1} = 2 \qquad 0 \le c_1, c_2, \dots, c_d, c_{d+1}$$

پس تعداد این تک‌جمله‌ای ها برابر خواهد بود با تعداد جواب های معادله بالا که می‌دانیم برابر است با:

$$\binom{d+1+2-1}{2} = \binom{d+2}{2} = \frac{(d+1)(d+2)}{2}$$

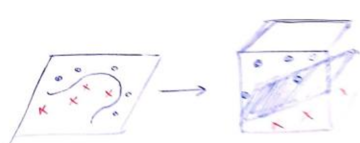
سوال ششم)

۱) در رویکرد generative ابتدا توزیع joint متغیرها بدست آمده و سپس از روی آن توزیع‌های شرطی محاسبه می‌شوند(در واقع تابع discriminant را با استفاده از تخمین توزیع‌های مربوط به هر کلاس تخمین می‌زنند) اما در رویکرد discriminative به صورت مستقیم تابع توزیع‌های شرطی محاسبه می‌شوند(در واقع تابع discriminant را به طور مستقیم تخمین می‌زنند) رویکرد generative نیاز به داده‌های زیادی دارد بنابراین رویکرد discriminative عملی تر است.

۲) مزیت رویکرد one-vs-rest این است که تعداد طبقه‌بند هایی که استفاده می‌کنیم برابر تعداد کلاس‌ها (یا یکی کمتر از آن‌ها) است و در نتیجه با افزایش تعداد کلاس‌ها، تعداد طبقه‌بندهای مورد نیاز به صورت خطی افزایش می‌یابد و بنابراین هزینه کمتری دارد. اگر در هر مرحله یکی از کلاس‌ها را کنار بگذاریم ایراد این روش این است نتیجه به ترتیب کلاس‌ها حساس خواهد بود. اگر در مرحله کلاس را کنار نگذاریم ایراد این خواهد بود که یا بعضی از نواحی به بیش از یک کلاس دسته‌بندی خواهند شد و یا بعضی نواحی به هیچ کلاسی دسته‌بندی نخواهند شد(ambiguous area). رویکرد one-vs-another معایب ذکر شده را ندارد اما یک عیب آن این است که برای هر دو کلاس یک طبقه بند نیاز دارد(در کل $\binom{C}{2}$ تا طبقه بند برای C کلاس). معمولا روش one-vs-another عملکرد بهتری نسبت به one-vs-rest دارد. در ماشین خطی نیز از همان اول مساله را چند-کلاسه حل می‌کنیم و در نهایت max میگیریم که این باعث می‌شود مشکلات بالا پیش نیاید اما در حالت کلی دو روش اول نتایج بهتری دارند.

۳)

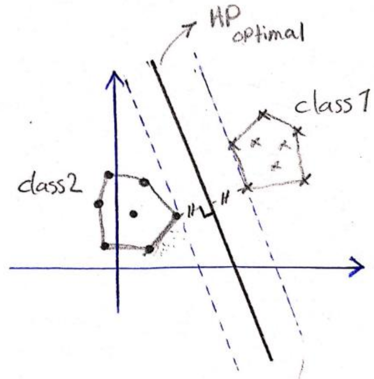
کاهش بُعد یا dimensionality reduction با هدف جلوگیری از نحسی ابعاد(curse of dimensionality) انجام می‌شود. نحسی ابعاد وقتی اتفاق می‌اُفتد که تعداد ویژگی‌ها در مقایسه با تعداد داده ها زیاد باشد که باعث sparse شدن داده ها و در نهایت overfitting و بدتر شدن عملکرد طبقه بند می‌شود. افزایش بُعد با این هدف انجام می‌شود تا بتوان داده‌هایی که در بعد پایین به صورت خطی جداپذیر نیستند به ابعاد بالاتر برد تا به صورت خطی از هم جداپذیر باشند.



(از اسلایدها)

(۴

در مساله طبقه‌بندی دو کلاسه، ابر صفحه optimal(بیشترین حاشیه) بر کوتاه ترین خطی که convex hull مجموعه داده‌های دو کلاس را به هم وصل میکند عمود است و آن را از وسط به دو نصف تقسیم می‌کند.



(۵

اولیه: minimize $\frac{1}{2}\|w\|^2$ subject to $y_i(w^T x_i + b) \geq 1$

maximize → $L(w, b, \alpha) = \frac{1}{2} w^T w + \sum_{i=1}^{n} \alpha_i (1 - y_i (w^T x_i + b))$

$$\begin{cases} \nabla_w L = 0 \rightarrow w + \sum_{i=1}^{n} \alpha_i (-y_i) x_i = 0 \rightarrow \underbrace{w = \sum_{i=1}^{n} \alpha_i y_i x_i}_{\text{جاگذاری در } L} \\ \nabla_b L = 0 \rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

$$\rightarrow L = \frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i x_i^T \sum_{j=1}^{n} \alpha_j y_j x_j + \sum_{i=1}^{n} \alpha_i (1 - y_i (\sum_{j=1}^{n} \alpha_j y_j x_j^T x_i + b))$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \sum_{j=1}^{n} \alpha_j y_j x_j^T x_i$$

$$- b \sum_{i=1}^{n} \alpha_i y_i = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\xrightarrow{\text{مسأله ثانویه}} \begin{cases} \text{maximize } W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{subject to: } \alpha_i \geq 0 \text{ , } \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

۶)

بله- در hard margin بردارهای پشتیبان نقاطی هستند که که نامساوی به حالت تساوی برقرار خواهد بود و در واقع ضریب آلفا متناظر با آن‌ها غیر صفر است. این نقاط در واقع نقاطی هستند که روی خط های حاشیه(margin lines) قرار می‌گیرند. اما در حالت soft-margin علاوه بر این نقاط، نقاطی که در داخل ناحیهmargin قرار می‌گیرند و نیز نقاطی که mis-classify می‌شوند نیز بردار پشتیبان هستند.

پایان