تمرین 4 یادگیری ماشین 99109393 علیرضا حبیب زاده

1.

۱. $h(x,y) = \frac{1}{4}\left(g(x+y,\, x+y) - g(x-y,\, x-y) \right)$

با استفاده از:

ب) $= \frac{1}{4}\left(g(x,\, x+y) + g(y,\, x+y) - g(x,\, x-y) - g(-y,\, x-y) \right)$

$$f(x, y+z) = f(y+z, x) = f(y,x) + f(z,x) = f(x,y) + f(x,z)$$

$$f(x, \alpha y) = f(\alpha y, x) = \alpha f(y,x) = \alpha f(x,y)$$

ازآنجایی که تقارن دائم خطی ب و ج برای هر دو آرگومان برقرار است.

$$h(x,y) = \frac{1}{4}\Big(\cancel{g(x,x)} + g(x,y) + g(y,x) + \cancel{g(y,y)} - \cancel{g(x,x)} + g(x,y) + g(y,x) - \cancel{g(y,y)} \Big)$$

$$= \frac{1}{4}\Big(2g(x,y) + 2\underbrace{g(y,x)}_{=g(x,y)} \Big) = g(x,y)$$

پس $h(x,y) = g(x,y)$ و اگر g هسته‌ی معتبر است h هم معتبر است.

2. در تمامی قسمت ها از آنجایی که k_1 و k_2 هسته های معتبر هستند پس تابع $\phi_1(x)$ و $\phi_2(x)$ وجود دارد به طوری که:

$$K_1(x,y) = \phi_1^T(x)\,\phi_1(y)$$

$$, \quad K_2(x,y) = \phi_2^T(x)\,\phi_2(y)$$

$$K_3(x,y) = k_1(x,y) + k_2(x,y) = \phi_1^T(x)\,\phi_1(y) + \phi_2^T(x)\,\phi_2(y)$$ آ)

ϕ_3 را چنین تعریف میکنیم:

$$\phi_3^T(x) = [\,\phi_1(x)_1 \,,\, \phi_2(x)_1 \,,\, \phi_1(x)_2 \,,\, \phi_2(x)_2 \,,\, \cdots\,]$$

(ابتدا ϕ_1 سپس ϕ_2 قرار دادم چون از متناهی بودن بعد ϕ_1 و ϕ_2 مطمئن نیستیم)

حال به سادگی میتوان دید:

$$\phi_3^T(x)\,\phi_3(y) = \sum \phi_1(x)_i\,\phi_1(y)_i + \sum \phi_2(x)_i\,\phi_2(y)_i$$

$$= \phi_1^T(x)\,\phi_1(y) + \phi_2^T(x)\,\phi_2(y) = K_3(x,y)$$

پس k_3 را به فرم $\phi_3^T(x)\phi(y)$ نوشتیم که در کنار تقارن لازم و کافی هستند.

$$K_3(y,x) = k_1(y,x) + k_2(y,x) = k_1(x,y) + k_2(x,y) = k_3(x,y)$$

پس (آ) یک هسته ی معتبر است.

ب)

$$K_4(x,y) = k_1(x,y)\, k_2(x,y)$$

تقارن : $K_4(y,x) = k_1(y,x)\, k_2(y,x) = k_1(x,y)\, k_2(x,y) = k_4(x,y)$

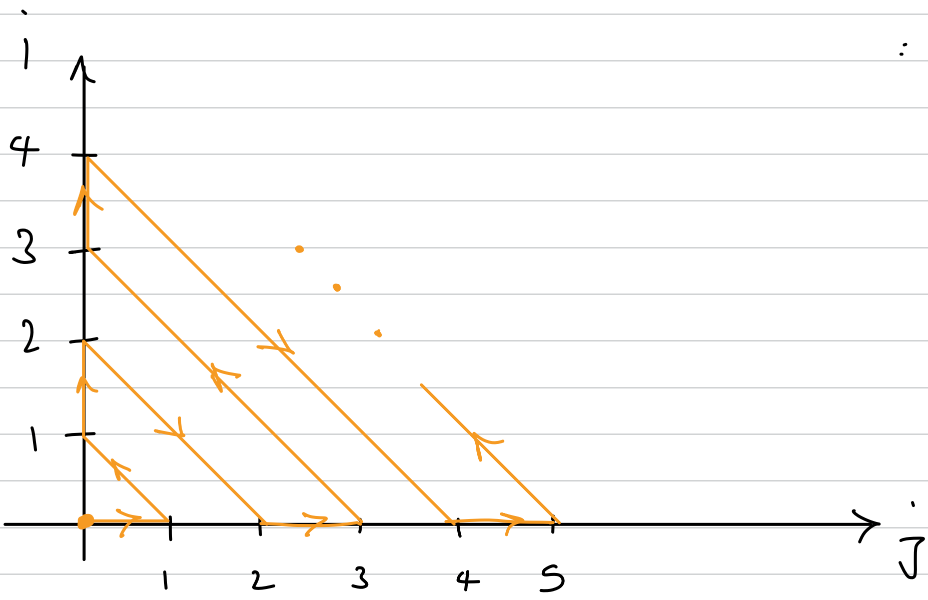
پیدا کردن نگاشت :

$$K_4(x,y) = \phi_1^T(x)\, \phi_1(y) \cdot \phi_2^T(x)\, \phi_2(y)$$

$$= \sum_i \phi_1(x)_i\, \phi_1(y)_i \cdot \sum_j \phi_2(x)_j\, \phi_2(y)_j$$

$$= \sum_{i,j} \phi_1(x)_i\, \phi_2(x)_j \cdot \phi_1(y)_i\, \phi_2(y)_j$$

حال کافی است یک ترتیب روی دوتایی های (i,j) پیدا کنیم که سپس جمع را روی یک متغیر بزنیم. درصورتی که ϕ_1 و ϕ_2 ابعاد محدود داشته باشند که کار راحت است. ابتدا روی یک ϕ میرویم و سپس روی دیگری. اما درهرحال یک مسیر وجود دارد که در ابعاد بی نهایت هم کار کند. n را ترتیب رسیدن به دوتایی ها در این مسیر تعریف میکنیم :



مثال

$$
\begin{array}{ccc}
(i,j) & \longrightarrow & n \\
(0,0) & \longrightarrow & 0 \\
(0,1) & \longrightarrow & 1 \\
(1,0) & \longrightarrow & 2 \\
(2,0) & \longrightarrow & 3 \\
 & \vdots &
\end{array}
$$

حال تعریف میکنیم :

$$\phi_4(x)_n \equiv \phi_1(x)_{i(n)}\, \phi_2(x)_{j(n)}$$

و جمع میشود :

$$K_4(x,y) = \sum_n \phi_4(x)_n\, \phi_4(y)_n$$

$$= \phi_4^T(x)\, \phi_4(y)$$

و هسته معتبر است .

نکته : اگر ابعاد یکی از ϕ ها (بدون کاست از کلیت ϕ_2) محدود است میتوانیم از ضرب کرونکر استفاده کنیم (ترتیب n طوری است که ابتدا اولین مولفه‌ی ϕ_1 در همه‌ی ϕ_2 ضرب میشود و سپس دومین ϕ_1 در همه‌ی ϕ_2 و ... . و با اتحادهای معروف کرونکر حاصل بدیهی است :

$$\phi_4(x) \equiv \phi_1(x) \otimes \phi_2(x)$$

$$\phi_4^T(x)\, \phi_4(y) = [\phi_1(x) \otimes \phi_2(x)]^T [\phi_1(y) \otimes \phi_2(y)]$$

$$= [\phi_1^T(x) \otimes \phi_2^T(x)][\phi_1(y) \otimes \phi_2(y)] \qquad = K_4(x,y)$$

$$= \underbrace{[\phi_1^T(x)\, \phi_1(y)]}_{\text{عدد } 1\times 1} \otimes \underbrace{[\phi_2^T(x)\, \phi_2(y)]}_{\text{عدد } 1\times 1} = \phi_1^T(x)\, \phi_1(y)\, \phi_2^T(x)\, \phi_2(y)$$

ج)

$$k_5(x,y) = e^{k_1(x,y)}$$

تقارن برقرار است: $K_5(y,x) = e^{k_1(y,x)} = e^{k_1(x,y)} = k_5(x,y)$

قضیه: اگر $\alpha > 0$ باشد و K هسته معتبر باشد αK هم هسته معتبر است.

اثبات: $\alpha k(x,y) = \alpha \phi^T(x)\, \phi(y) = (\sqrt{\alpha}\,\phi(x))^T (\sqrt{\alpha}\,\phi(y))$

تقارن هم برقرار است: $\alpha k(y,x) = \alpha k(x,y)$

حال:

$$k_5(x,y) = e^{k_1(x,y)} = 1 + k_1(x,y) + \frac{k_1(x,y)^2}{2!} + \frac{k_1(x,y)^3}{3!} + \cdots$$

$$= \sum_{n=0}^{\infty} \frac{k_1(x,y)^n}{n!}$$

ادعا میکنم هر جمله‌ی مجموع یک کرنل معتبر است:

$$\frac{k_1(x,y)^n}{n!} = \underbrace{\underbrace{\Big[\underbrace{\underbrace{k_1(x,y)}_{\text{معتبر}} \cdot k_1(x,y)}_{\text{معتبر طبق ب}} \cdot \;\cdots\; \cdot k_1(x,y)\Big]}_{\text{معتبر طبق ب}} \times \frac{1}{n!}}_{\text{معتبر طبق قضیه ابتدای صفحه } \left(\frac{1}{n!} > 0\right)}$$

ا) هم هسته‌ی معتبر است چرا که هم متقارن است هم مثبت معین.

حال که هرکدام از جملات جمع هسته‌ی معتبر هستند کل جمع هم هسته‌ی معتبر است

چون طبق (آ) جمع دو هسته‌ی معتبر یک هسته‌ی معتبر است.

$$K_6(x,y) = (1-x^Ty)^{-1}$$ د)

تقارن برقرار است:

$$K_6(y,x) = (1-y^Tx)^{-1} = (1-\underbrace{(x^Ty)^T}_{\substack{\text{عدد} \\ = x^Ty}})^{-1} = K_6(x,y)$$

هسته معتبر نیست. مثال نقض می‌آوریم. فرض کنید بعد فضا ۱ است و

$$K_6(x,y) = \frac{1}{1-x^Ty} = \frac{1}{1-x_0y_0}$$

ماتریس گرام هسته روی تک داده تک بعدی x_0 میشود:

$$A_{1\times1} = \left[\frac{1}{1-x_0^2}\right]$$

حالا قرار می‌دهیم $x_0 = 2$

$$A = \left[\frac{-1}{3}\right]$$

که برای $z_0 = 1$

$$z^TAz = [z_0]\left[\frac{-1}{3}\right][z_0] = \frac{-1}{3}z_0^2$$

کوچکتر از صفر است.

3. ورودی $\hat{k}$ دو مجموعه است. که در فضای مورد علاقه ی ما این دو زیر مجموعه ی دلخواه از X هستند.

تقارن:

$$\hat{k}(B,A) = \sum_{\substack{x \in B \\ x' \in A}} k(x,x')$$

با توجه به متقارن بودن k و dummy بودن x و x' اسم آن ها را عوض می کنیم. (در در k هم می‌چرخانیم)

$$\hat{k}(B,A) = \sum_{\substack{x' \in B \\ x \in A}} k(x,x') = \hat{k}(A,B)$$

k معتبر است پس ϕ وجود دارد به طوری که:

$$\hat{k}(A,B) = \sum_{\substack{x \in A \\ x' \in B}} \phi^T(x)\,\phi(x') = \sum_{x \in A} \sum_{x' \in B} \phi^T(x)\,\phi(x')$$

$$= \left[\sum_{x \in A} \phi^T(x)\right]\left[\sum_{x' \in B} \phi(x')\right] = \left[\sum_{x \in A} \phi(x)\right]^T \left[\sum_{x' \in B} \phi(x')\right]$$

پس تبدیل مناسب در فضای $\hat{X}$ را یافتیم:

$$\hat{\phi}(A) = \sum_{x \in A} \phi(x)$$

حال:

$$\hat{k}(A,B) = \hat{\phi}^T(A)\,\hat{\phi}(B)$$

این در کنار تقارن معتبر بودن هسته را می دهد.

2. ۱. در درس دیدیم با نرمالایز کردن میزان حاشیه برابر میشود با :

$$\rho = \frac{1}{\|w\|} \implies \frac{1}{\rho^2} = \|w\|^2$$

$$\frac{1}{\rho^2} = \Big\|\sum_i \alpha_i y_i x_i\Big\|^2 = \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

جمع تنها برای جملاتی که α ناصفر دارند سالم است که SV ها هستند. میدانیم

برای SV ها این شرط برقرار است.

$$y_i(w^T x_i + b) = 1 \qquad \text{KKT} \ \text{شرط}$$

$$w^T x_i + b = \frac{1}{y_i} = \sum_j \alpha_j y_j K(x_i, x_j) + b \quad *$$

حال برگردیم به $\frac{1}{\rho^2}$:

$$\frac{1}{\rho^2} = \sum_i \alpha_i y_i \underbrace{\sum_j \alpha_j y_j K(x_i, x_j)}_{*}$$

$$= \frac{1}{y_i} - b$$

$$= \sum_i \alpha_i y_i \left(\frac{1}{y_i} - b\right) = \sum_i \alpha_i - \cancelto{0}{b \sum_i \alpha_i y_i}$$

$$\implies \rho = \frac{1}{\sqrt{\sum_i \alpha_i}} \qquad \square$$

2. به نظر من robust است چرا که فلسفه‌ی SVM انتخاب جداکننده‌ای از بین جداکننده های ممکن است که margin کمینه در آن بیشینه میشود.

حال فرض کنید به داده هایی که داریم یک ویژگی جدید اضافه کنیم که تاثیری در فاصله‌ی آن‌ها از ابرصفحه (margin) ندارد. پس support vector ها همان قبلی ها میشوند و در نتیجه همان α های قبلی پاسخ مسئله هستند.

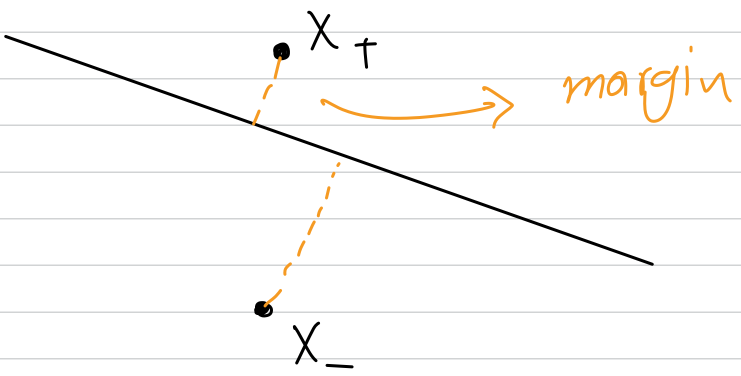
با توجه به بخش قبل میتوان دید که همان α های قبلی مطابق انتظار ρ را تغییر نمی دهند (ρ تنها تابع α هاست).

البته در w ها یک بعد جدید اضافه میشود که انتظار داریم نزدیک صفر باشد چون هر حال در classification تاثیری نگذارد زیرا داریم margin داده‌ها عوض نشدند.

(مگر ویژگی جدید داده‌ها را نسبت به جداکننده طوری آینه ای کند که نتیجه عوض شود و به کلی دگرگون شود ولی margin با توجه به آینه ای شدن عوض نشود که با توجه به اینکه حجم جدید بی ربط است بعید میدانم منظور این باشد).

3. فرض کنید (دقیقا) دو داده‌ی آموزش داریم که لیبلها + و − دارند (متفاوت)

P_+ ، P_- و میانه خط مستقیم سه نقطه در فضای d بعدی هستند که یک ابرصفحه (؟) تشکیل می دهند. صفحه‌ی این سه نقطه را بررسی میکنیم. (آیا اگر هرکدام از P_+ و P_- روی میانه باشند این دو و یک نقطه‌ی دلخواه دیگر)



به طور شهودی واضح است margin بیشینه اگر صفحه بخواهد بین دو نقطه باشد

زمانی است که صفحه برابر با صفحه‌ی عمود منصف پاره خط واصل P_+

و P_- باشد و مقدار margin برابر با نصف فاصله‌ی بین دو نقطه است.

و ضمناً حتماً x_+ و x_- support vector هستند.

حال که صفحه را داریم همه چیز آن معلوم است.

اما می‌خواهیم نشان بدهیم:

$$\sum \alpha_i y_i = \alpha_+ - \alpha_- = 0$$

$$\Rightarrow \alpha_+ = \alpha_- \equiv \alpha$$

$$\Rightarrow W = \sum \alpha_i y_i x_i = \alpha (x_+ - x_-)$$

بردار عمود صفحه در جهت خط واصل دو نقطه است. همان چیزی که حدس می‌زدیم!

$$W^T x_+ + b = 1 = \alpha x_+^2 - \alpha x_+^T x_- + b$$

$$W^T x_- + b = -1 = \alpha x_-^2 - \alpha x_+^T x_- + b$$

$$\mathcal{L} = \alpha_1 + \alpha_2 - \frac{1}{2}\left[\alpha_1^2 X_+^2 + \alpha_2^2 X_-^2 - 2\alpha_1\alpha_2 X_+^T X_-\right]$$

$$= 2\alpha - \frac{1}{2}\alpha^2 \|X_+ - X_-\|^2$$

$$\frac{\partial \mathcal{L}}{\partial \alpha} = 2 - \alpha\|X_+ - X_-\|^2 \Longrightarrow \alpha = \frac{2}{\|X_+ - X_-\|^2} > 0 \quad \checkmark$$

$$\rho = \frac{1}{\sqrt{\Sigma\alpha}} = \sqrt{\frac{\|X_+ - X_-\|^2}{4}} = \left\|\frac{X_+ - X_-}{2}\right\|$$ جالب! از نصف طول بیم

که یعنی صفحه بین دو نقطه است.

$$\Longrightarrow W = 2\,\frac{X_+ - X_-}{\|X_+ - X_-\|^2} \Longrightarrow \|W\| = \frac{2}{\|X_+ - X_-\|}$$

$$X_+^T \frac{2}{\|X_+ - X_-\|^2}(X_+ - X_-) + b = 1$$ [X_+ برای kTT]

$$\Longrightarrow b = \frac{X_+^2 + X_-^2 - \cancel{2X_+^T X_-} - 2X_+^2 + \cancel{2X_+^T X_-}}{\|X_+ - X_-\|^2}$$

$$d = \frac{|W^T X + b|}{\|W\|}$$ فاصله‌ی نقطه‌ی X از صفحه:

$$d(X=0) = \frac{|b|}{\|W\|} = \frac{\|X_+ - X_-\|}{2} \cdot \frac{|X_+^2 - X_-^2|}{\|X_+ - X_-\|^2}$$

$$\boxed{d(x=0) = \left| \left[\frac{x_+ - x_-}{\|x_+ - x_-\|} \right]^T \frac{x_+ + x_-}{2} \right|}$$

نقطه‌ی وسط بین دو داده را

(که میدانستیم صفحه از آن عبور میکرد)

پروجکت کن در جهت $\hat{w}$

(که میدانستیم همان جهت خط واصل دو نقطه است.)

3. ۱.

$$L_\epsilon(x^{(i)}, y^{(i)}, f) = \max(0, |y^{(i)} - f(x^{(i)})| - \epsilon)$$

$$y^{(i)} - f(x^{(i)}) - \epsilon \geqslant \xi_n^* \qquad (y^{(i)} \geqslant f(x^{(i)}))$$

$$f(x^{(i)}) - y^{(i)} - \epsilon \geqslant \xi_n \qquad (y^{(i)} < f(x^{(i)}))$$

تعاریف ξ

با شرط: $\xi_n^*, \xi_n \geqslant 0$

در هر حالت یکی از ξ_n یا ξ_n^* صفر است چون با اندازه یا بیشتر است یا کمتر و ممکن است هر دو صفر شوند (آن‌هایی که داخل بازه‌ی ϵ هستند)

پس: $\max(0, |y^{(i)} - f(x^{(i)})| - \epsilon) = \xi_i^* + \xi_i$

نهایتا:

$$\min \; \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{N}(\xi_i^* + \xi_i)$$

$$\text{subject to} \quad \xi_i, \xi_i^* \geqslant 0$$

2.

$$\mathcal{L} = \frac{1}{2} W^T W + C\sum_i (\xi_i^* + \xi_i)$$

$$- \sum \alpha_i^* (y_i - W^T x_i + \epsilon + \xi_i^*)$$

$$- \sum \alpha_i (W^T x_i - y_i + \epsilon + \xi_i)$$

$$- \sum \beta_i \xi_i - \sum \beta_i^* \xi_i^*$$

شرط ها به فرم $\geqslant 0$

$$\nabla_w L = w - \sum \alpha_i x_i + \sum \alpha_i^* x_i = 0$$

$$w = \sum (\alpha_i - \alpha_i^*) x_i$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \xrightarrow{\beta_i \geq 0} \alpha_i \leq C$$

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

با ضرب کردن یک علامت منفی مسئله به بیشینه تبدیل بشود. (صرفا جاگذاری مشتق بر حسب پله ها است) جاگذاری در لاگرانژین منفی شده:

$$-\frac{1}{2} w^T w = -\frac{1}{2} \left[\sum (\alpha_i - \alpha_i^*) x_i\right]^T \left[\sum (\alpha_i - \alpha_i^*) x_i\right]$$

$$= -1/2 \sum (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle$$

لاگرانژین بر حسب ξ_i و ξ_i^* ها خطی است پس ضرایب آن ها همان مشتق شان است و صفر می‌شود. ضریب w^T هم همین طور.

$$\max \; -\frac{1}{2} \sum (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle$$

$$- \epsilon \sum (\alpha_i + \alpha_i^*) + \sum y_i (\alpha_i - \alpha_i^*)$$

که شرط $\alpha_i, \alpha_i^* \leq C$ را داریم و شرط $\alpha_i, \alpha_i^* \geq 0$ هم از قبل داشتیم:

$$\alpha_i, \alpha_i^* \in [0, C]$$

3. حالا درصورتی که α_i ها و α_i^* ها را در یک دسته قرار بدهیم مثلا λ (که $2N$ تا هستند)
مسئله برحسب λ ها quadratic است چون همه جملات به شکل $\lambda_i \lambda_j \sim$
یا $\lambda_i \sim$ هستند.

4. شروط KKT که برای این لاگرانژین داریم چنین هستند:

$$\alpha_i \left(\xi_i + \epsilon + y^{(i)} - f(x^{(i)}) \right) = 0$$

$$\alpha_i^* \left(\xi_i^* + \epsilon + f(x^{(i)}) - y^{(i)} \right) = 0$$

support vector ها آن هایی هستند که یکی از α_i یا α_i^* آن ها ناصفر باشد یعنی
عبارت سمت راست α_i یا α_i^* برای آن ها باید صفر شود:

$$\xi_i + \epsilon + y^{(i)} - f(x^{(i)}) = 0$$

یا

$$\xi_i^* + \epsilon + f(x^{(i)}) - y^{(i)} = 0$$

با اعمال $\xi_i , \xi_i^* \geq 0$

$$\xi_i = f(x^{(i)}) - y^{(i)} - \epsilon \geq 0$$

یا

$$\xi_i^* = y^{(i)} - f(x^{(i)}) - \epsilon \geq 0$$

$$\Longrightarrow \left| y^{(i)} - f(x^{(i)}) \right| \geq \epsilon$$

support vector ها آن هایی هستند که رو یا خارج سازی خطای ϵ هستند.
(مقدار پیش بینی شده برای آن ها ϵ یا بیشتر بیشتر یا ϵ - یا کمتر کمتر است.

5. $w = \sum_i (\alpha_i - \alpha_i^*) x_i \Longrightarrow w^T x' = \sum_i (\alpha_i - \alpha_i^*) x_i^T x'$

به تنها ضرب داخلی x_i های یادگیری ظاهر میشوند با تبدیل $x \to \phi(x)$ این

حالا به تبدیل میشوند $K(x_i, x')$ که یعنی برای نگاشت تنها دانستن کرنل فضای

متصور کافی است. (در تکنیک کرنل کار میکند)

6. ϵ بازه‌ای است که در داخل آن داده‌ها دیگر penalize نمی‌شوند. در حد $\epsilon \to 0$

همه داده‌ها دخیل میشوند و در واقع جمع فواصل از بیشینه را کمینه میکنیم (شبیه رگرسیون

عادی ولی خود فواصل نه توان دو شان)

C در واقع معادل همان $1/\lambda$ بهینه‌سازی با مولفه‌ی $\lambda \|W\|^2$ است.

در حد $C \to \infty$ تنها حواشی داده‌ها مهم میشود و اوردی که مطرح کردیم و

در C های کوچک $\|W\|^2$ مهم میشود که با کمینه کردن $\|W\|^2$، $\frac{1}{\|W\|}$ را

بیشینه میکنیم که معادل margin است. (البته اینجا اگر بتوان به margin

به همان شکل قبلی در حوالی داده تعریف کرد C کوچک همان $\|W\|^2$ اندازه‌ی

بردار وزن‌ها را کم میکند که از overfit جلوگیری میکند.)

4. ۱. اگر support vector نباشد هیچ تغییری رخ نمی‌دهد. اما اگر باشد مرز به سمت آن حرکت می‌کند چرا که آن نقطه مرزی بوده و الان مرز می‌تواند بزرگتر شود.

البته حالت هایی هم هست که در آن حذف حتی نقاط support vector (یکی) باعث تغییر جواب گسسته SVM و margin نمی‌شود. (مثلا در مجموعه سه نقطه از هر طرف روی margin داریم.

به نظرم در logistic regression همواره مرز تصمیم با حذف شدن یکی از داده‌ها جابجا می‌شود. چرا که در این رگرسیون همه‌ی داده‌ها همواره دخیل هستند و چیزی به عنوان support vector نداریم.

۲. هر کسی که ξ آن نامنفی یعنی margin را دارد کرده. این لزوما به معنی اشتباه classify شدن نیست. ممکن است بین دسته بند تا مرز margin باشد

اما حد بالای خوبی است:

$$N_{wrong} \leqslant \sum_i \theta(\xi_i) \qquad \theta(x) = \begin{cases} 0 & x \leqslant 0 \\ 1 & x > 0 \end{cases}$$

حد بالای بهتری وجود دارد.

$$y_i(w^T x_i) < 0 \Rightarrow 1 - \xi_i < 0$$

$$\Rightarrow \xi_i > 1 \Longleftrightarrow \text{دسته بندی اشتباه}$$

$$N_{wrong} = \sum_i \Theta(\xi_i - 1) \leqslant \boxed{\sum \xi_i}$$

$$\xi_i \geqslant 0$$

3. اینجا دو عامل داریم که با هم مبارزه می‌کنند. یکی زیاد کردن margin تا جای ممکن و دیگری درست classify کردن داده‌ها. می‌توان margin را خیلی زیاد کرد به بهای این که بسیاری از داده‌ها درست classify نشوند یا margin را خیلی کم کرد و تعداد بیشتری از داده‌ها (hopefully همه‌شان) درست classify شوند.

هر کدام از این دو عامل هم تاثیر منفی روی E_{out} دارند. (امید داریم با وجود margin بزرگ E_{out} کم شود و با کم شدن E_{in}، E_{out} هم کم شود) پس جای بهینه‌ای برای فرامقدار C وجود دارد.

$C \to 0$: خطا اهمیت ندارد تنها زیاد شدن margin مهم است.

$C \to \infty$: $\underbrace{\text{margin}}_{\|w\|^2}$ مهم نیست تا جای ممکن درست طبقه‌بندی کن.

4. اگر جواب‌ها بهینه باشند SVM بین جواب‌های ممکن جوابی را می‌دهد که بیشترین margin را دارد اما LR لزوما این کار را نمی‌کند بلکه ممکن است در فضای جواب‌ها هر جایی را بگیرد (به نظرم در عمده موارد) در این حالت SVM بهتر است هرچند هزینه بیشتری دارد.

5. Soft SVM نسبت به این که چه C برایش انتخاب کنیم بین دو فاکتوری که بیشتر گفتم (دسته‌بندی درست و margin زیاد) یک جواب می‌دهد ولی LR همچنان همان خطای خودش را کمینه می‌کند.

LR احتمال (likelihood) وقوع این دیتاست را بیشینه می‌کند ولی soft SVM کار دیگری می‌کند. soft SVM در صورتی که بدون نگاشت استفاده شود یک جداکننده‌ی خطی به ما می‌دهد که نتیجه قابل فهم‌تر و ساده‌تری از LR است.