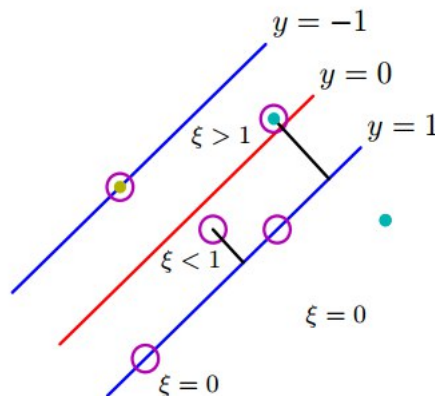




## سؤال ۱. ماشین بردار پشتیبان

• الف)

- یک) خیر. اگر  $C \rightarrow \infty$  آن‌گاه برای مجموعه داده‌هایی که به صورت خطی جداپذیر باشند، این اتفاق رخ خواهد داد.<sup>۱</sup>
- دو) همان‌طور که در شکل زیر مشخص است، حالات مختلف ممکن برای  $\epsilon$  عبارتند از:
  - \*  $\epsilon > 1$ : داده به‌طور اشتباه دسته‌بندی شده است.
  - \*  $\epsilon = 1$ : یعنی داده روی مرز تصمیم<sup>۲</sup> قرار دارد.
  - \*  $0 < \epsilon < 1$ : داده درست دسته‌بندی شده است اما در margin قرار دارد.
  - \*  $\epsilon = 0$ : داده درست دسته‌بندی شده است اما ممکن است روی margin یا بیرون آن باشد.



شکل ۱: توصیف معنایی حالات مختلف  $\epsilon$

- سه) حل مسئله با  $M$  متغیر به طور کلی پیچیدگی از  $O(M^3)$  دارد. در دوگان مسئله‌ی بهینه‌سازی را می‌خواهیم حل کنیم که به جای  $M$  متغیر،  $N$  متغیر دارد. برای تعداد ثابت توابع basis چون که  $M < N$  است، استفاده از روش دوگان به دلیل سرعت کمتر خوب نیست.
- استفاده روش دوگان به ما امکان استفاده از هسته‌ها را می‌دهد و می‌توانیم دسته‌بند حاشیه بیشینه<sup>۳</sup> را که ابعاد ویژگی‌های آن بیش‌تر از تعداد نقاط است پیدا کنیم و بسیار کارا است و حتی جواب‌گوی مواقعی که تعداد ویژگی‌ها به بی‌نهایت میل می‌کند می‌باشد.<sup>۴</sup>

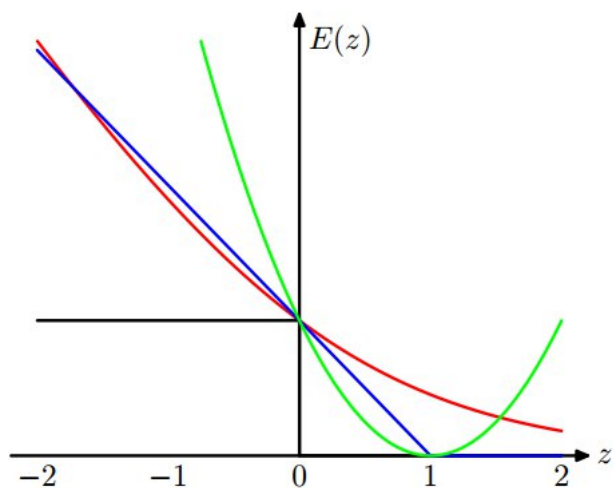
<sup>۱</sup> ر.ک. صفحه ۳۳۲ کتاب

<sup>۲</sup> decision boundary

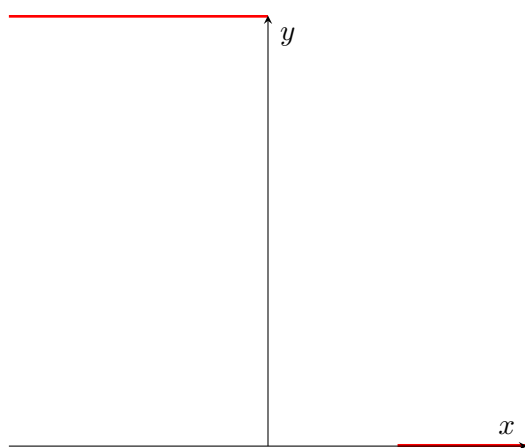
<sup>۳</sup> maximum margin classifier

<sup>۴</sup> ر.ک صفحه ۳۲۹ کتاب

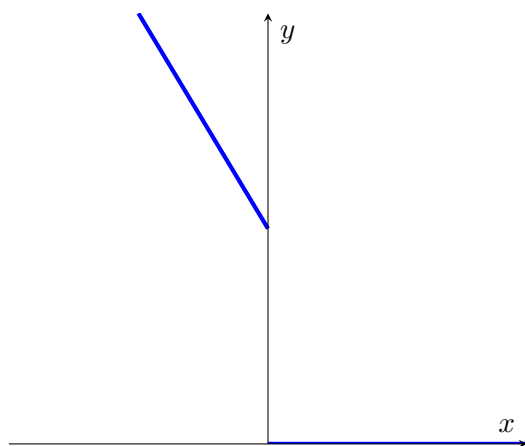
- چهار)



شکل ۲: نمودار آبی: ماشین بردار پشتیبان نرم، نمودار قرمز: logistic regression، نمودار مشکی: خطای دسته‌بندی و نمودار سبز: مربع خطاها



شکل ۳: ماشین بردار پشتیبان سخت



شکل ۴: perceptron

- (ب) در این رابطه حساسیت به داده‌های نویز بیشتر می‌شود؛ زیرا در مدل هدف این است که مجموع مجذورات را کمینه کنیم، بنابراین داده‌های کم‌تری باید به‌طور اشتباه دسته‌بندی شوند و داده‌های نویز اثر و وزن زیادی نسبت به حالت عادی تابع هزینه دارند.
- (پ)

– یک)

$$\tilde{L}(a) = \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(x_n m x_m)$$

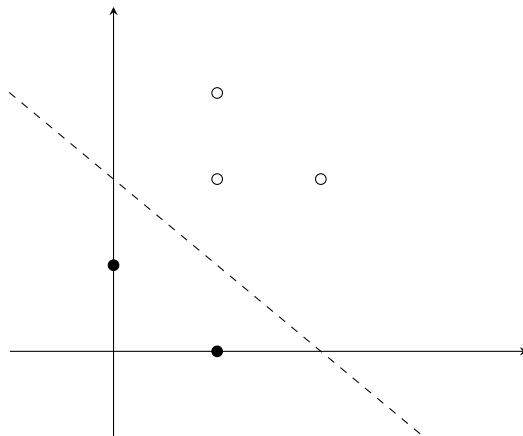
از آن جایی که تعداد داده‌ها ۵ است، بنابراین: <sup>۵</sup>

$$\tilde{L}(a) = \sum_{n=1}^5 a_n - \frac{1}{2} \sum_{n=1}^5 \sum_{m=1}^5 a_n a_m t_n t_m k(x_n m x_m)$$

$$\tilde{L}(a) = \sum_{n=1}^5 a_n - \frac{1}{2} (a_1^2 - 4a_1 a_4 - 4a_1 a_3 - 6a_1 a_5 + a_2^2 - 2a_2 a_3 - 4a_2 a_4 + 5a_3^2 - 2a_2 a_5 + 12a_3 a_4 + 14a_3 a_5 + 8a_4^2 + 16a_4 a_5 + 10a_5^2)$$

$$a_n \geq 0 \implies \sum_{n=1}^N a_n t_n = 0 \implies a_1 + a_2 - a_3 - a_4 - a_5 = 0$$

– (دو) چون نقطه‌های ۴ و ۵ نمی‌توانند بردار پشتیبان باشند، داریم:



$$a_4 = a_5 = 0 \implies a_1 + a_2 - a_3 = 0 \implies a_1 + a_2 = a_3$$

– (سه) با جای‌گذاری نتیجه‌ی بالا در معادله‌ی قسمت «۱» داریم:

$$\tilde{L}(a) = a_1 + a_2 + a_3 - \frac{1}{2} (a_1^2 + a_2^2 + 5a_3^2 - 4a_1 a_3 - 2a_2 a_3)$$

– (چهار) برای حل مسئله‌ی بهینه‌سازی باید لاگرانژ تابع را برابر با صفر قرار دهیم.

$$\nabla \tilde{L}(a) = 0 \implies \begin{cases} 1 - \frac{1}{2}(2a_1 - 4a_3) = 0 & \text{with respect to } a_1 \\ 1 - \frac{1}{2}(2a_2 - 2a_3) = 0 & \text{with respect to } a_2 \\ 1 - \frac{1}{2}(-4a_1 - 2a_2 + 10a_3) = 0 & \text{with respect to } a_3 \end{cases}$$

از قسمت قبلی دریافتیم که  $a_1 + a_2 = a_3$ . حال با حذف  $a_3$  از دو معادله داریم:

$$a_1 - a_2 = a_3$$

با جای‌گذاری این نتایج در معادلات بالا، به دست می‌آوریم:

$$a_1 = a_3 = \frac{1}{3}, a_2 = 0$$

– پنج)

$$y(x) = wx + b \implies w = \frac{1}{3}(x_1 - x_3) = \left(-\frac{1}{3}, -\frac{1}{3}\right), b = \frac{2}{3} \implies y(x) = \left(-\frac{1}{3}, -\frac{1}{3}\right)x + \frac{2}{3}$$

## سؤال ۲. هسته

• الف)

– یک) از آنجایی که می‌دانیم  $k_1(x, x')$  و  $k_2(x, x')$  دو هسته‌ی معتبر هستند؛ بنابراین ماتریس هسته  $K_1$  و  $K_2$  هر دو مثبت نیمه‌معین هستند. پس اگر فرض کنیم که ورودی‌های  $x$  و  $x'$  هر دو، دو بعدی باشند، داریم:

$$\begin{aligned} k_3(x, x') &= (x^T x')^2 = (x_1 x'_1 + x_2 x'_2)^2 = x_1^2 x_1'^2 + 2x_1 x'_1 x_2 x'_2 + x_2^2 x_2'^2 \\ \Rightarrow k_3(x, x') &= (x_1^2, \sqrt{2}x_1 x_2, x_2^2)(x_1'^2, \sqrt{2}x'_1 x'_2, x_2'^2) = \phi(x)^T \phi(x') \end{aligned}$$

با توجه به نتیجه‌ی بالا داریم:

$$k_3 = k_1 + k_2$$

پس  $k_3$  نیز مثبت نیمه‌معین بوده و ثابت شد که معتبر نیز هست.

– دو) با فرض این‌که تابع نگاشت هسته‌ی  $k_1$ ،  $\phi^{(1)}(x)$  با ابعاد  $M$  و تابع نگاشت هسته‌ی  $k_2$ ،  $\phi^{(2)}(x)$  با ابعاد  $N$  است، داریم:

$$\begin{aligned} k_4(x, x') &= k_1(x, x')k_2(x, x') = \phi^{(1)}(x)^T \phi^{(1)}(x') \phi^{(2)}(x)^T \phi^{(2)}(x') = \sum_{i=1}^M \phi_i^{(1)}(x) \phi_i^{(1)}(x') \sum_{j=1}^N \phi_j^{(2)}(x) \phi_j^{(2)}(x') \\ \Rightarrow k_4(x, x') &= \sum_{i=1}^M \sum_{j=1}^N [\phi_i^{(1)}(x) \phi_j^{(2)}(x)] [\phi_i^{(1)}(x') \phi_j^{(2)}(x')] = \sum_{k=1}^M N \phi_k(x) \phi_k(x') = \phi(x)^T \phi(x') \end{aligned}$$

که در آن  $\phi_i^{(1)}(x)$ ، عنصر  $i$ ام  $\phi^{(1)}(x)$  و  $\phi_j^{(2)}(x)$  عنصر  $j$ ام  $\phi^{(2)}(x)$  است.

– سه)

از آنجایی که  $k_1$  یک هسته‌ی معتبر است، بنابراین آن را به صورت  $k_1(x, x') = \phi(x)^T \phi(x')$  می‌توان نوشت. پس:

$$k_5(x, x') = a k_1(x, x') = [\sqrt{a} \phi(x)]^T [\sqrt{a} \phi(x')]$$

که با توجه به فرض  $a \geq 0$  ثابت می‌شود.

– چهار) اگر بسط تیلور را بنویسیم داریم:

$$k_6(x, x') = a_n k_1(x, x')^n + a_{n-1} k_1(x, x')^{n-1} + \dots + a_1 k_1(x, x') + a_0$$

حال با استفاده از نتایج سه قسمت قبل و با توجه به این‌که همه‌ی ضرایب بسط تیلور مثبت است، پس این هسته نیز یک هسته‌ی معتبر است.

• ب) با توجه به این‌که تابع هسته را می‌توان آن را به شکل ضرب داخلی در فضای ویژگی نوشت، پس هسته‌ی معتبر است. برای اثبات باید به رابطه‌ی  $k(A, B) = 2^{|A \cap B|} = \phi(A)^T \phi(B)$  برسیم.

اگر

$$\phi_U(X) = \begin{cases} 1 & \text{if } U \subseteq X \\ 0 & \text{otherwise} \end{cases}$$

باشد، داریم:

$$\phi(A)^T \phi(B) = \sum_{U \subseteq A \cap B} \phi_U(A) \phi_U(B)$$

با استفاده از سیگما (جمع کردن) در رابطه‌ی بالا، همه‌ی زیرمجموعه‌های ممکن  $|A \cap B|$  را اگر و تنها اگر هم زیرمجموعه‌ی  $A$  و  $B$  باشد (مقدار برابر با ۱) داریم. با این کار تعداد زیرمجموعه‌های اشتراک  $A$  و  $B$  در فضای  $S$  را محاسبه کرده‌ایم. علاوه‌براین هم  $A$  و هم  $B$  به عنوان زیرمجموعه‌ی فضای  $S$  معرفی شده‌اند، بنابراین:

$$\phi(A)^T \phi(B) = 2^{|A \cap B|}$$

• پ)

یک -)

$$k(x, x') = (x^T \cdot x' + c)^2 = k\left(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}, \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_d \end{pmatrix}\right) = (c + x_1 x'_1 + x_2 x'_2 + \dots + x_d x'_d)^2$$

$$\Rightarrow k(x, x') = c^2 \sum_{i=1}^d x_i^2 x_i'^2 + \sum_{i=1}^d 2c x_i x_i' + \sum_{i=1}^{d-1} \sum_{j=i+1}^d 2x_i x_i' x_j x_j'$$

$$\Rightarrow k(x, x') = \begin{pmatrix} c & x_1^2, \dots, x_d^2, & \dots & \sqrt{2c}x_d \end{pmatrix} \begin{pmatrix} x'_1, \dots, x'_d \\ \vdots \\ \sqrt{2c}x'_d \end{pmatrix}$$

$$\Rightarrow k(x, x') = \phi(x)^T \phi(x')$$

– دو) اگر  $c = 0$  باشد، آنگاه فضای تبدیل  $d + 1$  بعد کاهش می‌یابد؛ زیرا تعداد جملاتی که در آن  $c$  ضرب شده است،  $d + 1$  است که با صفر شدن آن حذف می‌شوند.

– سه)

$$k(x, x') = (x^T \cdot x' + c)^M$$

با توجه به توضیحاتی که در این لینک داده شده است، تعداد جملات برابر با:

$$\# \text{ of expressions} = \binom{M + d + 1 - 1}{d + 1 - 1} = \binom{M + d}{d}$$

### سؤال ۳. درخت تصمیم

• الف)

• ب)

• پ)

– یک)

– دو)

– سه)

## سؤال ۴. یادگیری جمعی

• الف) نامساوی Jensen:

$$f(\sum_{i=1}^M \lambda_i x_i) \leq \sum_{i=1}^M \lambda_i f(x_i)$$

همان‌طور که در صورت سوال گفته شده است برای  $E_{avg}$  داریم:

$$E_{avg} = \frac{1}{M} \sum_{i=1}^M E_x[(h_m(x) - h(x))^2]$$

حال اگر  $\frac{1}{M}$  را به درون سیگما ببریم، داریم:

$$E_{avg} = E_x[\sum_{m=1}^M \frac{1}{M} (h_m(x) - h(x))^2]$$

حال با توجه به نامساوی Jensen و محدب بودن تابع داریم:

$$(\sum_{m=1}^M \frac{1}{M} (h_m(x) - h(x))^2) \leq \sum_{m=1}^M \frac{1}{M} (h_m(x) - h(x))^2$$

$$\implies E_{com} \leq E_{avg}$$

• ب)

$$E_{avg} = \frac{1}{M} \sum_{m=1}^M E_x[(h_m(x) - h(x))^2]$$

$$E_{com} = E_x[(\frac{1}{M} \sum_{m=1}^M h_m(x) - h(x))^2] = E_x[(\frac{1}{M} \sum_{m=1}^M h_m(x) - h(x))(\frac{1}{M} \sum_{l=1}^M h_l(x) - h(x))]$$

حال اگر یکی از عامل‌های  $\frac{1}{M}$  را به دلیل ثابت بودن از داخل امید ریاضی بیرون بیاوریم، به دلیل فرضیات یعنی

$$\forall m \neq l \ E[(h_m(x) - h(x))(h_l(x) - h(x))] = 0$$

داریم:

$$E_{com} = \frac{1}{M} (\frac{1}{M} \sum_{m=1}^M E_x[(h_m(x) - h(x))^2]) = \frac{1}{M} E_{avg}$$



• الف)

• ب) خطای تابع نمایی به صورت  $E = \sum_{n=1}^N e^{-t_n f_m(x_n)}$  است که  $f_m$  به شکل  $f_m(x) = \frac{1}{2} \sum_{l=1}^m \alpha_l y_l(x)$  تعریف می شود و  $t_n \in \{-1, 1\}$  می باشد. به جای کمینه کردن تابع خطای کلی، می توان با توجه به  $\alpha_m$  و  $y_m(x)$  این کار را انجام داد:

$$E = \sum_{n=1}^N e^{-t_n f_{m-1}(x_n) - \frac{1}{2} t_n \alpha_m y_m(x_n)} = \sum_{n=1}^N w_n^{(m)} e^{-\frac{1}{2} t_n \alpha_m y_m(x_n)}$$

که در آن  $w_n^{(m)} = e^{-t_n f_{m-1}(x_n)}$  بوده و آن را می توان ثابت در نظر گرفت زیرا تنها  $\alpha_m$  و  $y_m(x)$  را بهبود می دهیم. اگر  $\tau_m$  را داده هایی که درست دسته بندی شده اند و  $M_m$  را داده هایی بگیریم که غلط دسته بندی شده اند، بگیریم داریم:

$$E = e^{-\frac{\alpha_m}{2} \sum_{n \in \tau_m} w_n^{(m)}} + e^{\frac{\alpha_m}{2} \sum_{n \in M_m} w_n^{(m)}} \\ \Rightarrow E = (e^{\frac{\alpha_m}{2}} - e^{-\frac{\alpha_m}{2}}) \sum_{n=1}^N w_n^{(m)} I(y_m(x_n) \neq t_n) + e^{-\frac{\alpha_m}{2} \sum_{n=1}^N w_n^{(m)}}$$

حال اگر بخواهیم با توجه به  $y_m$  کمینه کنیم، عبارت دوم ثابت خواهد بود بنابراین با عنایت به نتایجی که معادلات بالا به دست آمد داریم:

$$w_n^{(m+1)} = w_n^{(m)} e^{-\frac{1}{2} t_n \alpha_m y_m(x_n)}$$

با توجه به  $(1 - 2I(y_m(x_n) \neq t_n))$  می توان عبارت به دست آوردن وزن جدید را به صورت زیر بازنویسی کرد:

$$w_n^{(m+1)} = w_n^{(m)} e^{-\frac{\alpha_m}{2}} e^{\alpha_m I(y_m(x_n) \neq t_n)}$$

به دلیل استقلال  $e^{-\frac{\alpha_m}{2}}$  از  $n$  چون همه ی عبارات چنین عامل مشترکی دارند می توان آن را در نظر نگرفت.