دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

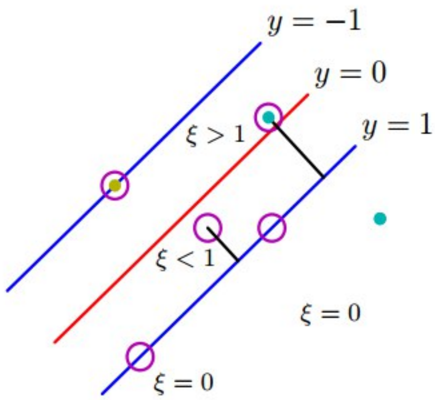
درس یادگیری ماشین

استاد سید عباس حسینی

تمرین ششم | Decision Tree and Ensemble Learning | مصطفی قدیمی

سؤال ۱. ماشین بردار پشتیبان

- الف)
  - یک) خیر. اگر $C \to \infty$ آن‌گاه برای مجموعه داده‌هایی که به صورت خطی جداپذیر باشند، این اتفاق رخ خواهد داد. [۱]
  - دو) همان‌طور که در شکل زیر مشخص است، حالات مختلف ممکن برای ϵ عبارتند از:
    * $\epsilon > 1$: داده به‌طور اشتباه دسته‌بندی شده است.
    * $\epsilon = 1$: یعنی داده روی مرز تصمیم [۲] 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.

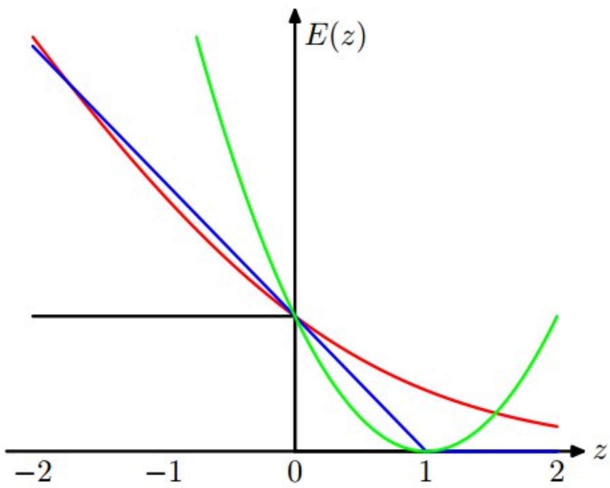


شکل ۱: توصیف معنایی حالات مختلف ϵ

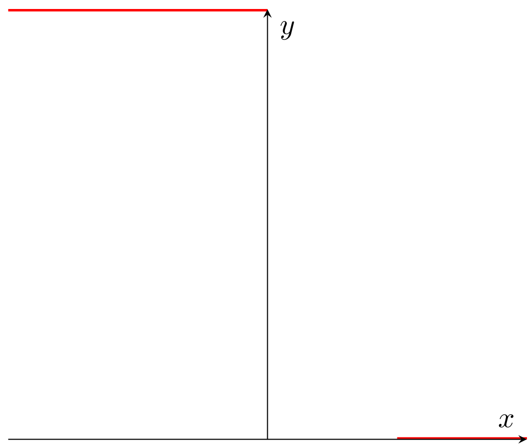
  - سه) حل مسئله با M متغیر به طور کلی پیچیدگی از $O(M^3)$ دارد. در دوگان مسئله‌ی بهینه‌سازی را می‌خواهیم حل کنیم که به‌جای M متغیر، N متغیر دارد. برای تعداد ثابت توابع basis چون که $M < N$ است، استفاده از روش دوگان به دلیل سرعت کم‌تر خوب نیست.
  استفاده روش دوگان به ما امکان استفاده از هسته‌ها را می‌دهد و می‌توانیم دسته‌بند حاشیه بیشینه [۳] 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. [۴]

---
[۱] ر.ک. صفحه ۳۳۲ کتاب
[۲] decision boundry
[۳] maximum margin classifier
[۴] ر.ک صفحه ۳۲۹ کتاب

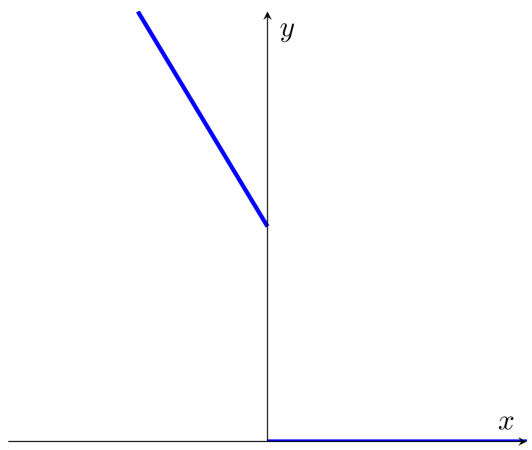
- چهار)



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.

$$L(w, w_0, \Sigma, \alpha, \beta) = \frac{1}{2}||w||^2 + \Sigma_{n=1}^{N}(\alpha_n)(1 - \Sigma_n - y^{(n)}(w^T x^{(n)} + w_0) + c\Sigma_{n=1}^{N}\Sigma_n^2 - \Sigma_{n=1}^{N}\beta_n\Sigma_n$$

- پ)
  - یک)

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

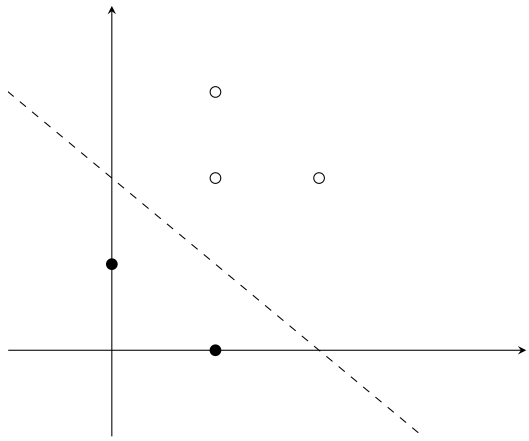
از آن‌جایی که تعداد داده‌ها ۵ است، بنابراین: [۵]

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

$$\tilde{L}(a) = \Sigma_{n=1}^{5}a_n - \frac{1}{2}(a_1^2 - 4a_1a_4 - 4a_1a_3 - 6a_1a_5 + a_2^2 - 2a_2a_3 - 4a_2a_4 + 5a_3^2 - 2a_2a_5 + 12a_3a_4 + 14a_3a_5 + 8a_4^2 + 16a_4a_5 + 10a_5^2)$$

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

  - 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:

$$\tilde{L}(a) = a_1 + a_2 + a_3 - \frac{1}{2}(a_1^2 + a_2^2 + 5a_3^2 - 4a_1a_3 - 2a_2a_3)$$

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

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

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

$$a_1 - a_2 = a_3$$

[۵] ر.ک. صفحه ۳۲۹ کتاب

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

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

- پنج)

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

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:

$$k_3(x,x') = (x^T x')^2 = (x_1x_1' + x_2x_2')^2 = x_1^2x_1'^2 + 2x_1x_1'x_2x_2' + x_2^2x_2'^2$$
$$\implies k_3(x,x') = (x_1^2, \sqrt{2}x_1x_2, x_2^2)(x_1'^2, \sqrt{2}x_1'x_2', x_2'^2) = \phi(x)^T\phi(x')$$

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:

$$k_4(x,x') = k_1(x,x')k_2(x,x') = \phi^{(1)}(x)^T\phi^{(1)}(x')\phi^{(2)}(x)^T\phi^{(2)}(x') = \Sigma_{i=1}^{M}\phi_i^{(1)}(x)\phi_i^{(1)}(x')\Sigma_{j=1}^{N}\phi_j^{(2)}(x)\phi_j^{(2)}(x')$$
$$\implies k_4(x,x') = \Sigma_{i=1}^{M}\Sigma_{j=1}^{N}[\phi_i^{(1)}(x)\phi_j^{(2)}(x)][\phi_i^{(1)}(x')\phi_j^{(2)}(x')] = \Sigma_{k=1}^{M}N\phi_k(x)\phi_k(x') = \phi(x)^T\phi(x')$$

که در آن $\phi_i^{(1)}(x)$، عنصر iام $\phi^{(1)}(x)$ و $\phi_j^{(2)}(x)$ عنصر jام $\phi^{(2)}(x)$ 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:

$$k_5(x,x') = ak_1(x,x') = [\sqrt{a}\phi(x)]^T[\sqrt{a}\phi(x')]$$

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

  - چهار) اگر بسط تیلور را بنویسیم داریم:

$$k_6(x,x') = a_nk_1(x,x')^n + a_{n-1}k_1(x,x')^{n-1} + \cdots + a_1k_1(x,x') + a_0$$

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

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

باشد، داریم:

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

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)

$$k(x,x') = (x^T.x' + c)^2 = k(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}, \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_d \end{pmatrix}) = (c + x_1x'_1 + x_2x'_2 + \cdots + x_dx'_d)^2$$

$$\implies k(x,x') = c^2\Sigma_{i=1}^{d} x_i^2 x_i'^2 + \Sigma_{i=1}^{d} 2cx_ix'_i + \Sigma_{i=1}^{d-1}\Sigma_{j=i+1}^{d} 2x_ix'_ix_jx'_j$$

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

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

– 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)

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

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

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

سؤال ۳. درخت تصمیم

- الف)

$$Gain(y, x_d) = \Sigma_i \Sigma_j p(y = j, x_d = i) log \frac{p(x_d = i)p(y = j)}{p(y = j, x_d = i)}$$

با توجه به استقلال y و x_d داریم:

$$p(y, x_d) = p(y)p(x_d)$$

$$\implies Gain(y, x_d) = \Sigma_i \Sigma_j p(y = j, x_d = i) log \frac{p(x_d = i)p(y = j)}{p(y = j)p(x_d = i)} = \Sigma_i \Sigma_j p(y = j, x_d = i) log(1) = 0$$

- ب) اگر ملاک ساختن درخت تصمیم‌گیری را براساس مقدار information gain قرار دهیم، به دلیل استقلال و یکتا بودن آن (همانند قسمت قبل)، مقدار gain برابر با صفر می‌شود. پس به عنوان ریشه‌ی درخت قرار نمی‌گیرد. حال اگر فرض کنیم این ویژگی به عنوان ریشه انتخاب شده، overfitting به دلیل یکتایی ویژگی‌ها حتمی خواهد بود. برای جلوگیری از overfitting عمدتا دو راه‌کار وجود دارد:

1. زودتر متوقف کردن درخت برای این‌که به نقطه‌ای نرسد که همه‌ی داده‌های آموزش را به‌طور کامل و عالی یاد گرفته باشد.
2. استفاده از تکنیک post-pruning. به روی‌کردی می‌گویند که داده‌ی آموزش و اعتبارسنجی وجود دارد که برای هرس کردن از آن استفاده می‌شود.

- پ)
  - یک)
  - دو)
  - سه)

سؤال ۴. یادگیری جمعی

- الف) نامساوی Jensen:

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

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

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

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

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

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

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

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

- ب)

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

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

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:

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

سؤال ۵. Adaboost

- الف) فرض خلف: $h_t = h_{t+1}$

$$\epsilon_t = \frac{\Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} \neq h_t(x^{(i)})}{\Sigma_{i=1}^{N} w_t^{(i)}} = \frac{\Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} \neq h_t(x^{(i)})}{\Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} \neq h_t(x^{(i)}) + \Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} = h_t(x^{(i)})} < \frac{1}{2}$$

با توجه به نحوه‌ی عمل‌کرد الگوریتم Adaboost و فرض اولیه داریم:

$$\epsilon_{t+1} = \frac{\Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} \neq h_t(x^{(i)})}{\Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} \neq h_t(x^{(i)}) + \Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} = h_t(x^{(i)})} = \frac{\frac{1-\epsilon_t}{\epsilon_t} \Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} \neq h_t(x^{(i)})}{\frac{1-\epsilon_t}{\epsilon_t} \Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} \neq h_t(x^{(i)}) + \Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} = h_t(x^{(i)})}$$

$$\Longrightarrow \epsilon_{t+1} = \frac{\Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} = h_t(x^{(i)})}{2\Sigma_{i=1}^{N} w_t^{(i)} I(y^{(i)} = h_t(x^{(i)})} = \frac{1}{2}$$

که خلاف فرض است و حکم ثابت می‌شود.

- ب) خطای تابع نمایی به صورت $E = \Sigma_{n=1}^{N} e^{-t_n f_m(x_n)}$ است که f_m به شکل $f_m(x) = \frac{1}{2}\Sigma_{l=1}^{m} \alpha_l y_l(x)$ تعریف می‌شود و $t_n \in \{-1, 1\}$ می‌باشد. به‌جای کمینه کردن تابع خطای کلی، می‌توان با توجه به α_m و $y_m(x)$ این کار را انجام داد:

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

که در آن $w_n^{(m)} = e^{-t_n f_{m-1}(x_n)}$ بوده و آن را می‌توان ثابت در نظر گرفت زیرا تنها α_m و $y_m(x)$ را بهبود می‌دهیم. اگر τ_m را داده‌هایی که درست دسته‌بندی شده‌اند و M_m را داده‌هایی بگیریم که غلط دسته‌بندی شده‌اند، بگیریم داریم:

$$E = e^{-\frac{\alpha_m}{2}} \Sigma_{n \in \tau_m} w_n^{(m)} + e^{\frac{\alpha_m}{2}} \Sigma_{n \in M_m} w_n^{(m)}$$

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

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 $t_n y_m(x_n) 1 - 2I(y_m(x_n) \neq t_n)$ می‌توان 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 همه‌ی عبارات چنین عامل مشترکی دارند می‌توان آن را در نظر نگرفت.