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)
  - سه)
  - چهار)
- ب)
- پ)
  - یک)
  - دو)
  - سه)
  - چهار)
  - پنج)

[۱] ر.ک. صفحه ۳۳۲ کتاب

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:

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

  با توجه به نتیجه‌ی بالا داریم:

$$k_3 = k_1 + k_2$$

  پس k_3 نیز مثبت نیمه‌معین بوده و ثایت 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:

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

  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

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

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:

$$\#\, of\ expressions = \begin{pmatrix} M + d + 1 - 1 \\ d + 1 - 1 \end{pmatrix} = \begin{pmatrix} M + d \\ d \end{pmatrix}$$

سؤال ۳. درخت تصمیم

- الف)
- ب)
- پ)
  - یک)
  - دو)
  - سه)

سؤال ۴. یادگیری جمعی

- الف) نامساوی Jensen:

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

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:

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

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

- 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)
- 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:

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

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

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.