دانشگاه صنعتی شریف	**مقدمه ای بر یادگیری ماشین 25737**
دانشکده مهندسی برق	
نیمسال پاییز 1401–1402	**مدرس: سید جمال الدین گلستانی**

بخشی از مسائل تئوریک تکالیف (ویراست چهارم)

مساله T1:

فوق صفحه یا hyperplane (اختصارا HP) مشخص شده با رابطه‌ی $\omega^T x + b = 0\ ,\ \omega, x \epsilon \mathbb{R}^n\ ,\ b \epsilon \mathbb{R}$ را در فضای $X = \mathbb{R}^n$ در نظر بگیرید.

الف - نشان دهید که بردار ω بر این HP عمود است. به عبارت دیگر نشان دهید که به ازای هر دو بردار u و v در این HP، خط واصل بین u و v (یعنی بردار v-u) بر ω عمود است.

ب - نشان دهید که جهت بردار ω به سمت نیم فضای $\omega^T x + b > 0$ است. برای اینکار کافی است نشان دهید که اگر از هر نقطه x بر روی HP در جهت ω حرکت کنیم، یعنی به نقطه $u = x + \alpha\omega\ ,\ \alpha > 0$ برویم، u در نیم فضای مذکور قرار دارد.

ج - ملاحظه کنید که اگر ω را به $\omega' = \alpha\omega$ و b را به $b' = \alpha b$ تغییر دهیم که α یک عدد حقیقی است، HP تغییر نمی‌کند، اما اگر α منفی باشد، جای دو نیم فضا با هم عوض می‌شود.

د - فاصله یک نقطه دلخواه u را از فوق صفحه $\omega^T x + b = 0$ بدست آورید. با توجه به اینکه ω بر فوق صفحه عمود است، فاصله u از فوق صفحه برابر است با مسافتی که باید از نقطه u در جهت $u + \alpha\omega$ حرکت کرد تا به نقطه ای بر روی فوق صفحه رسید (α می‌تواند مثبت یا منفی باشد)

مساله T2:

فرض کنید $X = \mathbb{R}$ و $Y = \mathbb{R}$ باشد و مجموعه داده آموزشی به صورت $S = \{(0,1),(1,0),(2,4)\}$ در اختیار است. می‌خواهیم یک چندجمله‌ای درجه دوم $h(x) = a_0 + a_1 x + a_2 x^2$ بدست آوریم که بر اساس خطای mean square یعنی $l\big(h,(x,y)\big) = \big(h(x) - y\big)^2$ بهترین انطباق را با داده آموزشی S داشته باشد.

الف – تابع ریسک تجربی $L_S(h)$ را برحسب ضرایب a_2,a_1,a_0 بیان کنید.

ب – از این تابع مستقیما نسبت به ضرایب a_2,a_1,a_0 مشتق بگیرید و با صفر نهادن مشتقات و حل دستگاه معادله بدست آمده، ضرایب را بدست آورید.

ج – حال مساله را با استفاده از رابطه ماتریسی بدست آمده در درس حل نمایید و ضرایب بدست آمده را با بند 'ب' مقایسه کنید.

مساله T3:

برای یادگیری یک مساله Binary Classification در فضای $\chi = \mathbb{R}^2$ ، مجموعه داده آموزشی S شامل 6 نقطه به شرح زیر در دست است:

$$S = \{((0,1) - 1),((1,0), -1), ((6,6), -1),((2,1), +1),((1,2), +1),((5,5), +1)\}$$

می‌دانیم که یادگیری این مساله بر اساس یک منحنی درجه 2 در χ به خوبی صورت می‌گیرد و داده آموزشی فوق با منحنی درجه 2 قابل جداسازی است.

الف – با تعریف بردار $\psi(x)$ مناسب، این مساله را به یک مساله طبقه بندی خطی تبدیل نمایید.

ب – همه قیود لازم بر روی بردار w را به نحوی که جداسازی موردنظر در S صورت گیرد، بیان نمایید.

ج – دو گام از الگوریتم Perceptron را به طور دستی اجرا نمایید.

مساله T4:

در متن درس مساله SVM را در حالت Separable به صورت یک مساله بهینه سازی بیان نمودیم و چند فرم معادل برای این مساله نوشتیم. فرم اول و فرم پنجم این مساله به نحوی که در کلاس بحث شد، به صورت زیر است:

$$\max_{(w,b)} d = \min_i \frac{1}{||w||} |w^T x_i + b| \quad (1)$$
$$such\ that \quad y_i(w^T x_i + b) > 0\ , \quad i = 1, \dots, m$$

$$\min ||w||^2 \quad (5)$$
$$such\ that \quad y_i(w^T x_i + b) \geq 1\ , \quad i = 1, \dots, m$$

که در (1)، d فاصله نزدیکترین نقطه x_i به HP است.

در این مساله نشان می‌دهیم که با فرض Seperability ، هر نقطه بهینه مساله (5)، نقطه بهینه مساله (1) نیز هست.

الف – فرض کنید (w^*,b^*) یک نقطه بهینه برای (5) است. ثابت کنید که در اینصورت، حداقل یکی از m قید در (5)، با علامت مساوی برقرار است. یعنی یک نقطه x_j وجود دارد که برای آن $y_j w^T x_j = 1$.

ب – ملاحظه کنید که (w^*,b^*) قیود مساله (1) را برآورده می‌کند.

ج – نشان دهید مقدار d در مساله (1) به ازای $(w,b) = (w^*,b^*)$ برابر است با $d^* = \frac{1}{||w^*||}$.

د – فرض کنید $(\tilde{w},\tilde{b})$ یک نقطه بهینه برای مساله (1) باشد. مقدار d بدست آمده در این نقطه را با $\tilde{d}$ نشان می‌دهیم. ثابت کنید که $\tilde{d}$ نمی‌تواند بزرگتر از d^* باشد و بنابراین با توجه به نتیجه بند "ب"، (w^*,b^*) یک پاسخ برای مساله (1) هم هست. برای این کار از اثبات خلف استفاده کنید. فرض کنید $\tilde{d} > d^*$ باشد. (w',b') را به صورت $w' = \alpha\tilde{w}$ و $b' = \alpha\tilde{b}$ تعریف کنید که در آن $\alpha = \frac{1}{||\tilde{w}||.\tilde{d}}$. ثابت کنید که اولا (w',b') در قیود مساله (5) صدق می‌کند و ثانیا $||w'||^2 < ||w^*||^2$ که به این ترتیب فرض بهینه بودن (w^*,b^*) نقض می‌شود.

مساله T5:

در این مساله Sample Complexity مربوط به PAC Learning را در یک مدل یادگیری با فرضیات زیر بررسی می کنیم:

- شرط Realizability در مورد H برقرار است.
- H تعداد محدودی عضو دارد.
- تابع تلف $l(h,z)$ تابع دلخواهی با مقادیر بین 0 و 1 است.

$$0 \le l(h,z) \le 1 :\ \forall z \in Z, \forall h \in H$$

در حل این مساله می توانید z را به صورت زوج $z = (x,y), x \in X, y \in Y$ فرض نمایید. هر چند ضرورتی به اینکار نیست و می توان حالت عمومی تر را در نظر گرفت.

میدانید که در PAC Learning، تابع $m_H(\varepsilon,\delta)$ به عنوان حداقل تعداد لازم داده های آموزشی m، تعریف می شود به نحوی که تضمین نماید تلف حاصل $L_D(h_S) = E\big(l(h_S,z)\big)$ با احتمال حداقل $1-\delta$ از ε کمتر است. در اینجا h_S آن فرضیه (hypothesis) است که الگوریتم یادگیری مبتنی بر مینیمم سازی ریسک تجربی(ERM) از m داده آموزشی موجود در S می آموزد. در این مساله می خواهیم یک حد بالائی بر روی $m_H(\varepsilon,\delta)$ برای مدل یادگیری مورد بحث بدست آوریم. این کار را از سه طریق مختلف زیر انجام می‌دهیم و نتایج بدست آمده را مقایسه می‌کنیم:

<u>رویکرد اول:</u>

A.1. استدلال نمایید که PAC Learning حالت خاص Agnostic PAC Learning است. توضیح دهید که در این حالت خاص، در نامساوی بکار رفته در تعریف 3.4 در کتاب، جمله اول در طرف راست نامساوی چقدر می شود.

A.2. اکنون نتیجه مربوط به Agnostic PAC Learning در Corollary 4.6 از فصل 4 کتاب را بر این حالت خاص اعمال نمایید و یک حد بالائی بر روی $m_H(\varepsilon,\delta)$ بدست آورید.

<u>رویکرد دوم:</u>

در این رویکرد سعی می کنیم از برقراری شرط Realizability (به جای حالت کلی تر Agnostic) بهره جوییم و حد بالائی کوچکتری بر روی $m_H(\varepsilon,\delta)$ بدست آوریم.

B.1. نخست ملاحظه نمایید که با توجه به فرض Realizability در اینجا داریم $L_S(h^*) = L_D(h^*) = 0$، که در این رابطه $h^* = arg\min_{h\in H} L_D(h)$. حال با استفاده از این نتیجه، حد بالایی بهتری بر روی $L_D(h_S) - L_D(h^*) = L_D\,(h_S)$ (نسبت به آنچه در حالت کلی Agnostic PAC Learning یافتیم) بدست آورید.

B.2. اکنون نتیجه بگیرید که در شرایط مورد بحث، برای آنکه با اطمینان $1-\delta$، تلف فرضیه یادگیری h_S از ε کمتر باشد، کافی است تعداد داده های آموزشی m در شرط زیر صدق کند:

$$m \geq m_H^{UC}(\varepsilon,\delta)$$

B.3. با اعمال Corollary 4.6 کتاب بر نتیجه فوق، نشان دهید که حد بالائی زیر بر روی $m_H(\varepsilon,\delta)$ برقرار است:

$$m_H(\varepsilon,\delta) \leq \frac{\log\left(\frac{2|H|}{\delta}\right)}{2\varepsilon^2}$$

B.4. این نتیجه را با حد بالایی بدست آمده در بند A.2 مقایسه کنید. حد بالایی که از این طریق بدست آورده‌اید، با چه ضریبی نسبت به حد بالایی حاصل در رویکرد اول بهبود یافته است (یعنی کمتر شده است)؟

رویکرد سوم:

قبل از پرداختن به این رویکرد، بهتر است نخست بخش 2.3 کتاب را که مساله PAC Learning را برای حالت خاص Binary Classification بررسی می‌کند، مطالعه نمایید. این بخش مساله تعمیم آن نتایج برای حالت کلی تر از Binary Classification است.

در این رویکرد، موضوع را از اساس با شیوه ای متفاوت از فصل 3 و 4 و با دنبال کردن روش به کار رفته در فصل 2 بررسی می کنیم. توجه نمایید که چون شرط Realizability برقرار است، تنها تفاوت مدل حاضر با فصل دوم کتاب در این است که در اینجا تابع تلف $l\big(h,(x,y)\big)$ می تواند هر مقداری را در فاصله [0,1] اختیار کند، در حالی که در فصل 2، $l\big(h,(x,y)\big)$ تنها مقادیر صفر (برای $h(x) = y$) و یک (برای $h(x) \neq y$) را اختیار می کرد.

C.1. با توجه به این تفاوت، روابط ریاضی و استدلال های به کار رفته در دو و نیم صفحه آخر فصل دوم را به دقت بررسی کنید و مشخص نمایید کدامیک از این روابط برای مدل مورد بحث ما برقرار می مانند و کدامیک محتاج اصلاح هستند. به طور مشخص، تعیین کنید چه تغییری باید در رابطه 2.8 داد.

C.2. نشان دهید که برای یک متغیر تصادفی دلخواه α که بین 0 و 1 قرار دارد، $0 \leq \alpha \leq 1$ و متوسط آن $\bar{\alpha}$ است، همواره داریم:

$$\mathrm{Prob}[\alpha = 0] \leq 1 - \bar{\alpha}$$

C.3. η را به صورت $\eta = \mathrm{Prob}_{x\sim D}\big[l\big(h,(x,y)\big) = 0\big]$ تعریف کنید. از C.2 نتیجه بگیرید که:

$$L_{D,f}\big(h,(x,y)\big) \geq \varepsilon \rightarrow \eta \leq 1 - \varepsilon$$

C.4. با توجه به C.1 و C.3، رابطه 2.9 را برای مدل مورد بحث اصلاح نمایید و نتیجه بگیرید که رابطه زیر در اینجا نیز برقرار می‌ماند:

$$\text{Prob}[L_D(h,s) \geq \varepsilon] \ \leq |H|e^{-\varepsilon m}$$

C.5. از این رابطه نشان دهید که حد بالایی زیر بر روی $m_H(\varepsilon,\delta)$ برقرار است:

$$m_H(\varepsilon,\delta) \leq \frac{\log\left(\frac{|H|}{\delta}\right)}{\varepsilon}$$

C.6. حد بالایی فوق را با آنچه در رویکرد های اول و دوم بدست آوردید، مقایسه کنید. آیا بهبود مهمی حاصل شده است؟ بحث نمایید.

C.7.(اختیاری) اکنون فرض کنید تابع $l(h,z)$ بتواند مقادیر منفی نیز اختیار کند.

C.7.1. آیا رویکرد سوم در اینجا قابل استفاده است؟ چرا؟

C.7.2. نشان دهید که می توان از رویکرد اول و دوم (پس از اعمال تغییری مختصر) استفاده کرد.

مساله T6:

الف – بردارهای $r,u,v \in \mathbb{R}^n$ مفروض اند. یک شبکه عصبی با ورودی $x \in \mathbb{R}^n$ و با استفاده از تابع activation $\sigma = sign$ طراحی کنید به نحوی که خروجی آن $N(x)$ دو بیت باینری به صورت

$$N(x) = \begin{cases} 0\,0 & S = -3 \\ 0\,1 & S = -1 \\ 1\,0 & S = 1 \\ 1\,1 & S = 3 \end{cases}$$

باشد، که در اینجا S به صورت زیر تعریف شده است:

$$S = sign(r^T x + b) + sign(u^T x + c) + sign(v^T x + d)$$

d,c,b اعداد حقیقی مفروضی می باشند. توجه کنید که در تعریف خروجی $N(x)$ ، 0 و 1 جایگزین مقادیر معمول 1- و 1+ شده‌اند و می‌توانید به جای 0 و 1، 1- و 1+ قرار دهید.

در شبکه ای که طراحی می‌کنید، برای گره های با ورودی ثابت، ورودی را برابر 1 فرض کنید. در طرح خود سعی کنید که از کمترین تعداد گره و لایه ممکن استفاده کنید. وزن w هر لینک را بر روی شکل (بر حسب r , u , v , b , c و d) مشخص نمایید.

مساله T7:

در یک مساله Binary classification فرض کنید $\chi = \mathbb{R}^2$ و H مجموعه ای از فرضیه های مبتنی بر axis aligned rectangles یا مستطیل‌های موازی محورهای مختصات باشد. به عبارت دیگر، فرض کنید

$$H = \{h_{(a_1,a_2,b_1,b_2)}:\ \ a_1 < a_2\ , b_1 < b_2\}$$

که در آن

$$h_{(a_1,a_2,b_1,b_2)}(x^1, x^2) = \begin{cases} 1 & a_1 \le x^1 \le a_2\ \ , b_1 \le x^2 \le b_2 \\ 0 & \text{Otherwise} \end{cases}$$

به یاد آورید که H در تعریف فوق؛ همان مجموعه فرضیه هایی است که در مثال مربوط به حدس طعم انبه مورد استفاده قرار گرفت.

الف – بخش 6.3.3 کتاب را که در آن $VCdim(H)$ بدست آمده است، مطالعه کنید.

ب – با توجه به قضیه 6.8، یک حد بالایی و حد پایینی بر روی $m_H(\varepsilon, \delta)$ به دست آورید.

ج – با توجه به اینکه در حد بالایی و پایینی بدست آمده در بند قبل از ضرایب c_1 و c_2 استفاده شده است که مقادیر آنها مشخص نیست، توضیح دهید که حدهای بدست آمده چه فایده ای دارد. به عبارت بهتر، درحالیکه ضرایب نامشخص c_1 و c_2 می‌توانند هر مقداری داشته باشند، این حدهای بالایی و پایینی حاوی چه اطلاعات سودمندی هستند؟

مساله T8:

الف – مساله 3 از فصل 2 کتاب. تنها بندهای 1 و 2 مساله.

در این مساله با فرض $\chi = \mathbb{R}^2$ و H تعریف شده در مساله T7، با روش بررسی مستقیم (یعنی بدون استفاده از $VCdim(H)$ و قضیه 6.8) نشان می‌دهید که هرگاه تعداد نمونه های آموزشی در نامساوی $m > \frac{4log^4/_\delta}{\varepsilon}$ صدق نماید و شرط realizability برقرار باشد، در آنصورت با احتمال حداقل $1 - \delta$ ، ریسک واقعی فرضیه انتخاب شده بوسیله الگوریتم پیشنهادی در مساله (که یک الگوریتم ERM است) از ε کمتر خواهد بود: $L_D(h_S) \le \varepsilon$

ب – آیا رابطه $m > \frac{4log^4/_\delta}{\varepsilon}$ بیان‌گر یک حد بالایی یا یک حد پایینی بر روی $m_H(\varepsilon, \delta)$ است؟ این حد را با حد مشابه بدست آمده در سوال T7 مقایسه کنید.

مساله T9:

در متن درس توضیح دادیم که در صورتی که در یک مساله binary classification، توزیع $\mathcal{D}$ معلوم باشد، در آنصورت روش تخمین زیر (که به تخمین Bayes موسوم است) به کمترین مقدار ریسک واقعی True Risk منجر میگردد که آن را ε_{Bayes} می‌نامیم.

$$f(x) = \begin{cases} 1 & Prob\,[y = 1 \mid x] \geq 0.5 \\ 0 & Prob\,[y = 1 \mid x] < 0.5 \end{cases}$$

با نوشتن رابطه ریاضی لازم، ثابت کنید که ریسک واقعی فرضیه فوق $L_D(f)$ نسبت به ریسک واقعی هر فرضیه h دیگر $L_D(h)$ کمتر یا مساوی است: $\varepsilon_{Bayes} = L_D(f) \leq L_D(h),\ \forall h$

توضیح: می‌دانیم که در مساله Binary classification، تابع ریسک به صورت

$$l(h,(x,y)) = \begin{cases} 1 & h(x) \neq y \\ 0 & h(x) = y \end{cases}$$

تعریف می‌شود.

مساله T10:

مجموعه فرضیه H با $|H|$ محدود و بر روی دامنه χ مفروض است. مجموعه label ها را y می‌نامیم. تابع ریسک $l(h,(x,y))$ همواره مقداری بین 0 و 1 اختیار می‌کند. یک مجموعه داده آموزشی S شامل m نقطه به صورت $i.i.d$ و بر اساس توزیع دلخواه $\mathcal{D}$ در اختیار ما قرار دارد.

الف – یک فرضیه h به طور رندم از مجموعه H اختیار می‌کنیم و $L_S(h)$ را بدست آورده، آن را η می‌نامیم: $L_S(h) = \eta$. فرض کنید $0 < \eta < 0.1$ باشد. مایلیم بدانیم احتمال اینکه $L_D(h)$ از 2η بیشتر باشد چقدر است. بهترین حد بالایی را که می‌توانید بر روی $Prob\,[L_D(h) > 2\eta]$ بدست آورید.

ب – اکنون فرض کنید یک الگوریتم یادگیری A بر روی مجموعه H و با استفاده از داده آموزشی S عمل می‌کند و فرضیه $A(s) = h_s \in H$ را بدست می‌آورد. (A ضرورتا بر اساس رویکرد ERM کار نمی‌کند)، برای h_s بدست آمده داریم $L_S(h_s) = \gamma$. باز هم فرض می‌کنیم $0 < \gamma < 0.1$. بند الف را برای h_s تکرار کنید، یعنی بهترین حد بالایی که می‌توانید بر روی $Prob\,[L_D(h_s) > 2\gamma]$ بدست آورید.

ج – حال فرض کنید مساله یادگیری مورد بحث در بند الف و ب به صورت Binary classification است یعنی $y = \{\pm 1\}$. همچنین فرض کنید به نحوی بفهمیم که برای توزیع $\mathcal{D}$، ε_{Bayes} برابر 3η است. در اینصورت آیا پاسخ شما در بند الف دچار تغییر می‌شود؟ چگونه؟

د – اگر پاسخ شما در بند ج نسبت به بند الف تغییر می‌یابد، آیا این موضوع به معنی آن است که نامساوی Hoeffding (که مبنای نتیجه گیری شما در بند الف بود)، دیگر در بند ج برقرار نیست؟ توضیح دهید.

مساله T11: مساله 1 از فصل 9 کتاب

میدانیم که در رگرسیون به فرم هموژن داریم $h(x) = w^T x$. فرض کنید تابع ریسک l را به صورت نرم خطا (به جای مجذور نرم خطا) تعریف نمائیم: $l(h,(x,y)) = \|h(x) - y\|$

در نتیجه در روش ERM برای یادگیری h و یافتن w مناسب، باید مساله زیر را حل کرد:

$$\min_{w} L_s(w) = \frac{1}{m}\sum_{i=1}^{m}|w^T x_i - y_i| \qquad w\epsilon\mathbb{R}^d \ , x_i \in \mathbb{R}^d \ , y_i \in \mathbb{R} \quad (i = 1,2,\dots,m)$$

الف- آیا این مساله، یک مساله بهینه سازی خطی است؟

ب- آیا این مساله، یک مساله بهینه سازی محدب است؟

ج- آیا تابع هدف در این مساله (یعتی همان $L_s(w)$) نسبت به w در همه جا مشتق پذیر است؟

د- مساله فوق را به صورت یک مساله بهینه سازی خطی در آورید. راهنمائی: برای این کار، نخست نشان دهید که مساله بهینه سازی $\min_{w\in\mathbb{R}^d}|f(w)|$ معادل مساله زیر است:

$$\begin{aligned} &\min_{w\in\mathbb{R}^d, c\in\mathbb{R}} c \\ &s.t. \quad c \geq f(w) \\ &\qquad\ \ c \geq -f(w) \end{aligned}$$

مساله T12:

در روش Kernel گفتیم که وقتی از نگاشت $\psi: \mathcal{X} \to F$ استفاده می‌کنیم، تابع Kernel را بصورت $k(x,x') =< \psi(x), \psi(x') >$ تعریف می‌کنیم که ضرب داخلی در فضای F صورت می‌گیرد. اکنون روند معکوسی را در نظر بگیرید که در آن، نخست یک فرم مطلوب برای تابع $k(x,x')$ در نظر گرفته و سعی داریم $\psi(x)$ را بنحوی تعیین کنیم که تابع Kernel حاصل از آن برابر با $k(x,x')$ گردد. آیا این کار همواره امکان‌پذیر است؟

طبق لم 16.2 کتاب، اگر یک تابع مشخص $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ بعنوان تابع Kernel در نظر داشته باشیم، شرط لازم و کافی برای اینکه یک feature set (F) و یک نگاشت $\psi(x)$ وجود داشته باشد بنحوی که تابع Kernel مربوطه $k(x,x')$ باشد، آنست که:

1- $k(x,x')$ متقارن باشد، یعنی $k(x,x') = k(x',x), \ \forall\, x', x \in \mathcal{X}$.

2- $k(x,x')$ positive semi-definite باشد، یعنی

$$\sum_{j=1}^{m}\sum_{i=1}^{m}\alpha_i\alpha_j k(x_i,x_j) \geq 0, \forall\, x_i \in \mathcal{X}\ ,\ \alpha_i \in \mathbb{R}\ , i = 1, \ldots, m$$

الف: در این مساله لازم بودن دو شرط فوق را نشان دهید.(کافی بودن این دو شرط در کتاب اثبات شده است)

ب: نشان دهید که شرط دوم معادل آنست که ماتریس گرام G نظیر m ورودی $x_1, x_2, \ldots, x_m$ یک ماتریس (psd) positive semi-definite باشد.

مساله T13:

در مثال مربوط به کاربرد روش Kernel برای تشخیص ویروس در یک فایل، یک فایل x به صورت دنباله‌ای از تعداد l کاراکتر درنظر گرفتیم که l میتواند حداکثر برابر مقدار مفروض d باشد: $l \leq d$. مجموعه فایلهای ممکن را با $\mathcal{X}_d$ نشان می‌دهیم. فرض کنید هر کاراکتر در x بتواند یکی از k حالت ممکن را اختیار نماید. همچنین ویروس v خود میتواند هر یک از دنباله‌های موجود در $\mathcal{X}_d$ باشد: $v \in \mathcal{X}_d$. فرض ما در طول این مثال آن است که تنها با یک ویروس v مواجه هستیم که سعی داریم از طریق "یادگیری" آن را پیدا نماییم. در این مثال، feature space را به صورت $F = \mathbb{R}^s$ درنظر گرفتیم که s برابر تعداد اعضای $\mathcal{X}_d$ میباشد: $s = |\mathcal{X}_d|$. نگاشت $\psi(x)$ به این نحو تعریف گردید که هر مولفه $\psi^u(x)$ برابر "1" یا "0" میباشد.(توجه نمایید که به ازای هر دنباله $u \in \mathcal{X}_d$ یک مولفه $\psi^u(x)$ داریم.) $\psi^u(x)$ در صورتی برابر "1" میباشد که u یک زیر دنباله یا substring از x باشد.

الف) $|\mathcal{X}_d|$ را برحسب d, k تعیین نمایید و نشان دهید که تابعی نمایی از d است.

ب) فرض کنید ویروس v را می‌شناسیم. در اینصورت $b \in \mathbb{R}$ و $w \in F$ را به نحوی مشخص نمایید که اولا نرم w واحد باشد: $\|w\| = 1$؛ ثانیا فایلهای x دارای ویروس و بدون ویروس با margin برابر $\frac{1}{2}$ از هم جدا شوند، یعنی داشته باشیم

$$\langle w, \psi(x)\rangle + b \geq \frac{1}{2} \quad \text{اگر ویروس } v \text{ در } x \text{ وجود دارد}$$

$$\langle w, \psi(x)\rangle + b \leq -\frac{1}{2} \quad \text{اگر ویروس } v \text{ در } x \text{ وجود ندارد}$$

ج) برای یک فایل x با طول l، نرم $\|\psi(x)\|$ را برحسب l بدست بیاورید. آنگاه نتیجه بگیرید که $\|\psi(x)\| = O(d)$.

د) از پاسخ خود به بند قبل چه نتیجه‌ای می‌گیرید؟ آیا مساله یادگیری به شکل یک مساله separable در فضای F در می‌آید یا نه؟ از کدامیک از دو روش Hard SVM و Soft SVM میتوان در فضای F برای تعیین w و b استفاده کرد؟

ه) نشان دهید که تابع $k(x, x')$ برابر تعداد زیر دنباله‌های مشترک (common substring) بین x و x' می‌باشد.

(اختیاری) و) الگوریتمی برای محاسبه $k(x, x')$ پیشنهاد کنید (لازم نیست بهترین الگوریتم ممکن را بیابید) و نشان دهید که پیچیدگی محاسباتی اینکار از $O(d^4)$ یا بهتر می‌باشد.

ز) آیا شرط (γ, ρ) - separability در این مساله برقرار است؟ مقدار ρ و γ را بدست آورید.

ح) در صورتیکه از روش Hard SVM برای حل این مساله استفاده نماییم و داده‌ی آموزشی ما شامل m فایل $x_1, \dots, x_m$ (همراه با y نظیر آنها) باشد، ریسک تجربی حاصل (یعنی $L_S(w,b) = \frac{1}{m}\sum_{i=1}^{m} \ell^{0-1}((w,b),(x_i,y_i))$) چقدر است؟ همچنین یک حد بالا بر روی ریسک حقیقی $L_D(w,b)$ (که با احتمال $1-\delta$ برقرار است) تعیین نمایید.(راهنمایی: به قضیه ۱۵.۴ توجه نمایید)

ط) فرض کنید مایل باشیم $L_D(w,b)$ از یک درصد تجاوز نکند، اگر طول ماکسیمم هر فایل x، $d = 1000$ و $k = 256$ باشد، با فرض $\delta = 0.01$، sample complexity مساله (یعنی حداقل m لازم) بر اساس نتیجه بند ز چقدر است؟ بحث کنید!

ی) در صورتیکه از یک الگوریتم یادگیری برای تعیین w, b استفاده کنیم که به طبقه بندی خطی در فضای F بیانجامد ولی margin اعمال نگردد، در این حالت حداقل m لازم را به ازای مقادیر عددی مذکور در بند ط بدست آورید و با m بدست آمده در آنجا مقایسه کنید.

ک) اکنون روش مورد بحث برای تشخیص ویروس مبتنی بر نگاشت به فضای F را کنار می‌گذاریم. در فضای $\mathcal{X}$، یک مجموعه فرضیه H در نظر بگیرید که به ازای هر ویروس v یک فرضیه h_v دارد که به صورت زیر تعریف شده است:

$$h_v(x) = \begin{cases} 1 & v \text{ is a substring of } x \\ 0 & otherwise \end{cases}$$

ملاحظه نمایید که $|H| = |\mathcal{X}_d|$. آیا شرط Realizability در مورد H برقرار است؟ sample complexity مساله را برحسب k, d, ϵ, δ بدست آورید. این sample complexity مبتنی بر استفاده از کدام روش یادگیری است؟ نتیجه را به ازای مقادیر داده شده در بند ط محاسبه و با sample complexity بدست آمده در آنجا مقایسه کنید.

مساله T14:

طی درس ملاحظه کردیم که هرگاه در یک مساله طبقه بندی توزیع D را بدانیم، میتوانیم بر اساس آن، تابع احتمال مشروط $P(y|x)$ را بدست آوریم و از آنجا بهترین تخمین ممکن $\hat{y}$ برای هر x را تعیین نماییم. در مساله T9 ملاحظه کردید که ریسک واقعی این روش(که آنرا روش تخمین Bayes می نامیم) از هر روش تخمین دیگری کمتر است.

در این مساله، روش تخمین Bayes را برای یک مساله رگرسیون با تابع ریسک mean square بررسی میکنیم. برای سهولت فرض میکنیم $\mathcal{Y} = \mathbb{R}$ و $\mathcal{X}$ مجموعه‌ای دلخواه است.

الف) یک فرضیه $h(x)$، $h: \mathcal{X} \to \mathbb{R}$ و تابع ریسک حقیقی مربوط به آن $L_D(h) = \mathbb{E}_D\left[\ell\left(h,(x,y)\right)\right] = \mathbb{E}_D\left[\left(h-(x,y)\right)^2\right]$ را در نظر بگیرید. می‌خواهیم تابع $h(x)$ را (از میان تمام توابع ممکن از $\mathcal{X}$ به $\mathbb{R}$) طوری تعیین کنیم که ریسک حقیقی $L_D(h)$ مینیمم گردد. برای بهره بردن از دانش خود در مورد توزیع D (و تابع چگالی احتمال مشروط $P(y|x)$ که از D به دست می‌آید) $L_D(h)$ را به صورت زیر می نویسیم:

$$L_D(h) = \mathbb{E}_x \mathbb{E}_y\left[\left(y - h(x)\right)^2 \middle| x\right]$$

در اینجا نخست مقدار مفروضی برای x درنظر گرفته میشود و متوسط آماری ریسک $\ell(h,(x,y)) = (y-h(x))^2$ نسبت به متغیر تصادفی y مشروط به مقدار مفروض x محاسبه میشود و آنگاه از نتیجه بدست آمده نسبت به متغیر x متوسط گرفته میشود. اکنون توضیح دهید چرا برای یافتن بهترین $h(x)$ به نحوی که $L_D(h)$ را مینیمم نماید، کافی است $h(x)$ را به نحوی تعیین کنیم که $\mathbb{E}_y\left[\left(y-h(x)\right)^2\middle|x\right]$ را مینیمم نماید. این تابع بهینه $h(x)$ را که مبتنی بر توزیع D است، $h_D(x)$ می‌نامیم.

ب) x را مقدار مفروضی درنظر بگیرید و عبارت $\mathbb{E}_y\left[\left(y-h(x)\right)^2\middle|x\right]$ را بسط دهید. آنگاه $h_D(x)$ را به نحوی تعیین کنید که این عبارت مینیمم گردد. نشان دهید که پاسخ به صورت $h_D(x) = \mathbb{E}_y[y|x]$ است، یعنی وقتی توزیع D را بدانیم، بهترین تخمین برچسب y برای هر x، متوسط آماری y مشروط به آن مقدار x است.

ج) ملاحظه نمایید که با انتخاب $h_D(x) = \mathbb{E}_y[y|x]$ نتیجه می‌شود

$$\mathbb{E}_y\left[\left(y-h(x)\right)^2\middle|x\right] = Variance(y|x)$$

که طرف راست را با $\sigma^2_{y|x}$ نشان میدهیم. و در نتیجه

$$L_D(h_D) = \mathbb{E}_x\left[\sigma^2_{y|x}\right]$$

که مقدار ریسک مینیمم فوق را ϵ_{Bayes} می‌نامیم.

د) اگر توزیع D به نحوی باشد که بر اساس آن همواره یک رابطه قطعی $deterministic$ به صورت $y = f(x)$ بین x و برچسب آن برقرار است (که طبیعتا f از توزیع D بدست می‌آید)، در این حالت $h_D(x)$ و $L_D(h_D)$ را برحسب $f(x)$ ساده نمایید.

ه) اگر برعکس بند د، توزیع D به نحوی باشد که بر اساس آن y, x از نظر آماری از هم مستقل باشند، در این حالت روابط بدست آمده در بند ج برای $h_D(x)$ و $L_D(h_D)$ را ساده نمایید.

و) حال فرض کنید y, x از هم مستقل نیستند، اما ما به جای استفاده از تابع بهینه $h_D(x)$، برای همه xها یک برچسب یکسان $h(x) = c$ به کار ببریم. در اینصورت $h(x) = c$ را به نحوی تعیین کنید که ریسک حقیقی $L_D(h)$ مینیمم گردد. ریسک حقیقی بدست آمد به ازای بهترین c چقدر است؟

ز) با توجه به پاسخ بدست آمده در بند و، به نظر شما برای یک توزیع دلخواه D، کدام یک از روابط زیر بین $\sigma^2_y, \epsilon_{Bayes} = \mathbb{E}_x\left[\sigma^2_{y|x}\right]$ برقرار است؟

$$\epsilon_{Bayes} = \mathbb{E}_x\left[\sigma^2_{y|x}\right] \begin{matrix} \leq \\ = \\ \geq \end{matrix} \sigma^2_y$$

توضیح دهید.

ح) اکنون یک مجموعه فرضیه H در نظر بگیرید. می‌دانیم که در یادگیری براساس داده آموزشی S (و در غیاب اطلاع از D) سعی میکنیم حتی‌الامکان بهترین h را از مجموعه H انتخاب کنیم. به نظر شما آیا رابطه زیر درست است؟ چرا؟

$$\min_{h\in H} L_D(h) \geq \epsilon_{Bayes} = \mathbb{E}_x\left[\sigma^2_{y|x}\right]$$

ط) حال فرض کنید در مورد H شرط $realizability$ برقرار است. در اینصورت سمت چپ رابطه فوق برابر چه مقدار است؟ از اینجا مقدار ϵ_{Bayes} را در این حالت بدست آورید.

ی) با توجه به بند د، توضیح دهید که چرا وقتی برای یک مجموعه H شرط $realizability$ برقرار است، داریم $\epsilon_{Bayes} = 0$.

ک) فرض کنید فرضیه h_S بر اساس داده آموزشی S و از میان مجموعه فرضیه H بدست آمده است.(شرط $realizability$ در مورد H برقرار نیست.) در رابطه ۵.۷ کتاب داشتیم که

$$L_D(h_S) = \epsilon_{app} + \epsilon_{est}, where \ \ \epsilon_{app} = \min_{h \in H} L_D(h)\,, \epsilon_{est} = L_D(h_S) - \epsilon_{app}$$

اکنون با توجه به نامساوی بند ح، رابطه فوق زا به صورت مجموع سه جمله که هر سه مثبت هستند می‌نویسیم:

$$L_D(h_S) = \epsilon_{Bayes} + (\epsilon_{app} - \epsilon_{Bayes}) + \epsilon_{est}$$

نخست بگویید که چرا جمله دوم مثبت است، آنگاه مفهوم و نقش هر یک از سه جمله فوق در ایجاد خطای حقیقی $L_D(h_S)$ را توضیح دهید.

مساله T15:

صحت رابطه زیر را نشان دهید.

$$E\ \ [L_s(h)] = \ L_D(h)$$

مساله T16:

x_1	x_2	x_3	x_4	y
1	0	1	0	0
1	1	0	0	0
0	0	1	1	0
1	0	0	1	1
0	1	1	1	1
0	0	0	1	1

یک مساله یادگیری به روش درخت تصمیم گیری برای $\chi = \{0,1\}^4$ و $y = \{0,1\}$ در نظر بگیرید. داده آموزشی S مطابق جدول مقابل می باشد.

الف- با به کار بردن الگوریتم ID3 درخت تصمیم گیری h_s را بدست آورید. برای این درخت $L_S(h_S)$ چقدر است؟

ب- اکنون با توجه به جدول S، y را به صورت یک تابع منطقی از x_1, x_2, x_3, x_4 (در واقع به صورت جمع منطقی چند ضرب منطقی) بنویسید. آنگاه با توجه به عبارت منطقی بدست آمده، یک درخت تصمیم گیری h بسازید به نحوی که خطای آن $L_S(h)$ صفر باشد.

ج- چرا در حالت عمومی از روش بند ب که خطای تجربی حاصل از آن صفر است، استفاده نمی شود؟

مساله T17:

کلاس H را که شامل k فرضیه می‌باشد دو نظر بگیرید. در مجموعه داده S و T هر یک شامل به ترتیب m_s و m_t داده (x, y) که همگی به طور مستقل از هم و بر مبنای توزیع D انتخاب شده اند، در دست می‌باشد. الگوریتم یادگیری A فرضیه $h_s \in H$ را بر مبنای داده S یادگیری می‌کند. به عبارت دیگر $A(S) = h_s$.

الف – خطاهای $\theta_i = l(h_s, (x_i, y_i))$ را به ازای داده های (x_i, y_i) مختلف در مجموعه S در نظر بگیرید. آیا این خطاها از همدیگر مستقل هستند؟ چرا؟

ب – خطاهای $\theta_i = l(h_s, (x_i, y_i))$ را به ازای داده های (x_i, y_i) مختلف در مجموعه T در نظر بگیرید. آیا این خطاها از همدیگر مستقل هستند؟ چرا؟

ج - بهترین مقدار ϵ_1 را در رابطه زیر به نحوی تعیین کنید که این رابطه با احتمال $1 - \delta$ یا بیشتر برقرار باشد:

$$|L_D(h_s) - L_S(h_s)| < \epsilon_1$$

$L_D(h_s)$ خطا یا ریسک حقیقی h_s و $L_S(h_s)$ خطای تجربی بر روی مجموعه S می‌باشد.
در این سوال منظور از بهترین ϵ_1 چیست؟

د - بهترین مقدار ϵ_2 را در رابطه زیر به نحوی تعیین کنید که این رابطه با احتمال $1 - \delta$ یا بیشتر برقرار باشد:

$$|L_D(h_s) - L_T(h_s)| < \epsilon_2$$

در اینجا $L_S(h_s)$ خطای تجربی بر روی مجموعه T می‌باشد.

ه - فرض کنید $m_s = m_t = m$. با این فرض مقادیر ϵ_1 و ϵ_2 بدست آمده در بند ج و د را مقایسه کنید.

و - با توجه به مقایسه بند ه توضیح دهید که برای تخمین خطاهای واقعی $L_D(h_s)$ ، مزیت استفاده از مجموعه داده T (که از آن در یادگیری h_s استفاده نشده باشد) چیست.

مساله T18:

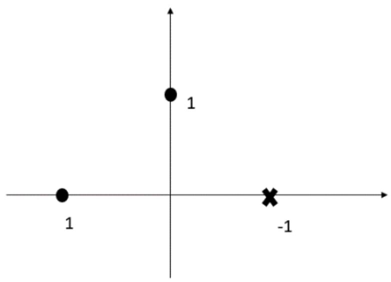
یک مساله یادگیری (به صورت طبقه بندی یا رگرسیون) روی مجموعه مشخصه های $\boldsymbol{X}$ و مجموعه برچسب های Y در نظر بگیرید. مجموعه آموزشی S شامل m نقطه $(x_i, y_i), i = 1, \ldots, m$ در اختیار است.

الف – میدانیم که برای توزیع D در حالت کلی $L_S(h)$ و $L_D(h)$ برابر نیستند. برای چه توزیع D تساوی $L_D(h) = L_S(h)$ به ازای هر فرضیه $h: X \to Y$ برقرار است؟ تذکر : برای مشخص کردن D ، D_x و $D_{y|x}$ را تعیین نمایید.

ب – یک فرضیه دلخواه h در نظر بگیرید. روشن است که مقادیر تلف $x_i = l\big(h, (x_i, y_i)\big), i = 1,2, \ldots, m$ برای h قابل محاسبه اند. توزیع D را به نحوی تعیین نمایید که $L_D(h)$ را بتوان برحسب مقادیر $\lambda_i, i = 1, \ldots, m$ بدست آورد. تذکر : کلی‌ترین فرم توزیع D با این ویژگی را بدست آورید. باز هم برای مشخص کردن D باید D_x و $D_{y|x}$ را مشخص نمایید.

مساله T19:

برای یادگیری یک مساله طبقه بندی خطی در فضای $X = \mathbb{R}^2$، مجموعه آموزشی S شامل $m = 3$ نقطه زیر در دست است:



$$x_1 = (1,0)^T, y_1 = -1$$
$$x_2 = (0,1)^T, y_1 = +1$$
$$x_3 = (-1,0)^T, y_1 = +1$$

الف – فرض کنید برای حل مساله $Soft\ SVM$ از الگوریتم SGD (صفحه 213 کتاب) با $\lambda = \frac{1}{2}$ استفاده نماییم. 6 گام الگوریتم را تا محاسبه ω^7 به صورت دستی انجام دهید. فرض کنید نمونه (x_i, y_i) که در هر گام الگوریتم از مجموعه S به صورت رندم انتخاب می‌شود به ترتیب (از راست به چپ) برابر است با $i = 2$ ، $i = 3$ ، $i = 2$ ، $i = 1$ ، $i = 3$ و $i = 1$ $\bar{\omega}$ را محاسبه و HP نظیر آن را در صفحه X در کنار نقاط x_1 ، x_2 و x_3 نمایش دهید.

ب – حال فرض کنید برای حل مساله $Soft\ SVM$ از الگوریتم GD (معادله 14.1 کتاب) استفاده نماییم و البته به جای گرادیان در این رابطه از سابگرادیان $v^t = \frac{1}{m}\sum_{i=1}^{m} v_i^t$ استفاده کنیم که v_i^t بردار سابگرادیان مربوط به ترم خطای $l(\omega, (x_i, y_i))$ در نقطه $\omega = \omega^t$ می‌باشد (رجوع به مثال 14.2) . با شروع از نقطه $\omega^1 = 0$ ، به صورت دستی 4 گام الگوریتم را تا رسیدن به نقطه ω^5 انجام دهید. $\bar{\omega}$ را محاسبه و HP نظیر آن را در صفحه X ترسیم کنید.