دانشگاه صنعتی شریف	**مقدمه ای بر یادگیری ماشین ۲۵۷۳۷**
دانشکده مهندسی برق	**گروه ۱**
نیمسال بهار ۹۸-۹۷	**مدرس: سید جمال الدین گلستانی**

تکلیف شماره ۲

موعد تحویل: یکشنبه ۹۷/۱۲/۱۹ ساعت ۱۰ صبح

توضیحات کلی

- در صورتی که برای عضو شدن در سایت‌های درس بر روی piazza.com و quera.ir یا برای آپلود کردن تکالیف خود دچار مشکل شدید، با آدرس ایمیل attarisadegh@yahoo.com تماس بگیرید.
- هر دو بخش کامپیوتری و تئوری هر تکلیف را بر روی سایت آپلود نمایید. تحویل به صورت کاغذی لازم نیست.
- در مورد هر تکلیف، هر یک از قسمت های مربوط به سوال کامپیوتری را در یک فایل به نام HWCijN.py و تمام فایلهای مربوط به سوالات تئوری را در فایلی به نام HWTiN.zip قرار دهید که i شماره تکلیف، j شماره هر قسمت از سوال کامپیوتری و N شماره دانشجویی شماست.
- به دلیل قابلیت‌های سایت piazza.com، از این سایت برای مدیریت سوال‌های مطرح شده استفاده می‌گردد. سوالات خود را تنها از طریق این سایت بفرستید و از سایت quera.ir صرفا برای آپلود تکالیف خود استفاده کنید. در صورت ایمیل کردن تکالیف به دستیاران آموزشی، نمره‌ای به آن تعلق نمی‌گیرد.
- برای سوال‌های کامپیوتری از زبان برنامه نویسی پایتون یا متلب استفاده کنید.

سوالات تئوری

سوال T4:

برای یادگیری یک مساله Binary Classification در فضای $\chi = \mathbb{R}^2$ ، مجموعه داده آموزشی S شامل ۶ 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 ۲ در χ به خوبی صورت می‌گیرد و داده آموزشی فوق با منحنی درجه ۲ 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.

سوال T5:

مساله ۲ از فصل ۱۵ کتاب.

سوال T6:

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:

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

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

که در (۱)، d فاصله نزدیکترین نقطه x_i به HP است.

در این مساله نشان می‌دهیم که هر نقطه بهینه مساله (۵)، نقطه بهینه مساله (۱) نیز هست.

الف – فرض کنید (w^*,b^*) یک نقطه بهینه برای (۵) است. ثابت کنید که در اینصورت، حداقل یکی از m قید در (۵)، 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 (۱) را برآورده می‌کند.

ج – نشان دهید مقدار d در مساله (۱) 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 (۱) 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 (۱) 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 (۵) صدق می‌کند و ثانیا $||w'||^2 < ||w^*||^2$ که به این ترتیب فرض بهینه بودن (w^*,b^*) نقض می‌شود.

سوال T7: (این مساله را بعد از درس یکشنبه ۱۲/۱۲ حل نمایید.)

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:

رویکرد اول:

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.

رویکرد دوم:

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:

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

C.3. η را به صورت $\eta = \text{Prob}_{x\sim D}\big[l\big(h,(x,y)\big) = 0\big]$ 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 ۲,۹ را برای مدل مورد بحث اصلاح نمایید و نتیجه بگیرید که رابطه زیر در اینجا نیز برقرار می-ماند:

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

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.

سوالات کامپیوتری

سوال C3:

در این سوال طبقه بندی خطی Linear classification را روی یک مجموعه داده آموزشی به چهار روش مختلف اعمال و نتیجه کار در هر روش را بر اساس یک مجموعه داده تست می‌سنجیم و با هم مقایسه میکنیم. از ۴ فایل داده که در اختیار شما قرار داده شده است، دو فایل شامل بردار سه بعدی x و برچسب $y = 1$ یا y-1 برای مجموعه آموزشی، و دو فایل دیگر حاوی همین اطلاعات برای مجموعه تست است. برای خواندن از دیتا بیس ها می توانید از کد موجود در فایل های Question3.py و Question3.m استفاده کنید

الف- در نخستین روش، با اعمال الگوریتم Perceptron بر داده آموزشی، بردار w مربوط به طبقه بندی خطی را بدست می-آوریم. البته قبل از این کار، لازم است به هر داده آموزشی یا تست x یک مولفه جدید با مقدار 1 اضافه کنید و به این ترتیب داده ها را چهار بعدی کنید که در این حالت آنها را x' مینامیم. هدف از اضافه کردن این بعد جدید چیست؟ حال کد لازم برای الگوریتم Perceptron را نوشته و آن را تا ۱۰۰۰۰ بار اجرا کنید. به ازای هر ۵۰۰ بار اجرا، خطا روی داده تست را محاسبه و ذخیره کنید و در آخر نمودار مقدار خطا بر حسب تعداد iteration ها را بکشید. بعد از پایان یادگیری، مقدار خطا روی داده تست را گزارش کنید. (منظور از خطا در این روش، نسبت تعداد داده هایی است که برچسب آنها اشتباه مشخص شده به تعداد کل داده ها.)

ب-در دومین روش، نخست هر داده x' (اعم از آموزشی یا تست) را بوسیله تبدیل $\psi(.)$ به داده چهار بعدی جدید $\psi(x')$ تصویر میکنیم. مولفه صفرم، اول، و دوم $\psi(x)$ برابر مولفه های نظیر در x' و مولفه سوم آن برابر مولفه سوم x به توان ۳ میباشد. بعد از این تغییر، بند الف را برای داده‌های بدست آمده جدید تکرار کنید.

ج- در سومین روش، الگوریتم SVM را که سعی میکند بهترین فاصله را بین نقاط و HP بدست آمده حفظ نماید، به کار می-بریم. برای این کار لازم نیست کدی بنویسید بلکه الگوریتم را با استفاده از کتابخانه های آماده بر روی داده اولیه x پیاده سازی

کنید. توجه کنید که کتابخانه های آماده موجود براساس همان فرم اولیه x که دارای سه بعد است (و نه x') عمل میکنند. خطا روی داده تست و آموزشی را گزارش دهید.

د- بالاخره در چهارمین روش، الگوریتم SVM را بر روی $\psi(x)$ اعمال نمایید. $\psi(x)$ مشابه $\psi(x')$ است جز این که دارای سه بعد است و فاقد مولفه صفرم میباشد. خطا روی داده تست و آموزشی را گزارش دهید.

ه-در پایان، بردارهای W و خطای بدست آمده در چها روش فوق را مورد بحث و بررسی و مقایسه قرار دهید.