دانشگاه صنعتی شریف	**مقدمه ای بر یادگیری ماشین 25737**
دانشکده مهندسی برق	**گروه 1**
نیمسال بهار 98-97	**مدرس: سید جمال الدین گلستانی**

تکلیف شماره 3

موعد تحویل: یکشنبه 98/1/18 ساعت 10 صبح

برای مطالعه - قسمت‌های زیر از کتاب درسی تا کنون مورد بحث قرار گرفته است:

- فصل 1 تا 4 به طور کامل
- فصل 5، بخش 5.2
- فصل 6، بخش های 6.2 و 6.3 و 6.4 (به جز corollary 6.4 و Theorem 6.6)
- فصل 9 به طور کامل
- فصل 11 بخش 11.2 (به جز 11.2.4)
- فصل 15 تا ابتدای 15.1.1
- فصل 20 بخش های 20.1 ، 20.2 و 20.3 (به جز قضایای 20.2، 20.3 و 20.5)

این تکلیف تنها شامل سوالات تئوری است.

سوال T8

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.

سوال T9

مساله 9 از فصل 6 کتاب

سوال T10

مساله 1 از فصل 6 کتاب

سوال T11

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

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

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؟

سوال T12

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

در این مساله، برای مساله مفروض در سوال T11، یعنی با فرض $\chi = \mathbb{R}^2$ و H تعریف شده در سوال T11، 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 T11 مقایسه کنید.

سوال T13

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.

سوال T14

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 بفهمم 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.