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 ۵

موعد تحویل: یکشنبه 98/3/5

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، تمام فایل‌های مربوط به سوالات کامپیوتری را در یک فایل به نام HWCiN.zip و تمام فایلهای مربوط به سوالات تئوری را در فایلی به نام HWTiN.zip قرار دهید که i شماره تکلیف 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 T20:

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

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

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 $x_1, x_2, \dots, x_m$ یک ماتریس (psd) positive semi-definite باشد.

سوال T21:

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 $\psi^{\mathrm{u}}(\mathrm{x})$ برابر "1" یا "0" میباشد.(توجه نمایید که به ازای هر دنباله $u \in \mathcal{X}_d$ یک مولفه $\psi^{\mathrm{u}}(\mathrm{x})$ داریم.) $\psi^{\mathrm{u}}(\mathrm{x})$ 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 صوریتکه 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.

سوال T22:

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 T13 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 $L_D(h) = \mathbb{E}_D\big[\ell\big(h,(x,y)\big)\big] = \mathbb{E}_D\left[\big(h-(x,y)\big)^2\right]$ 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 $\ell\big(h,(x,y)\big) = (y-h(x))^2$ 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.

سوال T23:

مساله ۱ از فصل ۱۲ کتاب