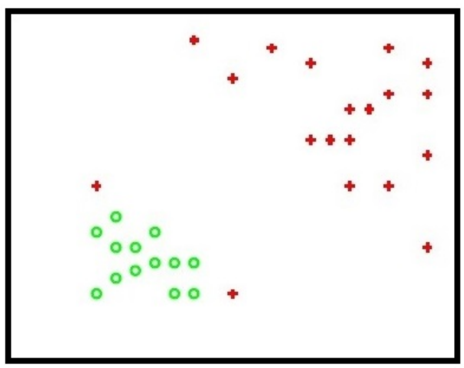
دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین		سه‌شنبه ۱ تیر ۱۴۰۰
	امتحان پایان‌ترم	
مدرس: دکتر محمّدحسین رهبان		زمان امتحان: ۱۰ + ۱۴۰ دقیقه

سوال ۱ *SVM & Feature Space* (۱۰ نمره، زمان پیشنهادی ۲۰ دقیقه)

آ) فرض کنید داده‌های شکل زیر را برای یادگیری یک دسته‌بند دودویی با استفاده از الگوریتم SVM با هسته چندجمله‌ای [1] درجه ۲ بکار ببریم.



۱. مرز تصمیم‌گیری را به ازای $C \to 0$ و $C \to \infty$ در دو شکل جدا رسم کنید. در کدام حالت به جواب بهتری می‌رسیم؟ در قسمت ۲ از این حالت استفاده کنید. (۲ نمره)

۲. یک داده آموزش جدید به شکل بالا اضافه کنید که تغییری در مرز تصمیم‌گیری ایجاد نکند. همچنین یک داده آموزش غیربدیهی دیگر به شکل بالا اضافه کنید که مرز تصمیم‌گیری را تغییر دهد. سپس مرز تصمیم‌گیری جدید را رسم کنید. (۲ نمره)

ب) دو عضو $\mathbf{a} = [a_1, a_2]^\top$ و $\mathbf{b} = [b_1, b_2]^\top$ از مجموعه دادگان $\mathcal{D} \subseteq \mathbb{R}^2 \times \{-1, 1\}$ را در نظر بگیرید.

۱. نشان دهید فضای ویژگی‌ای که هسته $k_1(\mathbf{a}, \mathbf{b}) = \left(\mathbf{a}^\top \mathbf{b}\right)^2$ تعریف می‌کند به شکل زیر است: (۳ نمره)

$$\begin{pmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{pmatrix}$$

۲. یک فضای ویژگی برای هسته $k_2(\mathbf{a}, \mathbf{b}) = \left(\mathbf{a}^\top \mathbf{b} + 1\right)^2$ بدست آورید. (۳ نمره)

[1] Polynomial

سوال ۲ *Linear Regression - PCA* (۱۲ نمره، زمان پیشنهادی ۳۰ دقیقه)

ماتریس $X \in \mathbb{R}^{D\times N}$ شامل N داده آموزش Dبعدی را به همراه بردار $Y \in \mathbb{R}^{N}$ که بیانگر برچسب متناظر با داده‌های آموزش است در نظر بگیرید.

آ) به صورت خلاصه و با بیان رابطه‌های ریاضی شرح دهید که مهمترین ویژگی‌هایی که باعث می‌شود از توزیع‌های گاوسی برای مدل در روش تمام بیزی [1] برای حل مساله استفاده کنیم، کدامند؟ (۱.۵ نمره)

ب) فرض کنید می‌توانیم مقدار و بردار ویژه‌های ماتریس $A \in \mathbb{R}^{D\times D}$ را در زمان $O\left(D^3\right)$ بدست آوریم. از طرفی می‌دانیم با دسترسی به بردار ویژه‌های ماتریس $XX^\top$ می‌توانیم با استفاده از روش PCA عملیات کاهش بعد را برای مسئله انجام دهیم. در حالتی که تعداد ویژگی‌های مسئله خیلی بیشتر از تعداد نمونه‌های آموزش است، بدست آوردن مقدار و بردار ویژه‌های ماتریس $XX^\top$ مقرون‌به‌صرفه نیست. در این حالت چگونه می‌توانیم با در نظر گرفتن ماتریس کواریانس $X^\top X$ و با پیچیدگی زمانی $O\left(N^3\right)$ مقدار و بردار ویژه‌ها را بدست آوریم؟ لازم است رابطه‌های ریاضی این بخش را کامل بنویسید. (۲.۵ نمره)

پ) درست یا غلط بودن هر یک از گزاره‌های زیر را با بیان دلیل مشخص کنید: (۳.۵ نمره)

۱. روش PCA به دوران [2] بردارهای ورودی حساس نیست. به عبارتی اگر بردارهای ورودی را دوران دهیم، خروجی این روش تغییری نمی‌کند.

۲. اگر از روش PCA برای کاهش بعد مسئله از D به K استفاده کنیم خروجی شبیه حالتی است که ابتدا این روش را برای کاهش بعد مسئله از D به L بکار ببریم و سپس ویژگی‌های ماتریس جدید را با استفاده از این روش از L به K کاهش بعد دهیم.

۳. اگر به همه نمونه‌های ماتریس X یک بعد با مقدار ثابت ۱ اضافه کنیم تغییری در نتیجه PCA ایجاد نمی‌شود.

ت) ماتریس نمونه $X = (-2\ 2\ v)$ که $v \in \mathbb{R}_+$ را به همراه بردار برچسب $Y = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ در نظر بگیرید. در این بخش می‌خواهیم مسئله رگرسیون خطی را با تابع هدف $\mathcal{L}(h) = |h(\mathbf{x}) - y|^2$ حل کنیم. از طرفی دو انتخاب به صورت $h_0(\mathbf{x}) = b$ و $h_1(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$ برای مدل داریم. مقدار پارامتر v را به گونه‌ای تنظیم کنید که میانگین تابع هدف برای این دو مدل در حالتی که اعتبارسنجی متقابل [3] را به صورت *Leave One Out* بکار می‌گیریم، برابر شوند. (۴.۵ نمره)

سوال ۳ *Nearest Neighbour* (۱۱ نمره، زمان پیشنهادی ۲۵ دقیقه)

مجموعه‌های $\mathcal{X} = [0,1]^d$ و $\mathcal{Y} = \{0,1\}$ و تابع‌های $\eta : \mathcal{X} \to [0,1]$ و $h^\star : \mathcal{X} \to \{0,1\}$ با ضابطه‌های زیر را در نظر بگیرید.

$$\eta(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x})$$

$$h^\star(\mathbf{x}) = \mathbb{1}_{[\eta(\mathbf{x})>0.5]}$$

در واقع $h^\star$ بیانگر دسته‌بند بهینه بیز [4] است. می‌دانیم η تابعی *c-Lipschitz* است یعنی به ازای هر $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ داریم:

$$\left|\eta(\mathbf{x}) - \eta(\mathbf{x}')\right| \leq c \left\|\mathbf{x} - \mathbf{x}'\right\|$$

حال m نمونه‌ی $S = \left\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\right\}$ را به صورت مستقل با توزیع یکسان [5] $\mathcal{D}$ روی $\mathcal{X} \times \mathcal{Y}$ نمونه‌برداری می‌کنیم. می‌خواهیم

[1] Fully Bayesian
[2] Rotation
[3] Cross Validation
[4] Bayes Optimal Rule
[5] Independent & Identically Distributed (i.i.d)

حالت $k = 1$ از مسئله‌ی KNN[1] را بررسی کنیم (حالتی که تنها نزدیکترین همسایه را در نظر می‌گیریم). در ادامه خروجی این الگوریتم را با تابع $h_S : \mathcal{X} \to \mathbb{R}$ نشان می‌دهیم.

آ) خطای عمومی به صورت $L_{\mathcal{D}}\left(h_S\right) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{1}_{\left[h_S(\mathbf{x})\neq y\right]}\right]$ تعریف می‌شود. ثابت کنید:

$$\mathbb{E}_S\left[L_{\mathcal{D}}\left(h_S\right)\right] = \mathbb{E}_{S_{\mathbf{x}}\sim D_{\mathcal{X}}^m\,,\,\mathbf{x}\sim D_{\mathcal{X}}}\left[\mathbb{P}_{y\sim\eta(\mathbf{x})\,,\,y'\sim\eta(\pi_1(\mathbf{x}))}\left(y\neq y'\right)\right]$$

که $\pi_1(\mathbf{x})$ بیانگر نزدیکترین همسایه $\mathbf{x}$ است. (۲ نمره)

ب) اثبات کنید: (۴ نمره)

$$\mathbb{P}_{y\sim\eta(\mathbf{x})\,,\,y'\sim\eta(\mathbf{x}')}\left(y\neq y'\right) = 2\,\eta(\mathbf{x})\left(1-\eta(\mathbf{x})\right) + \left(\eta(\mathbf{x})-\eta(\mathbf{x}')\right)\left(2\,\eta(\mathbf{x})-1\right)$$

پ) حال با استفاده از نتیجه بخش‌های **آ** و **ب** ثابت کنید: (۳ نمره)

$$\mathbb{E}_S\left[L_{\mathcal{D}}\left(h_S\right)\right] \leq 2\,L_{\mathcal{D}}\left(h^\star\right) + c\,\mathbb{E}_{S\sim D^m\,,\,\mathbf{x}\sim D}\left[\|\mathbf{x}-\mathbf{x}_{\pi_1(\mathbf{x})}\|\right]$$

ت) با فرض $\lim_{m\to\infty}\|\mathbf{x}-\mathbf{x}_{\pi_1(\mathbf{x})}\| = 0$ به چه نتیجه‌ای می‌رسیم؟ (۲ نمره)

سوال ۴ *Semi-Supervised Learning* (۱۳.۵ نمره، زمان پیشنهادی ۳۵ دقیقه)

در یکی از روش‌های یادگیری نیمه‌نظارتی[2] تابع منظم‌سازی[3] بکار می‌رود که هم از داده‌های دارای برچسب و هم از داده‌های بدون برچسب استفاده می‌کند. در ادامه تابع هزینه زیر را در نظر بگیرید:

$$\mathcal{L}_{total} = \ell(\mathcal{D}_l, \boldsymbol{\theta}) + \alpha\,\mathcal{R}(\mathcal{D}_l, \mathcal{D}_u, \boldsymbol{\theta})$$

در رابطه بالا $\boldsymbol{\theta}$ مجموعه پارامترهای مدل، $\mathcal{D}_u$ مجموعه دادگان بدون برچسب، $\mathcal{D}_l$ مجموعه دادگان دارای برچسب و α ضریبی ثابت را نشان می‌دهند. همچنین ℓ بیانگر تابع هزینه روی داده‌های دارای برچسب می‌باشد و $\mathcal{R}$ یک تابع منظم‌ساز به شکل زیر است:

$$\mathcal{R}(\mathcal{D}_l, \mathcal{D}_u, \boldsymbol{\theta}) = \frac{1}{n_l + n_u}\sum_{\mathbf{x}\in\mathcal{D}_l\cup\mathcal{D}_u}\underbrace{\mathtt{KL}\big(p(y\mid\mathbf{x},\boldsymbol{\theta})\,\|\,p(y\mid\mathbf{x}+\mathbf{r},\boldsymbol{\theta})\big)}_{M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})}$$

که n_l تعداد داده‌های دارای برچسب، n_u تعداد داده‌های بدون برچسب و $p(y\mid\mathbf{x},\boldsymbol{\theta})$ خروجی احتمالی مدل به ازای ورودی $\mathbf{x}$ و مجموعه پارامترهای $\boldsymbol{\theta}$ را نشان می‌دهند. همچنین $\mathtt{KL}$ بیانگر تابع *Kullback-Leibler*[4] است که برای مساله دسته‌بندی C کلاسه به شکل زیر تعریف می‌شود:

$$\mathtt{KL}(p\,\|\,q) = \sum_{i=1}^{C} p_i \log\left(\frac{p_i}{q_i}\right)$$

[1] K-Nearest Neighbors
[2] Semi-Supervised Learning
[3] Regularizer
[4] Kullback-Leibler Divergence

و $\mathbf{r}$ در عبارت $M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})$ نمایانگر برداری تصادفی است. در این سوال می‌خواهیم مسئله زیر را حل کنیم:

$$\mathbf{r}^\star = \underset{\mathbf{r}\in\mathbb{R}^d}{\boldsymbol{argmax}} \quad \mathtt{KL}\big(p(y \mid \mathbf{x},\boldsymbol{\theta}) \,\|\, p(y \mid \mathbf{x}+\mathbf{r},\boldsymbol{\theta})\big)$$
$$\textbf{\textit{subject to}} \quad \|\mathbf{r}\|_2 \le \varepsilon$$

آ) چرا می‌توان $M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})$ را با استفاده از بسط تیلور تا مرتبه ۲ به شکل زیر نوشت؟ (۲ نمره)

$$M(\mathbf{x},\mathbf{r},\boldsymbol{\theta}) \approx \frac{1}{2}\mathbf{r}^\top H(\mathbf{x},\boldsymbol{\theta})\,\mathbf{r}\,, \qquad H(\mathbf{x},\boldsymbol{\theta}) = \nabla^2_{\mathbf{r}} M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=\mathbf{0}}$$

ب) با در نظر گرفتن تقریب بخش **آ** ثابت کنید $\mathbf{r}^\star$ بردار ویژه نظیر بزرگترین مقدار ویژه ماتریس H است. به عبارتی نشان دهید:

$$\mathbf{r}^\star = \underset{\mathbf{r}\in\mathbb{R}^d}{\boldsymbol{argmax}} \quad \mathbf{r}^\top H(\mathbf{x},\boldsymbol{\theta})\,\mathbf{r} = \varepsilon\overline{\mathbf{u}(\mathbf{x},\boldsymbol{\theta})}$$
$$\textbf{\textit{subject to}} \quad \|\mathbf{r}\|_2 \le \varepsilon$$

که $\overline{\mathbf{u}(\mathbf{x},\boldsymbol{\theta})}$ نمایانگر بردار ویژه نرمال شده نظیر بزرگترین مقدار ویژه ماتریس $H(\mathbf{x},\boldsymbol{\theta})$ است $\left(\bar{\mathbf{u}} \triangleq \frac{\mathbf{u}}{\|\mathbf{u}\|}\right)$. (۵ نمره)

پ) همانطور که می‌دانیم در مسئله‌های دنیای واقعی با داده‌هایی با بعد بالا سروکار داریم. در این مسئله‌ها بدست آوردن ماتریس H و بردار ویژه‌هایش مقرون‌به‌صرفه نیست و نمی‌توان به ازای هر عضو مجموعه دادگان این محاسبه‌ها را انجام داد. در چنین شرایطی از الگوریتم‌های تکراری[1] استفاده می‌کنیم. با توجه به بخش **ب** برای بردار دلخواه $\mathbf{v}_0 \in \mathbb{R}^d$ با طول واحد، با فرض همگرایی ثابت کنید: (۴ نمره)

$$\bar{\mathbf{u}} = \lim_{n\to\infty} \mathbf{v}_n\,, \qquad \forall\, n \in \mathbb{Z}_{++} :\ \mathbf{v}_n = \overline{H\,\mathbf{v}_{n-1}} \tag{۱}$$

ت) با توجه به رابطه‌هایی که در بخش‌های قبل بدست آوردیم، ثابت کنید اگر تنها یک بار از رابطه ۱ استفاده کنیم، آن‌گاه می‌توان $\mathbf{r}^\star$ را به کمک رابطه زیر حساب کرد: (۲.۵ نمره)

$$\mathbf{r}^\star \approx \varepsilon\overline{\nabla_{\mathbf{r}} M(\mathbf{x},\mathbf{r},\boldsymbol{\theta})\big|_{\mathbf{r}=\xi\mathbf{v}_0}}$$

که ξ یک ضریب ثابت کوچک است.

سوال ۵ *RL* (۱۰ نمره، زمان پیشنهادی ۳۰ دقیقه)

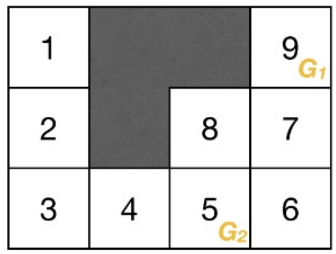
شکل زیر یک *Grid World* را نشان می‌دهد که برخی از خانه‌های آن مسدود است. عامل در هر یک از این خانه‌ها می‌تواند قرار بگیرد و هدف رسیدن به یکی از خانه‌های ۵ یا ۹ است. عامل در صورت ورود به هریک از این خانه‌ها و بدون انجام دادن کنش دیگری، امتیاز آن را دریافت می‌کنم. انواع کنش‌هایی که عامل می‌تواند در هر خانه به صورت قطعی انجام دهد به شرح زیر است:

Stay عامل در مکان فعلی خود بماند. در این صورت اگر عامل در خانه‌های ۹ یا ۵ باشد به ترتیب مقدار G_1 یا G_2 پاداش می‌گیرد. ماندن در خانه‌های دیگر ۱ واحد هزینه دارد.

[1] Iterative

Walk عامل با حرکت به سمت بالا، پایین، چپ یا راست به یکی از خانه‌های مجاور خود (در صورت مسدود نبودن) می‌رود. این کار هزینه‌ای به اندازه ۱ واحد دارد.

Jump عامل با یک حرکت L شکل (شبیه حرکت اسب در شطرنج) به یکی از خانه‌های دیگر (در صورت مسدود نبودن مسیر) می‌رود. برای مثال اگر عامل در خانه ۱ باشد می‌تواند با چنین حرکتی به خانه ۴ برود ولی به دلیل مسدود بودن مسیر نمی‌تواند به خانه ۸ برود. هزینه این حرکت ۲ واحد است.



آ) به ازای $G_1 = 10$، $G_2 = 3$ و $\gamma = 0.5$ با حساب کردن $V^\star(s)$ سیاست بهینه را بدست آورید. (۶ نمره)

ب) به ازای $G_1 = 10$ و $G_2 = 3$، در چه صورتی تغییر مقدار γ سیاست بهینه در خانه ۵ را تغییر می‌دهد؟ (۴ نمره)

پیروز باشید