دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین	پنج‌شنبه ۹ اردیبهشت ۱۴۰۰

امتحان میان‌ترم

مدرس: دکتر محمّدحسین رهبان	زمان امتحان: ۱۰ + ۱۴۰ دقیقه

سوال ۱ رگرسیون لجستیک (۲۰ نمره + ۳ نمره امتیازی، زمان پیشنهادی ۳۵ دقیقه)

تعریف: تابع $f : \mathcal{S} \to \mathbb{R}$ محدب است اگر:

$$\forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S},\ \alpha \in [0,1]:\ f\big(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2\big) \le \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2)$$

تابع $f : \mathcal{S} \to \mathbb{R}$ مقعر است اگر $-f$ محدب باشد.

فرض کنید در یک مساله دسته‌بندی احتمالاتی دودویی، برای هر نمونه $(\mathbf{x}, y)$ احتمال تعلق به هر کلاس به شکل زیر تعریف شود:

$$\mathbb{P}(y \mid \mathbf{x}; \mathbf{w}) = \begin{cases} f(\mathbf{x}; \mathbf{w}) & : \ y = +1 \\ 1 - f(\mathbf{x}; \mathbf{w}) & : \ y = -1 \end{cases}$$

که می‌دانیم $0 \le f(\mathbf{x}; \mathbf{w}) \le 1$ است و بردار $\mathbf{w}$ پارامترهای مدل را مشخص می‌کند.

آ) با در نظر گرفتن مجموعه دادگان $\mathcal{D} = \big\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\big\}$ ابتدا نشان دهید تابع لگاریتم درست‌نمایی [1] برای این مساله به شکل زیر است:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \Big[(1 + y_n) \log f(\mathbf{x}_n; \mathbf{w}) + (1 - y_n) \log \big(1 - f(\mathbf{x}_n; \mathbf{w})\big) \Big]$$

سپس نشان دهید به ازای تابع $f(\mathbf{x}; \mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x}}}$، تابع لگاریتم درست‌نمایی به شکل زیر ساده می‌شود:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \Big[-(1 + y_n) \log(1 + e^{-\mathbf{w}^\top \mathbf{x}_n}) - (1 - y_n) \log(1 + e^{\mathbf{w}^\top \mathbf{x}_n}) \Big] \tag{۱}$$

این مساله بهینه‌سازی به ازای تابع $f(\mathbf{x}; \mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x}}}$ دارای فرم بسته نیست. با این وجود می‌توان نشان داد تابع درست‌نمایی (۱) مقعر است. ویژگی پرکاربرد تابع‌های مقعر و محدب وجود تنها یک نقطه بهینه سراسری در دامنه تابع است. این ویژگی باعث می‌شود در به‌روزرسانی گام‌به‌گام پارامترهای مساله با روشی مانند *Gradient Descent* به پاسخ بهینه سراسری همگرا شویم. (۶ نمره)

[1] Log Likelihood

ب) تابع $g:\mathbb{R}\to\mathbb{R}$ محدب است، اگر و تنها اگر دو بار مشتق‌پذیر بوده و مشتق دوم آن به ازای هر $x\in\mathbb{R}$ مثبت باشد. ثابت کنید تابع $g(x)=\log(1+e^x)$ محدب است. (۴ نمره)

پ) با در نظر گرفتن دو تابع محدب $f,g:\mathbb{R}\to\mathbb{R}$ ثابت کنید صعودی بودن g یک شرط کافی برای محدب بودن ترکیب دو تابع یعنی $g\circ f$ است (راهنمایی: می‌توانید از ویژگی مثبت بودن مشتق دوم برای تابع‌های محدب استفاده کنید). (۴ نمره)

ت) ثابت کنید ترکیب خطی K تابع محدب $f_1,f_2,\ldots,f_K:\mathbb{R}\to\mathbb{R}$ با ضرایب ثابت و مثبت $c_1,c_2,\ldots,c_K$ به شکل $h=\sum_{k=1}^{K}c_kf_k$ نیز یک تابع محدب است. (۳ نمره امتیازی)

ث) نشان دهید تابع لگاریتم درست‌نمایی بخش آ (۱) مقعر است. (۶ نمره)

سوال ۲ دسته‌بندی دودویی در فضای دوبعدی (۲۰ نمره، زمان پیشنهادی ۲۰ دقیقه)

کلاس فرضیه $\mathcal{H}\subseteq 2^{\mathbb{R}^2}$ شامل همه دسته‌بندهای خطی گذرنده از مبدا را در نظر بگیرید. با این فرض که N بیانگر تعداد داده‌هاست به پرسش‌های زیر پاسخ دهید:

آ) تابع رشد $\mathcal{H}$ را برای $N\in\{1,2,3,4\}$ محاسبه کرده و دلیل انتخاب خود را بیان کنید. (۶ نمره)

ب) می‌دانیم ضابطه هر $h\in\mathcal{H}$ به شکل زیر است:

$$h(x_1,x_2)=\boldsymbol{sign}(w_1x_1+w_2x_2),\qquad (w_1,w_2)\in\mathbb{R}^2$$

دو تابع دسته‌بندی دلخواه $f,g\notin\mathcal{H}$ را در نظر بگیرید. کلاس فرضیه جدیدی به شکل $\mathcal{M}=\mathcal{H}\cup\{f,g\}$ تعریف می‌کنیم. بیشینه مقدار تابع رشد $\mathcal{M}$ را برای $N\in\{1,2,3,4\}$ به ازای همه انتخاب‌های ممکن f و g بیابید. (۴ نمره)

پ) با استفاده از مجموعه دادگان $\mathcal{D}=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_N,y_N)\}$، فرضیه $m\in\mathcal{M}$ با کمینه خطای آموزش را بدست می‌آوریم (داریم $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\overset{i.i.d}{\sim}\mathbf{P}_X$ و به ازای هر $i\in\{1,2,\ldots,N\}$ می‌دانیم y_i تابعی از $\mathbf{x}_i$ است). با چه احتمالی خطای تعمیم‌پذیری[1] دست‌کم $\frac{1}{\sqrt{2}}$ است؟ (۵ نمره)

ت) اگر خروجی الگوریتم یادگیری، فرضیه $h\in\mathcal{H}$ باشد که مقدار $\mathbf{E}_{in}(h)=\sum_{n=1}^{N}|h-y_n|$ را کمینه می‌کند، نشان دهید این الگوریتم تخمین‌گری از میانه $\{y_1,y_2,\ldots,y_N\}$ را یاد می‌گیرد. با این فرض که y_N داده‌ای خارج از محدوده[2] است، چه تغییری در خروجی الگوریتم ایجاد می‌شود (برای سادگی فرض کنید $y_N=\boldsymbol{max}\{y_1,y_2,\ldots,y_N\}$)؟ (۵ نمره)

سوال ۳ تابع هزینه (۱۵ نمره، زمان پیشنهادی ۲۰ دقیقه)

آ) به پرسش‌های زیر پاسخی کوتاه داده و خلاصه‌ای از دلیل خود را بیان کنید: (۴ نمره)

۱. چرا از مجموع مربعات به عنوان تابع خطا در مساله‌های دسته‌بندی استفاده نمی‌کنیم؟

۲. استفاده از تابع خطای *Logistic Regression* به جای *Perceptron* چه مزیت‌هایی دارد؟

[1] Generalization Error
[2] Outlier

ب) مجموعه داده‌ای N عضوی و مدلی ثابت در اختیار داریم. دو تابع هزینه‌ی میانگین مربعات [1] و میانگین قدرمطلق [2] که به شکل زیر تعریف می‌شوند را در نظر بگیرید: (۵ نمره)

$$L_{MSE}(\theta) = \frac{1}{N}\sum_{n=1}^{N}(x_n - \theta)^2, \qquad L_{MAE}(\theta) = \frac{1}{N}\sum_{n=1}^{N}|x_n - \theta|$$

۱. با کمینه کردن این دو تابع، مقدار $\theta^\star$ را برای هر یک پیدا کنید.

۲. با توجه به نتیجه قسمت قبل، خلاصه‌ای از مزایا و معایب هر تابع را بیان کنید.

پ) نشان دهید به ازای هر مجموعه داده خطی‌جداپذیر، پاسخ بیشینه درست‌نمایی [3] برای مدل *Logistic Regression*، با پیدا کردن بردار پارامتر $\mathbf{w}$ به‌طوری‌که مرز تصمیم‌گیری دو کلاس را جدا کند و سپس میل دادن اندازه‌ی $\mathbf{w}$ به بی‌نهایت، بدست می‌آید. (۶ نمره)

سوال ۴ شبکه عصبی (۱۸ نمره، زمان پیشنهادی ۳۰ دقیقه)

آ) فرض کنید می‌خواهیم تابع هدف f را تخمین بزنیم. از قبل می‌دانیم که پیچیدگی تابع هدف بالاست و دارای *Stochastic Noise* بسیار کمی است. در چه شرایطی بهتر است از مدلی نسبتا ساده استفاده کنیم (مثلا چندجمله‌ای‌های درجه ۲) و در چه شرایطی از مدلی نسبتا پیچیده (مثلا چند جمله‌ای‌های درجه ۲۰)؟ (۲ نمره)

ب) با درنظر گرفتن بردارهای ورودی $\mathbf{x}_1$ و $\mathbf{x}_2$ و ماتریس‌های وزن $\mathbf{W}_1$ و $\mathbf{W}_2$ به شکل زیر، گراف محاسباتی را برای $\mathbf{E}$ با رابطه‌ای که در ادامه می‌آید رسم کرده و با استفاده از روش *Backpropagation* مقدار گرادیان $\mathbf{E}$ را نسبت به بردارهای $\mathbf{x}_1$ و $\mathbf{x}_2$ و ماتریس‌های $\mathbf{W}_1$ و $\mathbf{W}_2$ بدست آورید.

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -2 \\ 3 \end{bmatrix}, \mathbf{W}_1 = \begin{bmatrix} -3 & 1 \\ 1 & 2 \end{bmatrix}, \mathbf{W}_2 = \begin{bmatrix} -2 & 2 \\ 3 & -1 \end{bmatrix}$$

$$\mathbf{E} = \boldsymbol{tanh}\big(0.1 \times \boldsymbol{max}(\mathbf{W}_1\mathbf{x}_1, 0) + \boldsymbol{max}(\mathbf{W}_2\mathbf{x}_2, 0)\big)$$

توجه کنید که تابع‌های $\boldsymbol{max}$ و $\boldsymbol{tanh}$ روی بردارها، عنصربه‌عنصر [4] عمل می‌کنند. (۴ نمره)

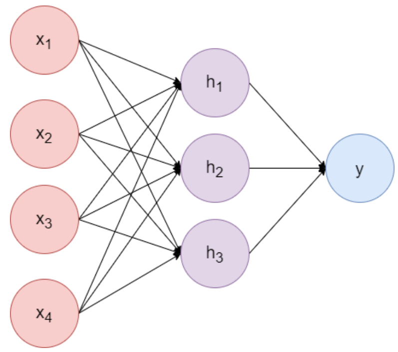
پ) فرض کنید شبکه‌ای داریم که ورودی‌های منطقی منفی یک یا یک دریافت کرده و تابع فعال‌ساز آن به ازای ورودی مثبت، عدد یک و به ازای ورودی منفی، عدد منفی یک را خروجی می‌دهد. با توجه به این که هر تابع منطقی را می توان به شکل *Sum of Products* نشان داد، بیشینه تعداد لایه‌های مورد نیاز برای نشان دادن یک تابع منطقی دلخواه چندتاست؟ مرتبه تعداد راس‌های لایه پنهان چیست؟ (۴ نمره)

ث) تابع f را در نظر بگیرید که اگر ورودی آن نامثبت باشد خروجی آن ۱ و در غیر این صورت خروجی آن صفر است. شبکه دو لایه زیر با دریافت چهار عدد حقیقی x_1، x_2، x_3 و x_4 تابع f را روی لایه میانی و نهایی اعمال می‌کند. وزن‌های شبکه را طوری تنظیم کنید که اگر ورودی در شرط زیر صدق کند شبکه خروجی یک و در غیر این صورت صفر را تولید کند. (۸ نمره)

$$x_1 \le x_2 \le x_3 \le x_4$$

---

[1] Mean Squared Error (MSE)
[2] Mean Absolute Error (MAE)
[3] Maximum Likelihood
[4] Element-Wise



سوال ۵ منظم سازی (۱۸ نمره + ۷ نمره امتیازی، زمان پیشنهادی ۳۵ دقیقه)

آ) اگر در مساله رگرسیون خطی از یک منظم‌ساز L_1 استفاده کنیم، با کاهش ضریب λ انتظار داریم که تغییرات خطای دادگان آموزش و خطای دادگان آزمون [1] چطور باشد؟ (۳ نمره)

ب) توضیح دهید که چرا استفاده از منظم‌ساز بهینه در مساله‌های یادگیری غیرواقعی و ناممکن است. اگر در مساله‌ای از یک منظم‌ساز نامناسب استفاده کنیم چه اتفاقی می‌افتد؟ (۴ نمره)

پ) برای سه بردار $\mathbf{x}_1$، $\mathbf{x}_2$ و $\mathbf{x}_3$ مقدار نرم‌های ۱ و ۲ را حساب کنید. حال با توجه به این مقدارها به صورت شهودی توضیح دهید که استفاده از کدامیک از منظم‌سازهای L_1 و L_2 برای وزن‌های مدل، تعداد وزن‌های با مقدار صفر را بیشتر می‌کند؟ (۷ نمره امتیازی)

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}, \ \mathbf{x}_2 = \begin{bmatrix} 4 \\ 3 \\ 0 \end{bmatrix}, \ \mathbf{x}_3 = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$$

ت) فرض کنید یک مساله رگرسیون خطی داریم که توزیع y برحسب $\mathbf{x}$ به صورت $\mathbf{p}(y \mid \mathbf{x}; \mathbf{w}, \beta) = f(\mathbf{x}, \mathbf{w}) + \epsilon$ بدست می‌آید که $\epsilon \sim \mathcal{N}(0, \beta^{-1})$. در اینجا $f(\mathbf{x}, \mathbf{w})$ یک تابع غیرخطی است که مقدار خروجی را به ازای ورودی $\mathbf{x}$ ارایه می‌دهد. با فرض توزیع احتمال پیشینی برای $\mathbf{w}$ به صورت $\mathbf{p}(\mathbf{w}; \alpha) = \mathcal{N}(0, \alpha^{-1} I)$، به کمک قاعده بیز مقدار بیشینه برای توزیع احتمال پسین $\mathbf{w}$ را با استفاده از لگاریتم درست‌نمایی بیشینه [2] بدست آورید. در ادامه نشان دهید این کار معادل حالتی است که خطای میانگین مربعات را کمینه کنیم و بجای داشتن توزیع پیشین $\mathbf{w}$ منظم‌ساز L_2 را بکار ببریم. (۱۱ نمره)

راهنمایی: تابع توزیع احتمال نرمال $\mathcal{N}(0, \sigma I)$ برای m متغییر به صورت $\left(\frac{1}{2\pi\sigma}\right)^{\frac{m+1}{2}} e^{-\frac{\mathbf{w}^\top \mathbf{w}}{2\sigma}}$ است.

پیروز باشید

[1] Test Data

[2] Maximum Log-Likelihood