بسمه‌تعالی

یادگیری ماشین **نیمسال دوم ۱۴۰۲-۱۴۰۱**

امتحان میان‌ترم **وقت آزمون: ۱۵۰ دقیقه**

توجه: استفاده از کتاب، جزوه، اسلایدهای درس، ماشین‌حساب و کلیه‌ی وسایل الکترونیکی غیر مجاز است و تقلب محسوب می‌شود.

توجه: امتحان از ۱۱۰ نمره است و برای کامل شدن باید ۱۰۰ نمره کسب شود.

سوال ۱ سوالات پاسخ کوتاه (۱۸ نمره)

در هر یک از موارد زیر درست یا غلط بودن آن را مشخص کنید و به صورت مختصر علت را توضیح دهید. (هر مورد ۳ نمره)

الف) اضافه کردن ترم منظم‌ساز[۱] به تابع هزینه، باعث افزایش خطای بایاس و کاهش خطای واریانس می‌شود.

ب) در روش کاهش گرادیان[۲] اگر اندازه‌ی پارامتر یادگیری[۳] به اندازه کافی کوچک باشد، همواره به global minimum می‌رسیم.

ج) در الگوریتم 1-NN اگر از ساختار داده‌ی KD-tree استفاده شود، هزینه یافتن نزدیک‌ترین همسایه $O(1)$ است.

د) شبکه‌های عصبی عمیق با به اشتراک گذاشتن وزن‌ها تعداد پارامترها را کم می‌کنند

ه) شبکه‌های عصبی RBF (Radial basis function network) قادر هستند هر تابعی را پیاده سازی کنند.

و) عموما شبکه های عصبی با خطای بایاس (Bias) مواجه هستند.

سوال ۲ طبقه‌بند بهینه بیز (۲۰ نمره)

در یک مسأله طبقه‌بندی دو کلاسه، توزیع احتمال دو کلاس را به صورت زیر در نظر بگیرید:

$$P(x|y=1) = \mathcal{N}\left(\boldsymbol{x}; \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\frac{1}{9} & 0\\ 0 & \frac{1}{8}\end{bmatrix}\right)$$

$$P(x|y=2) = \mathcal{N}\left(\boldsymbol{x}; \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\frac{1}{2} & 0\\ 0 & 1\end{bmatrix}\right)$$

که در آن $\boldsymbol{x} = \begin{bmatrix}x_1\\x_2\end{bmatrix}$ ورودی و y نشان دهنده‌ی کلاس است. احتمال پیشین دو کلاس را برابر فرض کنید.

در طبقه‌بند بهینه بیز، مرز جدا کننده دو کلاس را در صفحه دو بعدی به دست آورید و توضیح دهید که این مرز معادل چه شکل هندسی است (خط، دایره، بیضی، سهمی یا هذلولی).

راهنمایی: معکوس یک ماتریس قطری، با معکوس کردن درایه‌های قطر اصلی بدست می‌آید.

---

[۱] Regularization term

[۲] Gradient descent

[۳] Learning rate

سوال ۳ تخمین پارامتر (۲۵ نمره)

توزیع احتمال Pareto در اقتصاد کاربرد زیادی دارد. رابطه‌ی این توزیع به صورت زیر است:

$$P(x) = \frac{\theta\, b^{\theta}}{x^{\theta+1}}, \qquad x \geq b, \theta > 1$$

که در آن θ و b پارامترهای مدل هستند. فرض کنید نمونه‌های $D = \{x_1, \dots, x_n\}$ به صورت i.i.d. از توزیع احتمال Pareto آمده باشند.

الف) (۸ نمره) تخمین‌گر بیشینه‌ی درستنمایی[۴] (MLE) را برای پارامتر θ بدست آورید.

ب) (۸ نمره) توزیع احتمال پیشین[۵] زیر را برای پارامتر θ در نظر بگیرید:

$$p(\theta) = Gamma(\theta \mid \alpha, \beta) = c\, \theta^{\alpha-1}\, e^{-\beta\theta}$$

که در رابطه‌ی بالا، c یک ضریب ثابت است و α و β پارامترهای توزیع گاما هستند. توزیع احتمال پسین[۶] را برای پارامتر θ بدست آورید (نیازی به محاسبه ضریب ثابت توزیع پسین نیست):

$$p(\theta|D) = ?$$

راهنمایی: $a^b = e^{b \ln a}$

ج) (۳ نمره) آیا توزیع احتمال پیشین فوق، یک conjugate prior برای پارامتر θ است؟ توضیح دهید.

د) (۳ نمره) با استفاده از توزیع احتمال پیشین فوق، تخمین‌گر MAP برای پارامتر θ چیست؟ (**راهنمایی**: مقدار بیشینه توزیع گاما در نقطه $\theta = \frac{a-1}{\beta}$ رخ می‌دهد).

ه) (۳ نمره) آیا اگر $n \to +\infty$ آنگاه، تخمین‌گر MAP به تخمین‌گر ML میل می‌کند؟ چرا؟

سوال ۴ رگرسیون خطی (۱۶ نمره)

در یک مسأله رگرسیون خطی، مجموعه داده‌ی $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ را در اختیار داریم. رابطه‌ی احتمالاتی میان $x \in \mathbb{R}^d$ و y را به صورت زیر در نظر می‌گیریم:

$$y_i = w^T x_i + \epsilon_i$$

$$\epsilon_i = laplace(\epsilon_i|\mu, 1) = \frac{1}{2}\exp(-|\epsilon_i - \mu|)$$

که در آن $w \in \mathbb{R}^d$ پارامتر مدل و ϵ_i یک نویز لاپلاس با میانگین μ و واریانس ۲ است. مقدار میانگین نویز را صفر در نظر بگیرید: $\mu = 0$

الف) (۸ نمره) با فرض i.i.d. بودن داده‌ها، تابع log-likelihood را تشکیل دهید و نشان دهید که بیشینه کردن تابع log-likelihood روی پارامتر w معادل است با کمینه کردن مجموع قدرمطلق خطا. به عبارت دیگر نشان دهید که:

$$\arg\max_w \log P(D|w) = \arg\min_w \sum_{i=1}^{n} |y_i - w^T x_i|$$

---

[۴] Maximum Likelihood Estimation

[۵] Prior distribution

[۶] Posterior Distribution

ب) (۸ نمره) حال اگر بخواهیم از دیدگاه بیز به این مسأله نگاه کنیم، باید پارامتر w را یک متغیر تصادفی در نظر بگیریم و برای آن یک توزیع احتمال پیشین داشته باشیم. حال فرض کنید توزیع احتمال پیشین لاپلاس را برای w داشته باشیم:

$$P(w) = \frac{1}{(2b)^d} \exp\left(-\frac{\|w\|_1}{b}\right)$$

که در رابطه‌ی بالا، $\|w\|_1$ نرم ۱ بردار w است و d بُعد w است. نشان دهید که تخمین‌گر MAP برای w معادل است با کمینه کردن مجموع قدر مطلق خطا به اضافه‌ی یک ترم منظم‌ساز نرم ۱:

$$\arg\max_w \log P(w|D) = \arg\min_w \sum_{i=1}^{n} |y_i - w^T x_i| + \lambda \|w\|_1$$

رابطه‌ی بین λ و b چیست؟

سوال ۵ درخت تصمیم (۱۴ نمره)

داده‌های زیر نشان دهنده‌ی پاس شدن یا نشدن در درس یادگیری ماشین بر اساس معدل و میزان مطالعه برای امتحان است. برای معدل (GPA) سه حالت High، Medium و Low در نظر گرفته شده است.

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

لگاریتم‌ها را در مبنای ۲ محاسبه کنید. همچنین داریم: $\log_2 3 \simeq 1.6$ (استفاده از ماشین حساب مجاز نیست).

الف) (۳ نمره) آنتروپی $H(passed)$ چقدر است؟

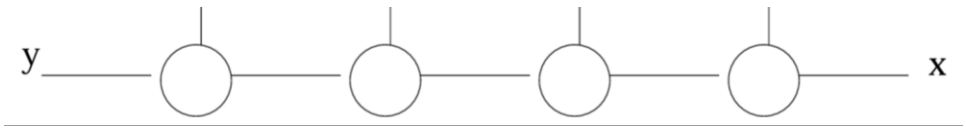
ب) (۴ نمره) آنتروپی $H(passed|GPA)$ چقدر است؟

ج) (۴ نمره) آنتروپی $H(passed|Studied)$ چقدر است؟

د) (۳ نمره) بر اساس مقادیر بدست آمده در قسمت‌های قبل، درخت تصمیم را مطابق الگوریتم ID3 ترسیم کنید. (ترسیم درخت به تنهایی کفایت می‌کند و نیازی به نوشتن محاسبات نیست).

سوال ۶ شبکه‌های عصبی (۱۷ نمره)

فرض کنید که به دنبال آموزش شبکه عصبی زیر به کمک back-propagation هستیم، تابع فعالسازی زیگموئید[۷] به کار گرفته شده است و مقدار اولیه تمام وزن‌ها برابر با 1 و تمام بایاس‌ها برابر با -0.5 است. ورودی $x = 0.5$ را به شبکه می‌دهیم، خروجی تمام نورون‌ها 0.5 خواهد شد. داده ورودی $x = 0.5$ برچسب $y = 1$ را دارد. در این شبکه میزان گرادیان در به روز شدن وزن‌ها را به دست آورده و در مورد سرعت تغییر وزن‌ها بحث کنید.



موفق و پیروز باشید.

[۷] Sigmoid activation function