دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

پاییز ۱۴۰۰

استاد: محمدحسین رهبان

گردآورندگان: امیرحسین باقری، امیرپویا معینی، علیرضا تاجمیرریاحی

تمرین چهارم — **رگرسیون، درخت تصمیم‌گیری** — مهلت ارسال: ۲ و ۱۴ دی

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۷ روز و در مجموع ۲۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد.
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

سوالات نظری (۱۰۰ نمره)

۱. (۲۵ نمره) مساله‌ی رگرسیون خطی ساده را درنظر بگیرید. در تعریف احتمالاتی این مساله فرض می‌کنیم رابطه‌ی زیر بین x_i و y_i وجود دارد به طوری که $\epsilon_i \sim \mathcal{N}(۰, \sigma^۲)$ توزیع می‌شود.

$$y_i = \beta_۰ + \beta_۱ x_i + \epsilon_i$$

می‌دانیم که $\beta_۰$، $\beta_۱$ و σ مقادیر ثابت نامنفی هستند.

الف) اثبات کنید که تخمین بیشینه درست نمایی دو پارامتر $\beta_۰$ و $\beta_۱$ برابر با کمینه کردن مجموع مربعات خطا است.

ب) اثبات کنید که تخمین‌های به دست آمده در قسمت قبل که براساس بیشینه درست نمایی بودند نااریب (unbiased) هستند و از توزیع‌های زیر پیروی می‌کنند.

$$\hat{\beta}_۱ \sim \mathbf{N}\left(\beta_۱, \frac{\sigma^۲}{\sum_j (x_j - \bar{x})^۲}\right) \quad , \quad \hat{\beta}_۰ \sim \mathbf{N}\left(\beta_۰, \frac{\sigma^۲ \sum_j x_j^Y}{n \sum (x_j - \bar{x})^۲}\right)$$

(راهنمایی: منظور از تخمینگر نااریب، تخمینگری است که امید ریاضی آن با مقدار واقعی متغیر موردنظر برابر باشد.)

پ) حال خانواده‌ای خطی از تخمینگرهای خطی برای تخمین پارامتر $\beta_۱$ مطابق زیر درنظر بگیرید.

$$\tilde{\beta}_۱ = \frac{\sum \gamma_i y_i}{\sum \gamma_i x_i} \quad \text{such that} \quad \sum_i \gamma_i = ۰$$

محاسبه کنید که آیا تخمینگر بیشینه درست نمایی عضوی از این خانواده است یا نه؟ و اگر هست رابطه‌ی γ_i به چه صورت است؟

ت) اثبات کنید هرتخمینی عضو این خانواده یک تخمینگر نااریب است.

ث) سپس درنهایت اثبات کنید که $\mathrm{Var}\left(\hat{\beta}_۱\right) \leqslant \mathrm{Var}\left(\tilde{\beta}_۱\right)$ برقرار است. سپس نتیجه به دست آمده را توضیح دهید.

۲. (۲۰ نمره) فرض کنید قصد داشته باشیم مساله‌ی رگرسیون چند متغیره را درنظر بگیریم. تابع هزینه‌ای که باید کمینه شود به فرم زیر خواهد بود.

$$\min_{W} F(W) = \lambda W^T W + \|XW - Y\|_۲^۲$$

الف) اگر بخواهیم این مساله را با الگوریتم Stochastic Gradient Descent حل کنیم، شبه کد آن را بنویسید.

ب) حال فرض کنید تعریف کنیم

$$W_۱ = \underset{W}{\operatorname{argmin}}\, L(W)$$

$$W_۲ = \underset{W}{\operatorname{argmin}}\, L(W) + \lambda W^T W$$

که $L(W)$ یک تابع نامنفی است. اثبات کنید که $\|W_۲\|_۲ \leq \|W_۱\|_۲$ و ارتباط آن را با فرمول بندی مساله بیان کنید.

۳. (۱۰ نمره) یک دیتاست چند متغیره را درنظر بگیرید، به این معنی که $x_i \in \mathbb{R}^p$ که $p > ۱$ است و $y_i \in \mathbb{R}$ است. فرض کنید مشاهده کرده‌ایم که یکی از ضرایب محاسبه شده یک مقدار خیلی بزرگ منفی نسبت به باقی متغیرها پیدا کرده است کدام یک از گزاره‌های زیر صحیح است؟ توضیح دهید.

- این ویژگی تاثیر زیادی روی مدل دارد و باید حفظ شود.
- این ویژگی تاثیر زیادی روی مدل ندارد و باید ایگنور شود.
- نمی‌توان بدون در دست داشتن اطلاعات بیشتر درمورد این ویژگی نظر داد.

۴. (۱۵ نمره) با ارائه دلیل صحیح یا غلط بودن هر یک از گزاره‌های زیر را ثابت کنید.

- اگر bias زیاد است اضافه کردن تعداد داده‌های آموزش کمک زیادی به کم کردن بایاس نمی‌کند.
- کم کردن خطای مدل روی داده‌های آموزش منجر به کاهش خطای مدل روی داده‌های تست می‌شود.
- افزایش پیچیدگی مدل رگرسیون همواره منجر به کاهش خطای مدل روی داده‌ی آموزش و افزایش خطای مدل روی داده‌ی تست می‌شود.

۵. (۲۰ نمره) بر روی ۶ بیمار قلبی، مطالعاتی صورت گرفته است و جدول زیر از نتایج آن به دست آمده است.

HEART ATTACK	EXERCISES	SMOKES	MALE	CHEST PAIN	PATIENT ID
yes	yes	no	yes	yes	۱
yes	no	yes	yes	yes	۲
yes	no	yes	no	no	۳
no	yes	no	yes	no	۴
yes	yes	yes	no	yes	۵
no	yes	yes	yes	no	۶

الف) با استفاده از این داده‌ها درخت تصمیم گیری پیش بینی حمله قلبی را تشکیل دهید.

ب) درخت به دست آمده را به صورت تعدادی گزاره‌ی تصمیم گیری ترجمه کنید.

۶. (۱۰ نمره) نشان دهید هر دسته بند دودویی به فرم $h : \{۰, ۱\}^d \mapsto \{۰, ۱\}$ می‌تواند به صورت یک درخت تصمیم گیری به عمق حداکثر $d + ۱$ با گره‌های به فرم $(x_i = ۰?)$ برای یک $i \in \{۱, \ldots, d\}$ پیاده سازی شود.

سوالات عملی (۱۰۰+۲۰ نمره)

۱. (۵۰+۱۰ نمره) در این سوال ابتدا با کمک مفهوم سود اطلاعات[۱] نحوه‌ی انتخاب ویژگی‌ها برای تشکیل یک درخت را پیاده‌سازی خواهید کرد؛ و سپس یک مدل درخت تصمیم با کمک ابزارهای موجود آموزش خواهید داد و بهترین پارامترها را برای آن‌ها انتخاب خواهید کرد. نوت‌بوک مربوط به این سوال (PQ1.ipynb) در اختیارتان قرار گرفته‌است.

۲. (۵۰+۱۰ نمره) در این سوال شما یک بار فرم بسته رگرسیون خطی را پیاده سازی می‌کنید و خواسته های مربوطه را در ژوپیتر فایل برای این بخش پیاده می‌کنید. سپس فرم gradient descent را با استفاده از بهترین نرخ یادگیری پیاده سازی کرده و خواسته های مربوط به این بخش را نیز انجام می‌دهید. نوت‌بوک مربوط به این بخش (PQ2.ipynb) در اختیارتان قرار گرفته‌است.

[۱] Information Gain