هوش مصنوعی

پاییز ۱۴۰۰
استاد: محمدحسین رهبان

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

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 $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ توزیع می‌شود.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

می‌دانیم که β_0، β_1 و σ مقادیر ثابت نامنفی هستند.

الف) اثبات کنید که تخمین بیشینه درست نمایی دو پارامتر β_0 و β_1 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.

$$\hat{\beta}_1 \sim \mathbf{N}\left(\beta_1, \frac{\sigma^2}{\sum_j (x_j - \bar{x})^2}\right) \quad , \quad \hat{\beta}_0 \sim \mathbf{N}\left(\beta_0, \frac{\sigma^2 \sum_j x_j^2}{n \sum (x_j - \bar{x})^2}\right)$$

(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 β_1 مطابق زیر درنظر بگیرید.

$$\tilde{\beta}_1 = \frac{\sum \gamma_i y_i}{\sum \gamma_i x_i} \quad \text{such that} \quad \sum_i \gamma_i = 0$$

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 $\mathrm{Var}\left(\hat{\beta}_1\right) \leqslant \mathrm{Var}\left(\tilde{\beta}_1\right)$ برقرار است. سپس نتیجه به دست آمده را توضیح دهید.

حل.

الف

$$f_L = f(y_1, ..., y_n|\beta_0, \beta_1) = \prod_{i=1}^{n} p(y_i|\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2)$$

$$\rightarrow f_L = (2\pi)^{\frac{-n}{2}} \sigma^{-n} \exp(-\frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2)$$

$$\mathbb{L} = \log(f_L) = c - \frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2)$$

$$\frac{\partial \mathbb{L}}{\partial \beta_0} \propto \sum (y_i - \beta_0 - \beta_1 x_i) = 0 \rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\frac{\partial \mathbb{L}}{\partial \beta_1} \propto \sum (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \rightarrow \beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

ب

رابطه‌ی را برای $\hat{\beta}_1$ اثبات می‌کنیم و باقی روند برای $\hat{\beta}_0$ نیز به همین صورت خواهد بود.

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2} = \sum_i \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} y_i$$

از آنجا که y_i خود یک توزیع گاوسی دارد، $\hat{\beta}_1$ نیز جمع تعدادی جمله‌ی گاوسی می‌شود پس توزیع آن نیز گاوسی می‌شود. پس داریم:

$$E[\hat{\beta}_1] = \sum_i \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i) = \beta_1$$

$$Var[\hat{\beta}_1] = \sum_i (\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2})^2 Var(y_i) = \frac{\sigma^2}{\sum (x_i - \bar{x})}$$

پ

مشخص است که با جایگذاری $\gamma_i = x_i - \bar{x}$ به همان تخمین کمینه‌ی مربعات خطا می‌رسیم.

ت

$$E[\tilde{\beta}_1] = \sum_i \frac{\gamma_i}{\sum \gamma_i x_i} E[y_i] = \sum_i \frac{\gamma_i}{\sum \gamma_i x_i} (\beta_0 + \beta_1 x_i) = 0 + \frac{\sum \gamma_i x_i}{\sum \gamma_i x_i} \beta_1 = \beta_1$$

ث

$$\frac{\gamma_i}{\sum \gamma_i x_i} := d_i$$

$$\rightarrow Var[\tilde{\beta}_۱] = \sigma^۲ \sum d_i^۲$$

$$\sigma^۲ \sum d_i^۲ \sum (x_i - \bar{x})^۲ >= \sigma^۲ (\sum d_i (x_i - \bar{x}))^۲ = \sigma^۲$$

$$\longrightarrow \sigma^۲ \sum d_i^۲ = Var[\tilde{\beta}_۱] \geq \frac{\sigma^۲}{\sum (x_i - \bar{x})^۲} = Var(\hat{\beta}_۱)$$

به نوعی انگار تخمین بیشینه درست نمایی این ضریب، بهترین تخمین است. زیرا هم نااریب است و هم در خانواده‌ای این تخمین‌ها کمترین واریانس را دارد.

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

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

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

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.

حل.

الف

- initialize W
- while not converged do
- for i in $1, \ldots, n$ do
- $W \leftarrow W - \alpha \times 2 \left(\lambda W + \left(W^T x - y\right) x\right)$
- return W

ب

$$w_2 = \arg\min_w L(w) + \lambda \|w\|^2 \Rightarrow L(w_2) + \lambda \|w_2\|^2 \leqslant L(w_1) + \lambda \|w_1\|^2$$
$$w_1 = \arg\min_w L(w) \Rightarrow L(w_1) \leqslant L(w_2)$$

از دو رابطه بالا می‌توان نتیجه گرفت

$$L(w_1) + \lambda \|w_2\|^2 \leqslant L(w_2) + \lambda \|w_1\|^2 \Rightarrow \lambda \|w_2\|^2 \leqslant \lambda \|w_1\|^2$$
$$\|w_2\|^2 \leqslant \|w_1\|^2 \Rightarrow \|w_2\| \leqslant \|w_1\|$$

۳. (۱۰ نمره) یک دیتاست چند متغیره را درنظر بگیرید، به این معنی که $x_i \in \mathbb{R}^p$ که $p > 1$ 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.

حل.
تنها مورد سوم درست است. هم ممکن است بخاطر overfitting ضریب زیاد شده باشد هم ممکن است که واقعا همبستگی زیادی وجود داشته باشد.

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.

حل.

- درست: زیرا وقتی بایاس زیاد است یعنی مدل نتوانسته ویژگی‌های داده را پیدا کند و افزایش داده فقط واریانس را کمتر می‌کند.
- نادرست: زیرا ممکن است overfit کنیم.
- نادرست: افزایش پیچیدگی در برخی مسائل ممکن است منجر به مدل بهتری شود.

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

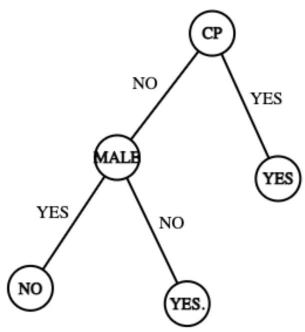
HEART ATTACK	EXERCISES	SMOKES	MALE	CHEST PAIN	PATIENT ID
yes	yes	no	yes	yes	۱
yes	no	yes	yes	yes	۲
yes	no	yes	no	no	۳
no	yes	no	yes	no	۴
yes	yes	yes	no	yes	۵
no	yes	yes	yes	no	۶

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.

حل.

الف



ب

- IF Chest Pain=Yes THEN Heart Attack = Yes
- IF Chest Pain=No AND Male=Yes THEN Heart Attack = No
- IF Chest Pain=No AND Male=No THEN Heart Attack = Yes

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 $i \in \{۱, \dots, d\}$ پیاده سازی شود.

حل. یک درخت دودویی کامل درنظر بگیرید که ریشه آن $(x_۱ = ۰?)$ باشد و گره‌های در سطح i به صورت $(x_{i+۱} = ۰?)$ باشند. این درخت ۲^d گره دارد و مسیر از ریشه به هر برگی به صورت گره‌های

$(x_1 = 0?) , (x_2 = 0?) , \ldots , (x_d = 0)$ است. اگر مثلا جواب yes به سوالات را جهت راست درنظر بگیریم، به سادگی مشخص است که این درخت معادل دسته بند دودویی گفته شده است.