دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

بهار ۱۴۰۲

استاد: محمدحسین رهبان

گردآورندگان: محمدرضا دویران، امیررضا میرزایی و محمد جواد هزاره

پاسخ تمرین پنجم — آشنایی با یادگیری ماشین، رگرسیون، درخت تصمیم‌گیری — مهلت ارسال: ۱۹ خرداد

- 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 ۱۵ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۲ 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 (۱۲۰ نمره)

۱. (۲۰ نمره) با توجه به مفاهیم فراگرفته شده در درس به سوالات زیر پاسخ دهید.

الف) برای مدل‌هایی که اریبی زیادی دارند دست کم دو راهکار ارائه دهید که مقدار این اریبی کاهش یابد.

ب) فرض کنید تعدادی از ویژگی‌های مدل دوبه‌دو با یکدیگر هم‌بسته باشند. این اتفاق را از دیدگاه bias-variance بررسی کنید. اگر ویژگی‌های هم‌بسته را حذف کنیم بیان کنید که چگونه bias و variance تغییر می‌کنند.

پ)کدام یک از گزاره‌های زیر درست می‌باشند؟ چرا؟

- اگر مقدار بایاس زیاد باشد، افزایش تعداد داده‌های آموزش می‌تواند باعث کاهش بایاس شود.
- افزایش پیچیدگی مدل در رگرسیون همواره باعث کاهش خطای آموزش و افزایش خطای تست می‌شود.

حل.

الف) برای حل این مشکل می‌توان از این راهکارها استفاده کرد: استفاده از تعداد بیشتری ویژگی - تنظیم پارامترها برای پیچیده‌تر شدن مدل - تنظیم داده‌ها - در نظر گرفتن مدل پیچیده‌تر

ب) یچرهای هم‌بسته به صورت کلی باعث افزایش واریانس می‌شوند و این بدین معناست که مدل ما از قدرت تعمیم‌سازی کمتری برخوردار است. این بعنی بایاس نیز کمتر است. با حذف ویژگی‌های هم‌بسته واریانس کم می‌شود و مقدار بایاس رشد می‌کند.

پ) گزاره اول درست است. زیرا افزایش سایز داده‌های آموزش می‌تواند باعث شود حالات بیشتری را مدل ببیند و در نتیجه با بایاس کمتری روبه‌رو شویم. البته گاهی اوقات تنها افزایش سایز داده‌های آموزش کافی نیست و محدودیت‌های مدل باید به شکل دیگری برطرف شوند. برای مثال می‌توانیم مدل را پیچیده‌تر کنیم.

گزاره دوم اشتباه است زیرا همواره عبارت داده‌شده برقرار نیست و افزایش پیچیدگی گاهی ممکن است باعث overfit شدن مدل شود و خطای تست را افزایش دهد اما رابطه بین این دو خطی نیست و گاهی افزایش پیچیدگی باعث می‌شود مدل بهتری داشته باشیم و باعث کاهش خطای تست شود.

۲. (۳۰ نمره)

(آ) فرض کنید برای داده‌های جدول ۱ یک درخت تصمیم آموزش می‌دهیم تا X را به وسیله‌ی A, B, C پیش‌بینی کنیم. درصد خطای مدل پس از آموزش بر روی داده‌های آموزش چقدر خواهد بود؟

C	B	A	X
۰	۰	۰	۰
۱	۰	۰	۰
۱	۰	۰	۰
۰	۱	۰	۰
۱	۱	۰	۰
۱	۱	۰	۱
۱	۱	۰	۱
۰	۰	۱	۰
۱	۰	۱	۱
۰	۱	۱	۱
۰	۱	۱	۱
۱	۱	۱	۰
۱	۱	۱	۱

جدول ۱: داده‌های مدل درخت تصمیم

(ب) فرض کنید روی مجموعه‌ی داده‌ی دلخواهی، درخت تصمیمی برای دسته‌بندی بین k کلاس، آموزش می‌دهیم. حداکثر خطایی که ممکن است این مدل روی داده‌های آموزش داشته باشد چقدر خواهد بود؟ (پاسخ را به صورت کسری بنویسید)

حل.

(آ) در درخت تصمیم، در دادگان آموزش خطا تنها زمانی اتفاق می‌افتد که چند داده با فیچرهای یکسان وجود داشته باشد اما برچسب آن‌ها متفاوت باشد.
در این جا سه داده‌ی $(A = ۰, B = ۱, C = ۱)$ داریم که برچسب دو تا از آن‌ها ۱ و دیگری صفر داده شده است. پس درخت مقدار ۱ را پیش‌بینی می‌کند و یک خطا به ازای این داده اتفاق می‌افتد.
همچنین دو داده‌ی $(A = ۱, B = ۱, C = ۱)$ داریم که برچسب یکی صفر و دیگری یک است و درخت به صورت رندم یکی از این برچسب‌ها را برای پیش‌بینی انتخاب می‌کند که باعث می‌شود یک خطا هم در اینجا داشته باشیم.
پس درخت در پیش‌بینی دو تا از داده‌ها خطا دارد که درصد خطای آن برابر خواهد بود با: $۲/۱۳ \times ۱۰۰ \approx ۱۵/۴\%$

(ب) بدترین حالت زمانی اتفاق می‌افتد که به ازای هر datapoint که داریم، برچسب همه‌ی دسته‌ها وجود داشته باشد که باعث می‌شود به ازای‌ی هر k داده، تنها یک پیش‌بینی درست داشته باشیم و $k - ۱$ پیش‌بینی غلط. که باعث خطای کلی $\frac{k-۱}{k}$ بر روی داده‌های آموزش می‌شود.

۳. (۴۰ نمره) در رابطه با الگوریتم Logistic regression به سوالات زیر پاسخ دهید.

الف) این الگوریتم را برای حالت K کلاسه تغییر دهید و احتمالات آن را بنویسید.

ب) همانطور که در قسمت الف به دست آوردید، در Logistic regression برای K کلاس، احتمال پسین به روش زیر محاسبه می‌شود:

$$P(Y=k|X=x)=\frac{e^{w_k^Tx}}{1+\Sigma_{i=1}^{K-1}e^{w_i^Tx}}, k=1,2,3,4,...,K-1$$
$$P(Y=K|X=x)=\frac{1}{1+\Sigma_{i=1}^{K-1}e^{w_i^Tx}}$$

برای راحتی فرض کردیم که $w_k=0$ می‌باشد. کدام یک از پارامترها باید تخمین زده شوند؟

پ) حال log-likelihood زیر را برای n نمونه‌ی زیر ساده کنید:

$$Samples:(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$$
$$L(w_1,...,w_{k-1})=\Sigma_{i=1}^{n}lnP(Y=y_i|X=x_i)$$

ت) گرادیان L را نسبت به هریک از w_k ها بیابید و آن را ساده کنید.

ث) تابع هدف زیر را در نظر بگیرید. گرادیان f را با توجه به هریک از w_k ها بیابید.

$$f(w_1,...,w_{k-1})=L(w_1,...,w_{k-1})-\frac{\lambda}{2}\Sigma_{j=1}^{K-1}||w_j||_2^2$$

حل. الف) به جای اینکه $p(C_1|x)$ را تخمین بزنیم کافی است که هر $p(C_i|x)$ را به صورت زیر در نظر بگیریم:

$$p(C_i|x)=y_i(x)=\frac{exp(-w_i^Tx)}{\sum_{j=1}^{K}exp(-w_j^Tx)}$$

حال T را ماتریسی در نظر می‌گیریم که هر سطر آن بردار one-hot انکود شده‌ی $t^{(i)}$ است و هر دیتا پوینت را به صورت $x^{(i)},t^{(i)}$ در نظر می‌گیریم. تابع loss برای این مدل به صورت زیر خواهد بود:

$$p(t^{(i)}|x^{(i)},w)=y_{t^{(i)}}(x^{(i)})$$
$$L(W)=-\sum_{i=1}^{N}(\sum_{k=1}^{N}T_{nk}log(y_k(x^{(n)})))$$

ب) نیاز است $K-1$ پارامتر شامل $w_1,w_2,\cdots,w_{k-1}$ ها تخمین زده شوند.

پ) ابتدا $Y^{(i)}$ را به عنوان برچسب بردار one-hot برای $x^{(i)}$ در نظر می‌گیریم. حال $o^{(i)}$ به صورت زیر خواهد بود:

$$o^{(i)}=\begin{bmatrix} w_1^Tx^{(i)} \\ w_2^Tx^{(i)} \\ \vdots \\ w_{k-1}^Tx^{(i)} \\ 0 \end{bmatrix} \tag{1}$$

حال اگر بدانیم که $x^{(i)} \in C_m$ برقرار باشد؛ آنگاه خواهیم داشت:

$$\begin{aligned}\log P(y^{(i)}|x^{(i)}) &= \log \frac{e^{o_k^{(i)}}}{\sum_j e^{o_j^{(i)}}} \\ &= o_k^{(i)} - \log(\sum_j e^{o_j^{(i)}}) \\ &= (Y^{(i)})^T o^{(i)} - \log(\sum_j e^{o_j^{(i)}})\end{aligned} \tag{۲}$$

حال برای L خواهیم داشت:

$$L(w_۱, \cdots, w_{k-۱}) = \sum_{i=۱}^{n} \left[(Y^{(i)})^T o^{(i)} - \log(\sum_j e^{o_j^{(i)}}) \right] \tag{۳}$$

ت) به صورت زیر گرادیان L را نسبت به تک تک w_k ها به دست می‌آوریم:

$$\begin{aligned}\frac{\partial L}{\partial w_{ml}} &= \sum_{i=۱}^{n} \left[Y_m^{(i)} x_l^{(i)} - \frac{e^{o_m^{(i)}}}{\sum_j e^{o_j^{(i)}}} x_l^{(i)} \right] \\ &= \sum_{i=۱}^{n} \left[Y_m^{(i)} - P_m^{(i)} \right] x_l^{(i)}\end{aligned} \tag{۴}$$

حال یا $P_m^{(i)} = P(Y = m|X = x^{(i)})$ برقرار است یا $x^{(i)}$ عضو کلاس m است:

$$\frac{\partial L}{\partial w_j} = \sum_{i=۱}^{n} \left(Y_j^{(i)} - P_j^{(i)} \right) x^{(i)} \implies \frac{\partial L}{\partial W} = (Y - P)^T X \tag{۵}$$

در رابطه بالا روابط زیر برقرارند:

$$\begin{aligned}X &= [x^{(۱)}, x^{(۲)}, \cdots, x^{(n)}]^T \\ Y &= [Y^{(۱)}, \cdots, Y^{(n)}]^T \\ P &= [P^{(۱)}, \cdots, P^{(n)}]^T \\ W &= [w_۱, \cdots, w_{k-۱}, ۰]^T\end{aligned}$$

ث) گردایان مورد نظر به صورت زیر خواهد بود:

$$\frac{\partial f}{\partial W} = (Y - P)^T X - \lambda W \tag{۶}$$

۴. (۳۰ نمره) فرض کنید n داده آموزش با m ویژگی داریم که که ماتریس این داده‌ها را $X_{n\times m}$ در نظر می‌گیریم. بردار مقدار هدف نیز برابر $y = [y^{(1)}, ..., y^{(n)}]$ می‌باشد. در ادامه منظور از x_j، j امین ستون ماتریس X است. حال با توجه به توضیحات داده شده به سوالات زیر پاسخ دهید.

الف) ابتدا ثابت کنید اگر رگرسیون را فقط بر روی یکی از m ویژگی موجود آموزش دهیم آنگاه خواهیم داشت:

$w_j = \frac{x_j^T y}{x_j^T x_j}$

ب) فرض کنید ستون‌های ماتریس X متعامد باشد. ثابت کنید که پارامترهای بهینه از آموزش رگرسیون بر روی همه ویژگی‌ها با پارامترهای بهینه حاصل از آموزش روی هر ویژگی به طور مستقل یکسان است.

پ) فرض کنید می‌خواهیم یک رگرسیون بر روی بایاس و یکی از ویژگی‌های نمونه داده‌ها آموزش دهیم. ($w = [w_j, w.]$) با توجه به اطلاعات داده شده عبارات زیر را اثبات کنید:

$w_j = \frac{cov[x_j, y]}{var[x_j]}$

$w. = E[y] - w_j E[x_j]$

حل. الف) در این حالت ماتریس داده ما برابر x_j است و با توجه به رابطه رگرسیون خطی رابطه زیر برقرار است:

$$w_j = (x_j^T x_j)^{-1} x_j y = \frac{x_j^T y}{x_j^T x_j}$$

ب) می‌دانیم که ستون‌های ماتریس X متعامد هستند. پس ضرب داخلی آن‌ها صفر است. به همین دلیل $X^T X$ قطری خواهد بود:

$$X^T X = diag(x_1^T x_1, ..., x_m^T x_m) \implies (X^T X)^{-1} = diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1})$$

حال خواهیم داشت:

$$w = (X^T X)^{-1} X^T y = diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1}) X^T y$$

$$\implies w_j = (diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1}) X^T y)_j = (diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1}))_j (X^T y)_j$$

$$\implies w_j = (x_j^T x_j)^{-1} (X^T y)_j = \frac{(X^T y)_j}{x_j^T x_j} = \frac{x_j^T y}{x_j^T x_j}$$

حال با توجه به قسمت قبل نتیجه می‌گیریم که پارامترهای بهینه از آموزش رگرسیون بر روی همه ویژگی‌ها با پارامترهای بهینه حاصل از آموزش بر روی هر ویژگی به طور مستقل یکسان است.

پ) در مد ماتریس داده‌ی ما به صورت زیر می‌باشد:

$$X = \begin{bmatrix} & 1 \\ & 1 \\ x_j & . \\ & . \\ & . \\ & 1 \end{bmatrix}$$

همانطور که می‌دانیم x_j یک ستون می‌باشد و ضرب داخلی آن با ستونی از یک برابر مجموع اعضای x_j است که آن را با $sum(x_j)$ نشان می‌دهیم. حال خواهیم داشت:

$$X^T X = \begin{bmatrix} x_j^T x_j & sum(x_j) \\ sum(x_j) & n \end{bmatrix} \implies (X^T X)^{-1} = \frac{1}{n||x_j||^2 - sum(x_j)^2} \begin{bmatrix} n & -sum(x_j) \\ -sum(x_j) & x_j^T x_j \end{bmatrix}$$

همچنین در نظر داشته باشید که داریم:

$$X^T y = \begin{bmatrix} x_j^T y \\ sum(y) \end{bmatrix}$$

حال دو عبارت خواسته شده را اثبات می‌کنیم:

$$[w_j, w_0] = w = (X^T X)^{-1} X^T y = \frac{1}{n||x_j||^2 - sum(x_j)^2} \begin{bmatrix} n & -sum(x_j) \\ -sum(x_j) & x_j^T x_j \end{bmatrix} \begin{bmatrix} x_j^T y \\ sum(y) \end{bmatrix}$$

حال صورت و مخرج را در n^2 تقسیم می‌کنیم:

$$w_j = \frac{E[x_j, y] - E[x_j]E[y]}{E[x_j^2] - E[x_j]^2} = \frac{cov(x_j, y)}{var(x_j)}$$

حال بخش دوم را اثبات می‌کنیم:

$$w_0 = \frac{sum(y)||x_j||^2 - sum(x_j)sum(x_j^T y)}{n||x_j||^2 - sum(x_j)^2} = E[y] + \frac{E[y]E[x_j^2] - E[x_j]E[x_j y]}{var(x_j)}$$

$$\implies w_0 = E[y] + E[x_j]\frac{E[y]E[x_j] - E[x_j y]}{var(x_j)} = E[y] - w_j E[x_j]$$