دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر
طراحان: سپهرغیاث، امیرکوشان فتاح، شایان بقایی‌نژاد، آرش ضیائی، رادین چراغی، علیرضا ملک حسینی

هوش مصنوعی

بهار۱۴۰۴
استاد: احسان تن قطاری

تمرین چهارم | یادگیری ماشین و شبکه عصبی | مهلت ارسال: ۴ خرداد

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین سقف ۴ روز و در مجموع ۱۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر ساعت تأخیر غیر مجاز نیم درصد از نمره ی تمرین کم 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 (۱۰۰ نمره)

۱. (۵ نمره) درستی یا نادرستی عبارت‌های زیر را با ذکر دلیل مشخص کنید.

(آ) افزایش پیچیدگی یک مدل رگرسیون، همواره منجر به کاهش خطای مدل در داده های آموزش و افزایش خطای مدل در برآورد دادهٔ تست می‌شود.

(ب) اگر bias مدل ما زیاد است، اضافه کردن تعداد داده های آموزش کمک زیادی به کم کردن bias نمی‌کند.

(ج) از آنجا که طبقه بندی حالت خاصی از رگرسیون است، رگرسیون لجیستیک نیز حالت خاصی از رگرسیون خطی می‌باشد.

(د) در صورتی که یک نمونه از داده‌های آموزشی توسط الگوریتم پرسپترون اشتباه کلاس بندی شود، پس از بروزرسانی وزن‌های مدل با استفاده از همان نمونه، الگوریتم آنرا به درستی کلاس‌بندی می‌کند.

حل.

(آ) غلط، چرا که اگر مدل در حالت underfitting باشد (یعنی بیش از حد ساده باشد و نتواند الگوی موجود در داده را یاد بگیرد)، افزایش پیچیدگی می‌تواند هم خطای آموزش و هم خطای تست را کاهش دهد، چون مدل بهتر با الگوی واقعی داده منطبق می‌شود. تنها زمانی که مدل به مرحله overfitting برسد، پیچیدگی بیشتر باعث افزایش خطای تست می‌شود.

(ب) غلط، در کل بایاس بالا معمولاً ناشی از ساده بودن مدل است (underfitting). در این حالت، حتی با اضافه کردن داده‌های بیشتر، مدل همچنان قادر به یادگیری الگوی پیچیده مسئله نیست. کاهش بایاس نیازمند افزایش ظرفیت مدل (مثلاً استفاده از مدل‌های غیرخطی‌تر یا پیچیده‌تر) است و افزایش داده احتمالا کمک خاصی به آن نمیکند. مگر اینکه تعداد داده ها قبل از این بیش از حد کم و پراکنده بوده و الگوی موجود در توزیع اصلی داده ها در سمپل اولیه یافت نمی‌شود (و به نحوی الگوی پیچیده تری ساخته شده). در این حالت ممکن است (به احتمال کمی) بایاس کم شود.

(ج) غلط، هرچند از نظر ساختار تابع، رگرسیون لجیستیک از یک ترکیب خطی از ویژگی‌ها استفاده می‌کند، اما خروجی آن احتمال (بین ۰ و ۱) است که از طریق تابع سیگموید حاصل می‌شود، نه یک مقدار عددی پیوسته مانند رگرسیون خطی. بنابراین، رگرسیون لجیستیک را نمی‌توان به طور کامل «حالت خاصی از رگرسیون خطی» دانست. بلکه می‌توان گفت از ترکیب خطی در مرحله تصمیم‌گیری بهره می‌برد ولی ماهیت آن متفاوت است.

(د) غلط، به‌روزرسانی وزن‌ها لزوماً باعث تصحیح فوری طبقه‌بندی نمونه نمی‌شود. ممکن است چندین تکرار لازم باشد.

یک مثال نقض میزنیم: فرض کنید وزن اولیه و داده به صورت زیر باشد

$$\mathbf{w} = \begin{bmatrix} ۵ \\ ۵ \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} ۱ \\ ۱ \end{bmatrix}, \quad y = -۱$$

ابتدا مدل یک پیش بینی اشتباه ارائه می‌دهد.

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x}) = \text{sign}(۵ \cdot ۱ + ۵ \cdot ۱) = \text{sign}(۱۰) = +۱$$

اشتباه طبقه‌بندی شده است.

حال به‌روزرسانی وزن‌ها انجام میشود.

$$\mathbf{w}_{\text{new}} = \mathbf{w} + y \cdot \mathbf{x} = \begin{bmatrix} ۵ \\ ۵ \end{bmatrix} + (-۱) \cdot \begin{bmatrix} ۱ \\ ۱ \end{bmatrix} = \begin{bmatrix} ۴ \\ ۴ \end{bmatrix}$$

بعد از بروزرسانی پیش بینی را دوباره انجام می‌دهیم.

$$f(\mathbf{x}) = \text{sign}(۴ \cdot ۱ + ۴ \cdot ۱) = \text{sign}(۸) = +۱$$

می‌بینیم که همچنان اشتباه طبقه‌بندی شده است.

۲. (۱۲ نمره) شما تصمیم گرفته‌اید معلم شوید. تنها مشکلی که وجود دارد این است که نمی‌خواهید زمان زیادی را صرف تصحیح مقاله‌ها کنید، بنابراین تصمیم می‌گیرید همه آن‌ها را با یک طبقه‌بند خطی نمره‌دهی کنید. طبقه‌بند شما تعداد کلمات ۷ حرفی ($f_۷$) و ۸ حرفی ($f_۸$) در یک مقاله را در نظر می‌گیرد و سپس براساس این دو عدد به مقاله نمره A یا F می‌دهد.

شما چهار مقاله نمره‌دار برای یادگیری در اختیار دارید:

تعداد کلمات ۷ حرفی ($f_۷$)	تعداد کلمات ۸ حرفی ($f_۸$)	نمره ((-)A (+), F)
۳	۲	A
۱	۱	F
۴	۳	A
۰	۲	F

(آ) ابتدا با ذکر دلیل مشخص کنید که داده‌های آموزشی با ویژگی‌های داده‌شده به‌صورت خطی جداپذیر هستند یا نه.

(ب) شما تصمیم می‌گیرید الگوریتم پرسپترون را اجرا کنید و با خوش‌بینی نسبت به توانایی‌های نگارش مقاله‌ی دانش‌آموزان، بردار وزن خود را به صورت $(۱, ۰, ۰)$ مقداردهی اولیه می‌کنید. اگر امتیاز خروجی طبقه‌بند شما بیشتر از ۰ باشد، نمره‌ی A می‌دهد؛ اگر ۰ یا کمتر باشد، نمره‌ی F می‌دهد. بردار وزن حاصل را پس از مشاهده‌ی اولین و دومین مثال آموزشی با استفاده از الگوریتم پرسپترون بنویسید.

(ج) (۴ نمره) برای هر یک از قوانین تصمیمگیری زیر، مشخص کنید آیا بردار وزنی وجود دارد که آن قانون را نمایش دهد. اگر پاسخ «بله» است، بردار وزنی مربوطه را بنویسید.

- مقاله نمره A میگیرد اگر و تنها اگر ($f_۷ + f_۸ \geq ۷$) برقرار باشد.
- مقاله نمره A میگیرد اگر و تنها اگر ($f_۸ \geq ۴$ و $f_۷ \geq ۵$) برقرار باشد.
- مقاله نمره A میگیرد اگر و تنها اگر ($f_۸ \geq ۴$ یا $f_۷ \geq ۵$) برقرار باشد.
- مقاله نمره A میگیرد اگر و تنها اگر بین ۴ تا ۶ (شامل هر دو) کلمه ۷ حرفی و بین ۳ تا ۵ کلمه ۸ حرفی داشته باشد.

حل.

(آ) درست. این داده‌ها با یک مرز خطی قابل جدا شدن هستند. به عنوان مثال، بردار وزن (۲، ۵، ۱) می‌تواند همه نقاط را به درستی طبقه‌بندی کند.

(ب) :وزن اولیه برابر است با

$$(۱,\ ۰,\ ۰)$$

پس از دیدن نمونه اول با بردار ویژگی $(۱,\ ۲,\ ۱)$ و برچسب مثبت، (A) حاصل ضرب داخلی برابر است با:

$$۱ \cdot ۱ + ۰ \cdot ۲ + ۰ \cdot ۱ = ۱ > ۰$$

پس پیش‌بینی درست بوده و وزن تغییر نمی‌کند:

$$(۱,\ ۰,\ ۰)$$

پس از دیدن نمونه دوم با بردار ویژگی $(۱,\ ۰,\ ۲)$ و برچسب منفی، (F) حاصل ضرب داخلی برابر است با:

$$۱ \cdot ۱ + ۰ \cdot ۰ + ۰ \cdot ۲ = ۱ > ۰$$

پیش‌بینی اشتباه بوده، پس به‌روزرسانی انجام می‌دهیم:

$$(۱,\ ۰,\ ۰) + (-۱) \cdot (۱,\ ۰,\ ۲) = (۰,\ ۰,\ -۲)$$

(ج)
- بله. شرط $f_7 + f_8 \geq 7$ یک مرز خطی ایجاد می‌کند و می‌توان آن را با بردار وزن (۱، ۱، -۷) نمایش داد.
- خیر. شرط $f_8 \geq 4$ و $f_7 \geq 5$ یک منطقه اشتراکی غیرخطی تعریف می‌کند و قابل نمایش با یک بردار وزن نیست.
- خیر. شرط $f_8 \geq 4$ یا $f_7 \geq 5$ یک اجتماع از دو ناحیه جداگانه است و قابل نمایش با یک مرز خطی نیست.
- خیر. داشتن تعداد مشخصی از کلمات بین دو بازه، یک ناحیه بسته و محدود (مستطیلی) تعریف می‌کند که نمی‌توان آن را با یک مرز خطی نمایش داد.

۳. (۱۵ نمره) فرض کنید یک مجموعه‌داده با سه ویژگی ورودی دودویی به صورت زیر A, B, C و یک ویژگی خروجی دودویی Y در اختیار داریم.

ویژگی‌های ورودی مقادیر $\{۰, ۱\}$ می‌گیرند، در حالی که Y مقادیر {نادرست, درست} را می‌گیرد.

A	B	C	Y
۰	۱	۱	درست
۱	۱	۰	درست
۱	۰	۱	نادرست
۱	۱	۱	نادرست
۰	۱	۱	درست
۰	۰	۰	درست
۰	۱	۱	نادرست
۱	۰	۱	نادرست
۰	۱	۰	درست
۱	۱	۱	درست

با توجه به اطلاعات ذکر شده، به موارد زیر پاسخ دهید.

(آ) Naive Bayes رکورد $(A=۱, B=۱, C=۰)$ را چگونه طبقه‌بندی می‌کند؟

(ب) Naive Bayes رکورد $(A=۰, B=۰, C=۱)$ را چگونه طبقه‌بندی می‌کند؟

(ج) Naive Bayes رکورد $(A=۰, B=۰, C=۰)$ را با استفاده از Laplace smoothing(k=1) چگونه طبقه‌بندی می‌کند؟

(د) آیا افزودن تنها یک رکورد می‌تواند باعث تغییر طبقه‌بندی رکورد $(A=۱, B=۰, C=۱)$ شود؟

حل.

(آ) «درست» $(1,1,0) \rightarrow$

$$P_{\mathrm{T}} = P(Y{=}\text{درست})\, P(A{=}۱|\mathrm{T})\, P(B{=}۱|\mathrm{T})\, P(C{=}۰|\mathrm{T}) = \frac{۶}{۱۰} \times \frac{۲}{۶} \times \frac{۵}{۶} \times \frac{۳}{۶} = \frac{۱}{۱۲},$$

$$P_{\mathrm{F}} = P(Y{=}\text{نادرست})\, P(A{=}۱|\mathrm{F})\, P(B{=}۱|\mathrm{F})\, P(C{=}۰|\mathrm{F}) = \frac{۴}{۱۰} \times \frac{۳}{۴} \times \frac{۲}{۴} \times \frac{۰}{۴} = ۰.$$

چون $P_{\mathrm{T}} > P_{\mathrm{F}}$، برچسب «درست» برگردانده می‌شود.

(ب) «نادرست» $(0,0,1) \rightarrow$

$$P_{\mathrm{T}} = \frac{۶}{۱۰} \times \frac{۴}{۶} \times \frac{۱}{۶} \times \frac{۳}{۶} = ۰/۰۳۳\ldots,$$

$$P_{\mathrm{F}} = \frac{۴}{۱۰} \times \frac{۱}{۴} \times \frac{۲}{۴} \times \frac{۴}{۴} = ۰/۰۵.$$

از آنجا که $P_{\mathrm{F}} > P_{\mathrm{T}}$ است، «نادرست» پیش‌بینی می‌شود.

(ج) «درست» $(0,0,0) \rightarrow$ فرمول هموارسازی لاپلاس برای متغیرهای دودویی:

$$P(X{=}x|Y{=}y) = \frac{\mathrm{count}(X{=}x, Y{=}y) + ۱}{N_y + ۲}, \qquad P(Y{=}y) = \frac{\mathrm{count}(Y{=}y) + ۱}{N + ۲}.$$

با $N = ۱۰$ و $|Y| = ۲$:

$$P(Y{=}\mathrm{T}) = \frac{۶+۱}{۱۰+۲} = \frac{۷}{۱۲},$$

$$P(Y{=}\mathrm{F}) = \frac{۴+۱}{۱۰+۲} = \frac{۵}{۱۲},$$

$$P(A{=}۰|\mathrm{T}) = \frac{۴+۱}{۶+۲} = \frac{۵}{۸},$$

$$P(B{=}۰|\mathrm{T}) = \frac{۱+۱}{۶+۲} = \frac{۲}{۸},$$

$$P(C{=}۰|\mathrm{T}) = \frac{۳+۱}{۶+۲} = \frac{۴}{۸},$$

$$P(A{=}۰|\mathrm{F}) = \frac{۱+۱}{۴+۲} = \frac{۲}{۶},$$

$$P(B{=}۰|\mathrm{F}) = \frac{۲+۱}{۴+۲} = \frac{۳}{۶},$$

$$P(C{=}۰|\mathrm{F}) = \frac{۰+۱}{۴+۲} = \frac{۱}{۶}.$$

$$P_{\mathrm{T}} = \frac{۷}{۱۲} \times \frac{۵}{۸} \times \frac{۲}{۸} \times \frac{۴}{۸} = \frac{۳۵}{۷۶۸} \approx ۰/۰۴۵۶,$$
$$P_{\mathrm{F}} = \frac{۵}{۱۲} \times \frac{۲}{۶} \times \frac{۳}{۶} \times \frac{۱}{۶} = \frac{۵}{۴۳۲} \approx ۰/۰۱۱۶.$$

از آنجا که $P_{\mathrm{T}} > P_{\mathrm{F}}$، برچسب نهایی «درست» است. (هموارسازی، احتمال صفر طرف «نادرست» را برطرف می‌کند اما نتیجه تغییری نمی‌کند.)

(د) «خیر» $(1,0,1) \rightarrow$ در وضعیت فعلی:

$$P_{\mathrm{T}} = \frac{۶}{۱۰} \times \frac{۲}{۶} \times \frac{۱}{۶} \times \frac{۳}{۶} = ۰/۰۱۶۷,$$
$$P_{\mathrm{F}} = \frac{۴}{۱۰} \times \frac{۳}{۴} \times \frac{۲}{۴} \times \frac{۴}{۴} = ۰/۱۵.$$

بعد از افزودن داده:

$$P_{\mathrm{T}} = \frac{۷}{۱۱} \times \frac{۳}{۷} \times \frac{۲}{۷} \times \frac{۴}{۷} = ۰/۰۴۴۵,$$
$$P_{\mathrm{F}} = \frac{۵}{۱۱} \times \frac{۳}{۵} \times \frac{۲}{۵} \times \frac{۴}{۵} = ۰/۰۸۷۳.$$

افزودن تنها یک نمونه (با هر برچسب یا ویژگی) حداکثر می‌تواند P_T را تا حدود $۰/۰۴۴۵ \approx$ افزایش داده یا P_F را تا حدود $۰/۰۸۷۳ \approx$ کاهش دهد؛
بااین‌حال همچنان $P_F > P_T$ باقی می‌ماند، در نتیجه پاسخ نهایی «خیر» است.

۴. (۱۶ نمره) به پرسش های زیر با استفاده از مجموعه دادهٔ سوال قبل و جدول آنتروپی زیر پاسخ دهید.

Specific Conditional Entropies		
$\mathbf{H(Y\|A = ۱, C = ۰) = ۰/۰۰}$	$\mathbf{H(Y\|A = ۰, B = ۰) = ۰/۰۰}$	$\mathbf{H(Y\|A = ۰) = ۰/۷۲}$
$\mathbf{H(Y\|A = ۱, C = ۱) = ۰/۸۱}$	$\mathbf{H(Y\|A = ۰, B = ۱) = ۰/۸۱}$	$\mathbf{H(Y\|A = ۱) = ۰/۹۷}$
$\mathbf{H(Y\|B = ۰, C = ۰) = ۰/۰۰}$	$\mathbf{H(Y\|A = ۱, B = ۰) = ۰/۰۰}$	$\mathbf{H(Y\|B = ۰) = ۰/۹۲}$
$\mathbf{H(Y\|B = ۰, C = ۱) = ۰/۰۰}$	$\mathbf{H(Y\|A = ۱, B = ۱) = ۰/۹۲}$	$\mathbf{H(Y\|B = ۱) = ۰/۸۶}$
$\mathbf{H(Y\|B = ۱, C = ۰) = ۰/۰۰}$	$\mathbf{H(Y\|A = ۰, C = ۰) = ۰/۰۰}$	$\mathbf{H(Y\|C = ۰) = ۰/۰۰}$
$\mathbf{H(Y\|B = ۱, C = ۱) = ۰/۹۷}$	$\mathbf{H(Y\|A = ۰, C = ۱) = ۰/۹۲}$	$\mathbf{H(Y\|C = ۱) = ۰/۹۹}$

(آ) کدام‌یک از A, B, C بیشترین information gain را دارد.

(ب) درخت تصمیم را بدون استفاده از عملیات هرس برای این سوال رسم کنید. در هنگام ساخت درخت، اگر در مرحله‌ایی به دو ویژگی یکسان برای تقسیم برخوردید، ویژگی را براساس ترتیب حروف الفبا انتخاب کنید. همچنین برای برچسب‌گذاری برگ‌ها، از True در صورت تساوی استفاده کنید.

(ج) خروجی درخت تصمیم شما برای ورودی‌های زیر چیست؟

(۱) $(A = ۰,\ B = ۰,\ C = ۱)$

(۲) $(A = ۱,\ B = ۰,\ C = ۰)$

(د) اگر تمامی رئوس به جز ریشه را از درخت تصمیم هرس کنید و فرض کنید در صورت تساوی True را انتخاب کنید. خروجی برای ورودی‌های زیر چه خواهد بود؟

(۱) $(A = ۰,\ B = ۰,\ C = ۱)$

(۲) $(A = ۱,\ B = ۰,\ C = ۰)$

حل.

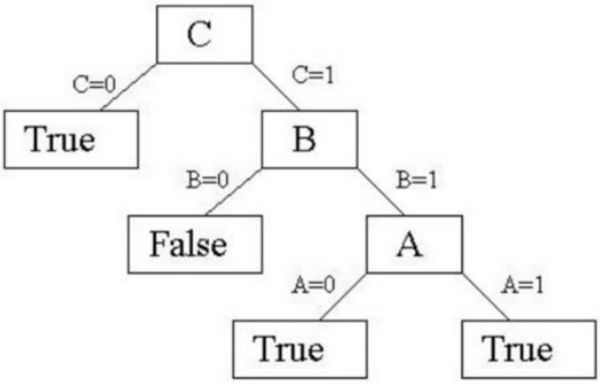
(آ)

$$H(Y|A) = P(A = ۰)*H(Y|A = ۰)+P(A = ۱)*H(Y|A = ۱) = \frac{۵}{۱۰}*۰/۷۲+\frac{۵}{۱۰}*۰/۹۷ = ۰/۸۴۵$$

$$H(Y|B) = P(B = ۰)*H(Y|B = ۰)+P(B = ۱)*H(Y|B = ۱) = \frac{۳}{۱۰}*۰/۹۲+\frac{۷}{۱۰}*۰/۸۶ = ۰/۸۷۸$$

$$H(Y|C) = P(C = ۰)*H(Y|C = ۰)+P(C = ۱)*H(Y|C = ۱) = \frac{۳}{۱۰}*۰/۰۰+\frac{۷}{۱۰}*۰/۹۹ = ۰/۶۹۳$$

information gain زمانی بیشینه میشود که $X = C$



(ب) توجه کنید که در حالتی که $C = ۱$، ویژگی اولی که انتخاب می‌شود B است نه A، چرا که:

$$H(Y|A,C=۱) = P(A=۰|C=۱)H(Y|A=۰,C=۱)+P(A=۱|C=۱)H(Y|A=۱,C=۱)$$
$$= \frac{۳}{۷} * ۰/۹۲ + \frac{۴}{۷} * ۰/۸۱ = ۰/۸۵۷$$

که بیشتر از:

$$H(Y|B,C=۱) = P(B=۰|C=۱)H(Y|B=۰,C=۱)+P(B=۱|C=۱)H(Y|B=۱,C=۱)$$
$$= \frac{۲}{۷} * ۰/۰۰ + \frac{۵}{۷} * ۰/۹۷ = ۰/۶۹۳$$

(ج)
- False
- True

(د)
- False
- True

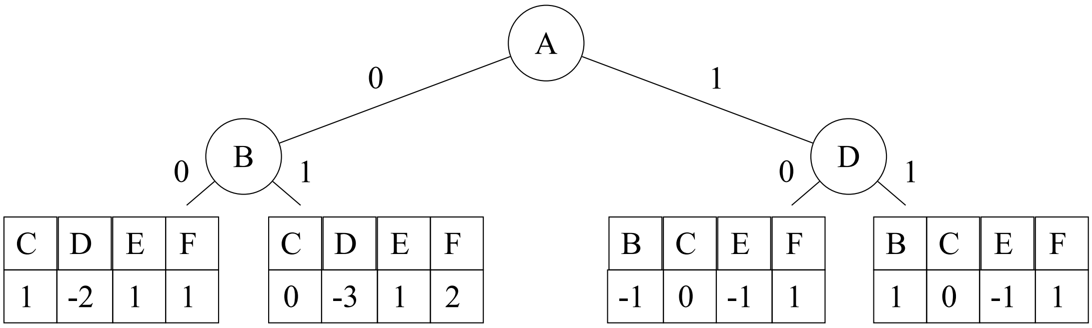
۵. (۱۲ نمره)

نیکا که به دنبال دسته‌بندی کننده‌ی مناسب خودش می‌گشت، الگوریتم جدیدی به نام «درخت پرسپترون» به ذهنش رسید که ویژگی‌هایی را از هر دو مدل درخت تصمیم و الگوریتم پرسپترون را باهم آمیخته می‌کند. درخت پرسپترون مشابه درخت تصمیم است، با این تفاوت که در هر برگ درخت به جای رای اکثریت [۱] از یک مدل پرسپترون استفاده می‌شود.

برای ساخت درخت پرسپترون، نخستین گام دنبال کردن روند عادی یک الگوریتم درخت تصمیم و جدا کردن بر اساس ویژگی‌هاست تا زمانی که به حداکثر عمق مجاز درخت برسیم. پس از رسیدن به این عمق، در هر برگ یک مدل پرسپترون **بر پایه‌ی ویژگی‌های استفاده نشده در آن شاخه** یاد گرفته می‌شود. در نتیجه با داشتن یک نمونه‌ی جدید، ابتدا نمونه را وارد درخت تصمیم خود می‌کنیم و با پیش‌روی به یکی از برگ‌ها می‌رسیم. سپس مدل پرسپترون آن برگ را بر روی ویژگی‌های بررسی نشده انجام می‌دهیم و دسته‌بندی می‌کنیم.

[۱] Majority Voting

فرض کنید که یک dataset با ۶ ویژگی دودویی [۲] {A, B, C, D, E, F} و دو دسته‌ی خروجی {۱، ۱-} دارید. یک درخت پرسپترون با عمق ۲ برای این dataset در شکل زیر داده شده است. وزن‌های پرسپترون در هر برگ داده شده است. فرض کنید مقدار bias برای هر برگ $b = ۱$ است.



آ) خروجی درخت پرسپترون برای نمونه‌ی $\text{x} = \{۱, ۱, ۰, ۱, ۰, ۱\}$ چیست؟

ب) درستی یا نادرستی عبارت‌های زیر را درباره‌ی درخت پرسپترون مشخص کنید.

۱. مرز تصمیم‌گیری درخت پرسپترون همواره خطی است.

۲. برای مقادیر کوچک حداکثر عمق (۲-۳)، احتمال underfit شدن درخت پرسپترون بیش‌تر از درخت تصمیم است.

ج) فرض کنید $\mathcal{D}$ که یک dataset است، به ما داده شده است. از دو ساختار درخت تصمیم متفاوت برای یادگیری این dataset استفاده می‌کنیم. یک بار از Information Gain و بار دیگر از Training Error Rate برای جدا کردن نمونه‌ها استفاده می‌کنیم (در حالت دوم در هر گام بر اساس ویژگی‌ای تقسیم‌بندی می‌کنیم که کم‌ترین Training Error را می‌دهد). با فرض این که هر دو درخت تا رسیدن به Training Error صفر آموزش داده می‌شوند، کدام موارد زیر درست هستند؟

- هر دو درخت در ریشه بر پایه‌ی ویژگی یکسانی تقسیم‌بندی می‌کنند.
- هر دو درخت عمق یکسان خواهند داشت.
- هر دو درخت برای هر Data Point در $\mathcal{D}$ خروجی یکسان خواهند داشت.
- هر دو درخت برای هر ورودی، خروجی یکسان خواهند داشت.
- هر دو درخت از تمام ویژگی‌های درون Dataset استفاده خواهند کرد.

حل.

آ) با توجه به این که $A = D = ۱$ نمونه به راست‌ترین برگ درخت می‌رود که در آن جا

$$\text{out} = ۱ \times ۱ + ۰ \times ۰ + (-۱) \times ۰ + ۱ \times ۱ + ۱ = ۳ \Rightarrow \text{sign}(۳) = ۱$$

پس خروجی برابر ۱ است.

ب) ۱. نادرست - درخت پرسپترون در ساختار خود درخت تصمیم را نیز دارد که یک Classifier غیر خطی است.

۲. نادرست - به طور کلی درخت پرسپترون ساختار پیچیده‌تری از درخت تصمیم با عمق یکسان دارد و در نتیجه بیش‌تر تمایل به overfit شدن دارد.

ج) با توجه به این که می‌دانیم در هر دو $\text{Training Error} = ۰$ نتیجه می‌گیریم هر دو تمام نمونه‌های درون dataset را درست دسته‌بندی می‌کنند. پس مورد سوم درست است. هیچ کدام از موارد دیگر لزوماً برقرار نیستند.

[۲] binary

۶. (۲۰ نمره) یکی از روش‌های منظم‌سازی در مسائل رگرسیون خطی، روش Lasso است. در این روش، نرم L1 وزن‌های مدل در تابع خطا وارد می‌شود. این کار باعث می‌شود که پاسخ نهایی مسئله به‌صورت پراکنده‌تری (sparse) باشد. در این مسئله، خواهیم دید که چگونه جمله‌ی نرم L1 منجر به افزایش پراکندگی می‌شود.
هدف پراکنده کردن جواب نهایی (تنک کردن آن) یا همون sparse کردن آن بهینه کردن فضای ذخیره سازی داده ها ، ساده کردن محاسبات بعدی و جلو گیری از بیش برازش (overfitting) هست.
$\mathbf{X} \in \mathbf{R}^{n\times d}$ ماتریسی هست که هر سطر آن یک مشاهده از d ویژگی می باشد و در کل n مشاهده داریم.
$\mathbf{y} \in \mathbf{R}^{n}$ بردار برچسب ما می باشد. حال فرض کنید که $\mathbf{w} \in \mathbf{R}^{d}$ بردار وزن مدل رگرسیون ما می باشد و w^* وزن های بهینه می باشند. همچنین فرض کنید مشاهدات ما دارای خاصیت $X^TX = \mathbf{I}$ می باشند. (سفید شده اند)
در رگرسیون لاسو بردار وزن های بهینه به صورت زیر به دست میاید :

$$\mathbf{w}^* = argmin_w J_\lambda(\mathbf{w})$$
$$J_\lambda = \frac{1}{2}||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_1$$

(آ) ابتدا نشان می‌دهیم که سفیدسازی داده‌ها باعث مستقل شدن ویژگی‌ها می‌شود، به‌طوری‌که w_i^* تنها از ویژگی iام نتیجه‌گیری می‌شود. برای اثبات این موضوع، ابتدا نشان دهید که J_λ را می‌توان به صورت زیر نوشت:

$$J_\lambda(\mathbf{w}) = g(\mathbf{y}) + \sum_{i=۱}^{d} f(\mathbf{X}_{:,i}, \mathbf{y}, w_i, \lambda)$$

که در آن $\mathbf{X}_{:,i}$ ستون iام ماتریس $\mathbf{X}$ است.

(ب) اگر $w_i \geq ۰$ باشد، مقدار w_i را بیابید.

(ج) اگر $w_i < ۰$ باشد، مقدار w_i را بیابید.

(د) با توجه به بخش‌های قبلی، تحت چه شرایطی w_i برابر صفر خواهد شد؟ این شرایط چگونه قابل اعمال هستند؟

(ه) همان‌طور که می‌دانیم، در رگرسیون ریج، جمله‌ی منظم‌سازی در تابع هزینه به صورت $\frac{۱}{۲}\lambda||\mathbf{w}||_۲^۲$ ظاهر می‌شود. در این حالت، w_i تحت چه شرایطی برابر صفر می‌شود؟ تفاوت این حالت با حالت قبلی چیست؟

حل.

(آ) بر اساس روابط داده شده در سوال داریم که :

$$\mathbf{w}^* = argmin_w J_\lambda(\mathbf{w}), \quad J_\lambda(\mathbf{w}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 + \lambda \cdot ||\mathbf{w}||_1 \quad (\lambda > 0)$$
$$J_\lambda(\mathbf{w}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \cdot ||\mathbf{w}||_1$$
$$\Rightarrow \frac{1}{2}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}) + \lambda \cdot ||\mathbf{w}||_1$$
$$J_\lambda(\mathbf{w}) = \frac{1}{2}(\mathbf{y}^T\mathbf{y} + (\mathbf{X}\mathbf{w})^T(\mathbf{X}\mathbf{w}) - 2(\mathbf{X}\mathbf{w})^Ty) + \lambda \cdot ||\mathbf{w}||_1$$

به خاطر سفید سازی ای که انجام داده بودیم داریم که :

$$\mathbf{X}^T\mathbf{X} = \mathbf{I} \Rightarrow J_\lambda(\mathbf{w}) = \frac{1}{2}\mathbf{y}^T\mathbf{y} + \frac{1}{2}\mathbf{w}^T\mathbf{w} - (\mathbf{X}\mathbf{w})^T\mathbf{y} + \lambda \cdot ||\mathbf{w}||_1$$

$$\mathbf{X}\mathbf{w} = \sum_{i=1}^{d} \mathbf{X}_i \cdot w_i$$

که در رابطه بالا $\mathbf{X}_i$ ستون های ماتریس $\mathbf{X}$ هستند. ضرب $\mathbf{X}\mathbf{w}$ به ما برداری $\mathbf{R^n}$ میدهد که هر سطر آن مجموع وزندار هر سطر ماتریس $\mathbf{X}$هست.

$$\mathbf{X}\mathbf{w} = \begin{bmatrix} X_{11} \cdot w_1 + X_{12} \cdot w_2 + X_{13} \cdot w_3 + ... + X_{1d} \cdot w_d \\ X_{21} \cdot w_1 + X_{22} \cdot w_2 + X_{23} \cdot w_3 + ... + X_{2d} \cdot w_d \\ X_{31} \cdot w_1 + X_{32} \cdot w_2 + X_{33} \cdot w_3 + ... + X_{3d} \cdot w_d \\ ... \\ ... \\ X_{n1} \cdot w_1 + X_{n2} \cdot w_2 + X_{n3} \cdot w_3 + ... + X_{nd} \cdot w_d \end{bmatrix}_n = \mathbf{Q}$$

$$(\mathbf{X}\mathbf{w})^T\mathbf{y} = \mathbf{Q}^T\mathbf{y} = \sum_{i=1}^{n} Q_i \cdot y_i$$

$$\mathbf{w}^T\mathbf{w} = \sum_{i=1}^{d} w_i^2$$

$$J_\lambda(\mathbf{w}) = \underbrace{\frac{1}{2}\mathbf{y}^T\mathbf{y}}_{g(\mathbf{y})} + \frac{1}{2}\sum_{i=1}^{d} w_i^2 - \sum_{i=1}^{n} (\mathbf{X}\mathbf{w})_i^T y_i + \lambda \cdot \sum_{i=1}^{d} |w_i|$$

برای برطرف کردن اندیس جمع $\sum_{i=۱}^{n}(\mathbf{X}\mathbf{w})_i^T y_i$ داریم که :

$$(\mathbf{X}\mathbf{w})^T\mathbf{y} = \sum_{i=1}^{d} w_i \mathbf{X^T y}$$

$$J_\lambda(\mathbf{w}) = \underbrace{\frac{1}{2}\mathbf{y}^T\mathbf{y}}_{g(\mathbf{y})} + \frac{1}{2}\sum_{i=1}^{d} w_i^2 - \sum_{i=1}^{d} w_i \mathbf{X^T y} + \lambda \cdot \sum_{i=1}^{d} |w_i|$$

$$\Rightarrow g(\mathbf{y}) + \sum_{i=1}^{d} \frac{1}{2} \cdot w_i^2 - w_i \cdot \mathbf{X}^T\mathbf{y} + \lambda \cdot |w_i| = g(\mathbf{y}) + \sum_{i=1}^{d} f(\mathbf{X}_{:,i}, \mathbf{y}, w_i, \lambda)$$

حال برای اینکه نشون دهیم w_i^* تنها از ویژگی i نتیجه میشود داریم که :

$$J_\lambda(\mathbf{w}) = \frac{1}{2}\mathbf{y}^T\mathbf{y} + \sum_{i=1}^{d} \frac{1}{2} w_i^2 - w_i \cdot \mathbf{X}_i^T\mathbf{y} + \lambda \cdot |w_i|$$

$$\frac{\partial J}{\partial w_i} = w_i - \mathbf{X}_i^T\mathbf{y} + \lambda sgn(w_i)$$

$$\frac{\partial J}{\partial w_i} = 0 \Rightarrow w_i = \mathbf{X}_i^T\mathbf{y} - \lambda \cdot sgn(w_i)$$

همانطور که در رابطه آخر میبینیم w_i تنها به ویژگی iام بستگی دارد. همچنین میبینیم که فرم بسته ای برای جواب وجود ندارد.

(ب)

$$w_i^* \geq ۰ \Rightarrow w_i^* = \mathbf{X_i}^T\mathbf{y} - \lambda$$

(ج)

$$w_i^* < ۰ \Rightarrow w_i^* = \mathbf{X_i}^T\mathbf{y} + \lambda$$

(د)

$$w_i^* = ۰ \Rightarrow \mathbf{X}_i^T\mathbf{y} = sgn(w_i^*)\lambda \Rightarrow \lambda = -sng(w_i^*)\mathbf{X}_i^T y$$

اگر مقدار λ را برابر مقدار یافت شده در قسمت قبل قرار دهیم ، w_i^* صفر میشود.

(ه)

$$\begin{aligned}
R = \frac{۱}{۲}\lambda||\mathbf{w}||_۲^۲ &\Rightarrow J_\lambda(\mathbf{w}) = \frac{۱}{۲}||\mathbf{y} - \mathbf{X}\mathbf{w}||_۲^۲ + \frac{۱}{۲}\lambda||\mathbf{w}||_۲^۲ \\
J_\lambda &= \frac{۱}{۲}(\mathbf{y}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - ۲(\mathbf{X}\mathbf{w})^T\mathbf{y}) + \frac{۱}{۲}\mathbf{w}^T\mathbf{w} \\
&= \frac{۱}{۲}\sum_{j=۱}^{n} y_i^۲ + \frac{۱}{۲}\sum_{i=۱}^{d} w_i^۲ - \sum_{i=۱}^{d} w_i \cdot (\mathbf{X}_{:,i}^T\mathbf{y}) + \frac{۱}{۲}\lambda\sum_{i=۱}^{d} w_i^۲ \\
\frac{\partial J}{\partial w_i} = ۰ &\Rightarrow w_i - \mathbf{X}_{:,i}^T\mathbf{y} + \lambda \cdot w_i = ۰ \Rightarrow w_i = \frac{\mathbf{X}_{:,i}^T\mathbf{y}}{۱ + \lambda}
\end{aligned}$$

در مسیله ridge تنها وقتی که برچسب داده ها بر ویژگی داده عمود باشد ، $w_i^* = ۰$ میشود ($\mathbf{X}_{:,i}^T\mathbf{y} = ۰$)
ولی در مسیله لاسو میتوانیم با انتخاب بهینه λ که در بخش قبل به دست آوردیم به این نتیجه برسیم.

۷. (۱۵ نمره) تابع زیر را در نظر بگیرید :

$$f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$$

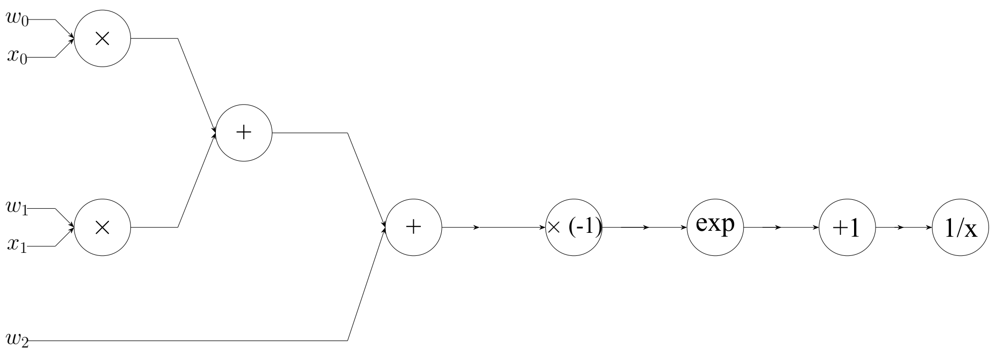
شبکه متناظر با این تابع در زیر آمده است. با فرض ورودی های

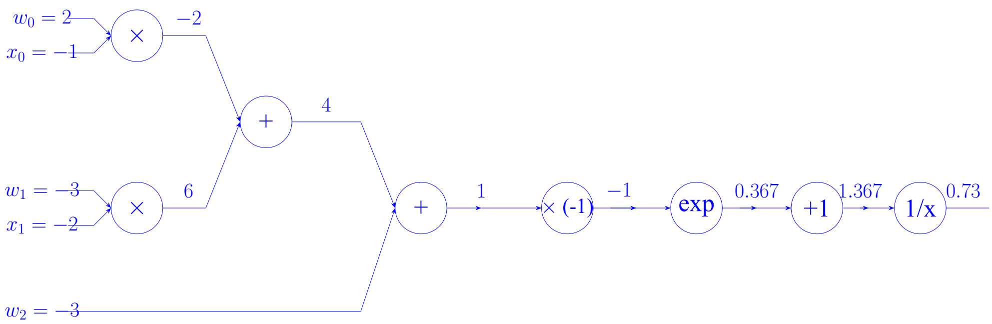
$$w_0 = 2,\ x_0 = -1,\ w_1 = -3,\ x_1 = -2,\ w_2 = -3$$

ابتدا خروجی نهایی را محاسبه کنید. سپس با استفاده از Backpropagation مشتق خروجی نهایی را نسبت به خروجی هر راس حساب کنید.

حل. در هر مرحله از forward-pass ما بر اساس گراف محاسباتی مان خروجی را محاسبه میکنیم که به فرم زیر در میاید :

حال در هر مرحله از Backpropagation نسبت به تابع گره ای که در آن هستیم مشتق میگیریم و به عقب حرکت میکنیم. (برای $w.$ را مینویسیم)





$$g(x) = \frac{1}{x},\ h(x) = exp(ax),\ Q(x) = ax,\ m(x) = b + x$$

$$\frac{\partial \hat{y}}{\partial w_0} = \frac{\partial \hat{y}}{\partial g} \cdot \frac{\partial g}{\partial w_0} = \frac{\partial \hat{y}}{\partial g} \cdot \frac{\partial g}{\partial m} \cdot \frac{\partial m}{\partial w_0} = \frac{\partial \hat{y}}{\partial g} \cdot \frac{\partial g}{\partial m} \cdot \frac{\partial m}{\partial h} \cdot \frac{\partial h}{\partial w_0} = \frac{\partial \hat{y}}{\partial g} \cdot \frac{\partial g}{\partial m} \cdot \frac{\partial m}{\partial h} \cdot \frac{\partial h}{\partial Q} \cdot \frac{\partial Q}{\partial w_0}$$

$$1)\ \frac{\partial \hat{y}}{\partial w_0} = \frac{\partial \hat{y}}{\partial g}\Big|_{g=1.367} \cdot \frac{\partial g}{\partial w_0} = \frac{-1}{x^2}\Big|_{x=1.367} \cdot \frac{\partial g}{\partial w_0} = -0.53 \cdot \frac{\partial g}{\partial w_0}$$

$$2)\ \frac{\partial \hat{y}}{\partial g} \cdot \frac{\partial g}{\partial m}\Big|_{m=0.367} \cdot \frac{\partial m}{\partial w_0} = -0.53 \cdot 1 \cdot \frac{\partial m}{\partial w_0}$$

$$3)\ \frac{\partial \hat{y}}{\partial g} \cdot \frac{\partial g}{\partial m} \cdot \frac{\partial m}{\partial h} \cdot \frac{\partial h}{\partial w_0} = -0.53 \cdot 1 \cdot exp(x)|_{-1} \frac{\partial h}{\partial w_0} = -0.53 \cdot 1 \cdot e^{-1} = -0.194 \frac{\partial h}{\partial w_0}$$

$$4)\ \frac{\partial \hat{y}}{\partial g} \cdot \frac{\partial g}{\partial m} \cdot \frac{\partial m}{\partial h} \cdot \frac{\partial h}{\partial Q} \cdot \frac{\partial Q}{\partial w_0} = -0.194 \cdot \frac{\partial Q}{\partial w_0}\Big|_{a=-1} = -0.194 \cdot -1 = 0.194 = T$$

$$5)\ \frac{\partial T}{\partial m} \cdot \frac{\partial m}{\partial w_0} = 0.194$$

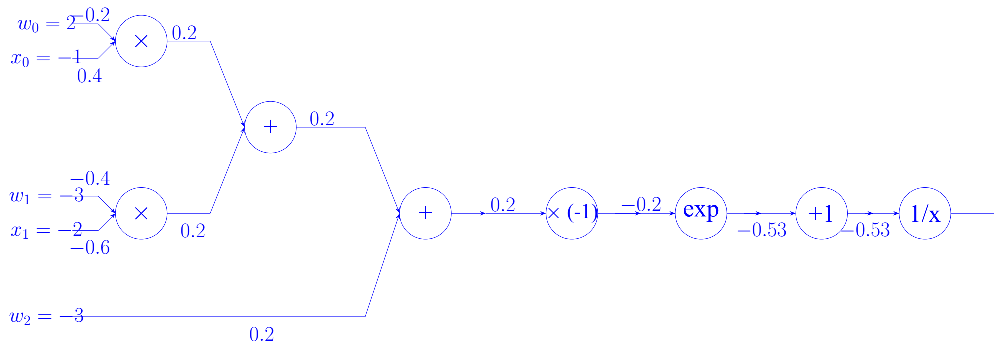
$$6)\ \frac{\partial T}{\partial m} \cdot \frac{\partial m}{\partial m} \cdot \frac{\partial m}{\partial w_0} = 0.194$$

$$7)\ \frac{\partial T}{\partial m} \cdot \frac{\partial m}{\partial m} \cdot \frac{\partial m}{\partial Q} \cdot \frac{\partial Q}{\partial w_0} = 0.194 \cdot x_0 | x_0 = -1 = -0.194$$

شماره گذاری برای راحت تر شدن خواندن مراحل انجام شده است و صرفا برای نمایش یک دور backpropagation می باشند و نیازی به نوشتن آنها در پاسخ نبوده است و درستی جواب آخر کافی می باشد. همانطور که دیدید ۷ مشتق جزیی نسبت به توابع و اجزای مختلف شکبه در مسیر متغیر w. گرفتیم تا نرخ تغییرات با توجه به متغیر مشخص شده را بیابیم که این تعداد برابر تعداد نود های بین متغیر مورد نظر و خروجی شبکه بود.

مقادیر نهایی Backpropagation در هر مرحله در زیر آمده است : (مقادیر به دو رقم اعشار گرد شده اند)

با استفاده از مقادیر به دست آمده میتوان متغیر ها را طوری تغییر داد تا به جواب بهینه مان برسیم (الگوریتم گرادیان کاهشی نمونه ای از این دسته روش ها می باشد.)



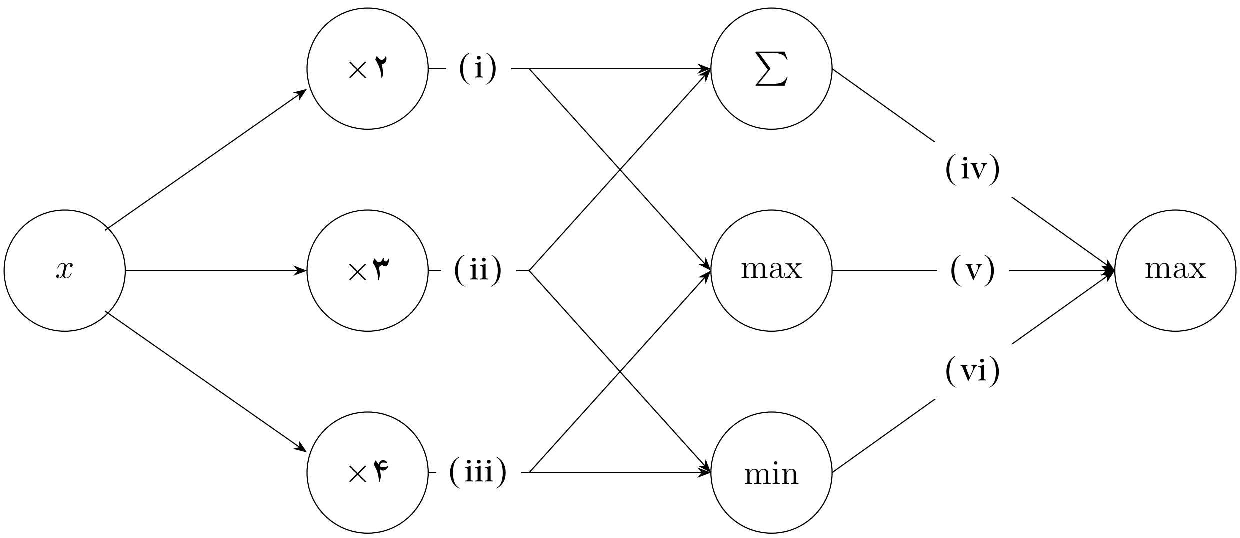
۸. (۱۵+۵ نمره)

(آ) عملیات Forward Propagation را برای شبکه‌ی عصبی زیر با ورودی $x = ۱$ انجام دهید.

۱) مقادیر مربوط به (i), (ii), (iii) را به دست آورید.

۲) مقادیر مربوط به (iv), (v), (vi) را به دست آورید.

۳) مقدار خروجی مدل را به دست آورید.



(ب) (امتیازی) شکل زیر یک شبکه‌ی عصبی با وزن‌های a، b، c، d و f را نشان می‌دهد. ورودی این شبکه $x_۱$ و $x_۲$ هستند. لایه‌ی مخفی نخست به شکل زیر محاسبه می‌شود.

$$r_۱ = \max(c \cdot x_۱ + e \cdot x_۲, ۰) \qquad r_۲ = \max(d \cdot x_۱ + f \cdot x_۲, ۰)$$

خروجی لایه‌ی مخفی دوم نیز به شکل زیر است.

$$s_۱ = \frac{۱}{۱ + \exp(-a \cdot r_۱)} \qquad s_۲ = \frac{۱}{۱ + \exp(-b \cdot r_۲)}$$

خروجی شبکه نیز برابر $y = s_۱ + s_۲$ است.

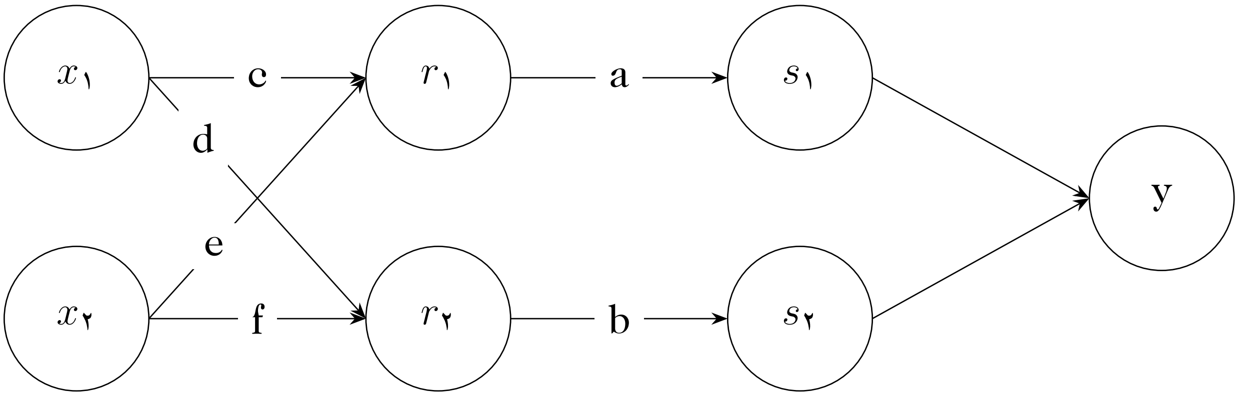
فرض کنید ورودی شبکه $x_۱ = ۱, x_۲ = -۱$ و مقدار وزن‌ها برابر

$$a = ۱, b = ۱, c = ۴, d = ۱, e = ۲, f = ۲$$

باشد. در این صورت خروجی تقریبی شبکه به صورت

$$r_۱ = ۲, r_۲ = ۰, s_۱ = ۰/۹, s_۲ = ۰/۵, y = ۱/۴$$

خواهد بود.



با استفاده از مقادیر تقریبی داده‌شده از الگوریتم backpropagation استفاده کنید تا مقدار مشتق‌های جزئی را حساب کنید. مقادیر را به صورت عددی به دست آورید.

۱) مقدار $\frac{\partial y}{\partial a}$ و $\frac{\partial y}{\partial b}$ را به دست آورید.

۲) مقدار $\frac{\partial r_۱}{\partial c}$ و $\frac{\partial r_۱}{\partial e}$ را به دست آورید.

۳) مقدار $\frac{\partial r_۲}{\partial d}$ و $\frac{\partial r_۲}{\partial f}$ را به دست آورید.

۴) مقدار $\frac{\partial y}{\partial r_۱}$ و $\frac{\partial y}{\partial r_۲}$ را به دست آورید.

۵) با استفاده از مقادیر به دست آمده در بخش ۴ و خاصیت زنجیره‌ای مشتقات، $\frac{\partial y}{\partial c}$ و $\frac{\partial y}{\partial e}$ را به دست آورید.

۶) با استفاده از مقادیر به دست آمده در بخش ۴ و خاصیت زنجیره‌ای مشتقات، $\frac{\partial y}{\partial d}$ و $\frac{\partial y}{\partial f}$ را به دست آورید.

۷) تمام وزن‌های شبکه را یک مرحله با نرخ یادگیری یک بروزرسانی کنید.

حل.

(آ) ۱)

$$(\text{i}) : x \times ۲ = ۲, \quad (\text{ii}) : x \times ۳ = ۳, \quad (\text{iii}) : x \times ۴ = ۴$$

۲)

$$(\text{iv}) : (\text{i}) + (\text{ii}) = ۵, \quad (\text{v}) : \max((\text{i}), (\text{iii})) = ۴, \quad (\text{vi}) : \min((\text{ii}), (\text{iii})) = ۳$$

۳)

$$\text{out} = \max((\text{iv}), (\text{v}), (\text{vi})) = ۵$$

(ب) ۱)

$$\frac{\partial y}{\partial a} = \frac{\partial y}{\partial s_۱}\frac{\partial s_۱}{\partial a} = ۱ \cdot r_۱ s_۱(۱ - s_۱) = ۰/۱۸$$

$$\frac{\partial y}{\partial b} = \frac{\partial y}{\partial s_۲}\frac{\partial s_۲}{\partial b} = ۱ \cdot r_۲ s_۲(۱ - s_۲) = ۰$$

۲)

$$c \cdot x_۱ + e \cdot x_۲ > ۰ \Rightarrow r_۱ = c \cdot x_۱ + e \cdot x_۲ \Rightarrow \frac{\partial r_۱}{\partial c} = x_۱ = ۱, \quad \frac{\partial r_۱}{\partial e} = x_۲ = -۱$$

۳)

$$d \cdot x_۱ + f \cdot x_۲ = -۱ < ۰ \Rightarrow \frac{\partial r_۲}{\partial d} = \frac{\partial r_۲}{\partial f} = ۰$$

۴)

$$\frac{\partial y}{\partial r_۱} = \frac{\partial y}{\partial s_۱}\frac{\partial s_۱}{\partial r_۱} = ۱ \cdot a s_۱(۱ - s_۱) = ۰/۰۹$$

$$\frac{\partial y}{\partial r_۲} = \frac{\partial y}{\partial s_۲}\frac{\partial s_۲}{\partial r_۲} = ۱ \cdot b s_۲(۱ - s_۲) = ۰/۲۵$$

۵)

$$\frac{\partial y}{\partial c} = \frac{\partial y}{\partial r_۱}\frac{\partial r_۱}{\partial c} = ۰/۰۹ \times ۱ = ۰/۰۹$$

$$\frac{\partial y}{\partial e} = \frac{\partial y}{\partial r_۱}\frac{\partial r_۱}{\partial e} = ۰/۰۹ \times -۱ = -۰/۰۹$$

۶)

$$\frac{\partial y}{\partial f} = \frac{\partial y}{\partial r_۲}\frac{\partial r_۲}{\partial f} = ۰$$

$$\frac{\partial y}{\partial d} = \frac{\partial y}{\partial r_۲}\frac{\partial r_۲}{\partial d} = ۰$$

۷)

$$a \leftarrow ۰/۸۲, b \leftarrow ۱, c \leftarrow ۳/۹۱, d \leftarrow ۱, e \leftarrow ۲/۰۹, f \leftarrow ۲$$