دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

**هوش مصنوعی**
**پاییز ۱۴۰۱**
استاد: محمدحسین رهبان

# آزمون پایان‌ترم

- زمان در نظر گرفته شده برای آزمون ۱۸۰ دقیقه است.
- لطفاً پاسخ‌های خود را خوانا و خوش‌خط بنویسید.

## سوالات (۱۰۰ نمره)

۱. (۱۵ نمره) به سوالات زیر به طور مختصر پاسخ دهید:

(آ) درستی یا نادرستی عبارت رو به رو را با ذکر دلیل مشخص کنید: در یک شبکه بیزی که در آن X به شرط Z از Y مستقل است؛ ممکن است فرض استقلال این دو متغیر با شرطی کردن شواهد اضافی برای متغیرهای دیگر در شبکه، برقرار نباشد.

(ب) دو مورد از روش‌های جلوگیری از بیش‌برازش (overfitting) در درخت تصمیم را ذکر کرده و مختصرا توضیح دهید.

(ج) آیا همواره با استفاده از مدل درخت تصمیم بدون محدودیت عمق می‌توان به دقت ۱۰۰٪ رسید؟ توضیح دهید.

(د) برای طبقه‌بندی تصاویر به دو کلاس سگ و گربه، از یک شبکه عصبی عمیق استفاده شده است. در انتهای این شبکه، از یک تابع فعال‌ساز $ReLU$ و سپس از تابع $sigmoid$ استفاده شده است. درصورتی که خروجی نهایی شبکه $\hat{y} \geq ۰/۵$ باشد آن را به عنوان گربه و در غیر این صورت به عنوان سگ در نظر می‌گیریم. آیا این شبکه ایرادی در برچسب‌گذاری تصاویر دارد؟ مختصرا توضیح دهید.

(ه) فرض کنید از رابطه‌ی

$$Q(s,a) \leftarrow (۱-\alpha)Q(s,a) + \alpha\left[R(s,a,s') + \gamma \max_{a'} f\big(Q(s',a'), N(s',a')\big)\right]$$

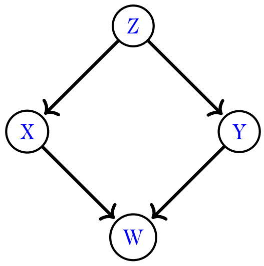
برای بروزرسانی Q-valueها استفاده کنیم که $N(s',a')$ تعداد دفعاتی است که از حالت $s'$ کنش $a'$ را انجام داده‌ایم و $f(x,y) = x + \frac{\beta}{y}$ که $\beta$ عدد ثابت مثبتی است. توضیح دهید که این تغییر چه تاثیری بر جست‌وجوی فضای حالت خواهد گذاشت؟

حل.

(آ) درست، چون مجموعه‌ای از متغیرها به نام W می‌توانند در شبکه موجود باشند که به دلیل وجود ساختار vشکل در میان آن‌ها، گزاره $X \perp\!\!\!\perp Y | Z, W$ برقرار نباشد. (شکل ۱)

(ب)
- جلوگیری از رشد بیش از حد درخت با استفاده از محدود کردن تعداد برگ‌ها و یا عمق درخت
- هرس کردن درخت با استفاده از تست‌های آماری ـ به این صورت که در راس‌های غیر برگ با استفاده از یک تست آماری بررسی می‌کنیم که آیا افزایش دقت مدل واقعا به علت تقسیم بر حسب معیار انتخابی بوده است یا احتمالا به علت وجود نویز در داده است.

(ج) خیر. در صورتی می‌توان به حالت بدون خطا دست یافت که داده‌های یکسان برچسب‌های متفاوتی نداشته باشند.




شکل ۱: سوال ۱ قسمت (آ)

(د) با توجه به این که مقدار خروجی ReLU برای هر ورودی بزرگ‌تر از صفر است، پس از اعمال تابع سیگموید روی خروجی آن، همواره عددی بزرگتر از ۰/۵ خواهیم داشت. در نتیجه برای تمامی تصاویر یک برچسب در نظر گرفته می‌شود که نامطلوب است.

(ه) تعداد دفعاتی که جفت $(s, a)$ را دیده‌ایم در مقدار Q-valueها تاثیرگذار است و چون این تاثیر از رابطه‌ی $f$ ناشی می‌شود که آن هم رابطه‌ی عکس با تعداد دیده شدن‌ها دارد، می‌توان نتیجه گرفت که در صورت مثبت بودن $\beta$، با این شیوه‌ی آپدیت کردن کنش‌هایی که به حالت‌هایی که زیاد دیده نشده‌اند منجر می‌شوند برتری خواهند داشت و این کمک می‌کند تا بهتر بتوان فضای حالت را جست‌وجو کرد.

۲. (۸ نمره) مدل مارکوف زیر را در نظر بگیرید که سه استیت ۱ و ۲ و ۳ دارد. این مدل دارای Transition Matrix به صورت زیر است:

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

(آ) برای این مدل یک نمودار حالت (state diagram) رسم کنید.

(ب) اگر بدانیم که $P(X_1 = 1) = P(X_1 = 2) = \frac{1}{4}$ مقدار $P(X_1 = 3, X_2 = 2, X_3 = 1)$ را محاسبه کنید.

حل.

(آ) نمودار حالت این مدل در شکل ۲ قابل مشاهده است.

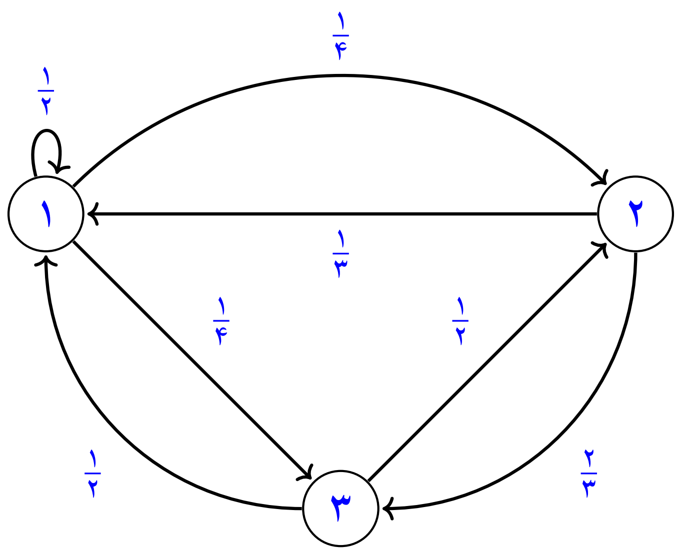
(ب) در ابتدا دقت کنیم که داریم:

$$P(X_1 = 3) = 1 - P(X_1 = 1) - P(X_1 = 2) = 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2}$$

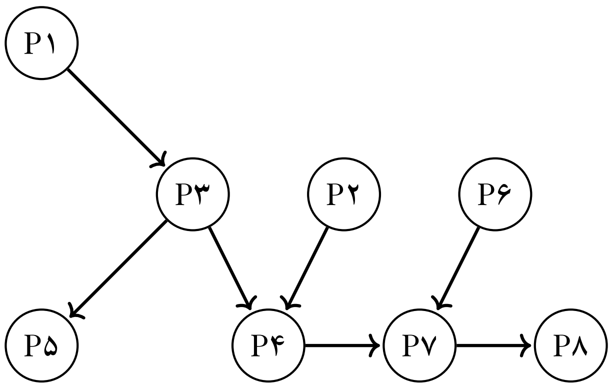
حال می‌توان نوشت:

$$P(X_1 = 3, X_2 = 2, X_3 = 1) = P(X_1 = 3) \times p_{32} \times p_{21} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{12}$$

۳. (۱۲ نمره) فرض کنید گراف زیر را به عنوان یک شبکه‌ی بیزین داریم:




شکل ۲: سوال ۲ قسمت (آ)




درستی یا نادرستی هر کدام از عبارت‌های زیر را با ذکر دلیل مشخص کنید.

(آ) $P_5 \perp\!\!\!\perp P_6|P_7$

(ب) $P_1 \perp\!\!\!\perp P_2|P_5$

(ج) $P_1 \perp\!\!\!\perp P_2|P_8$

(د) $P_5 \perp\!\!\!\perp P_8|P_7$

حل.

(آ) نادرست. همه‌ی سه‌تایی‌ها در مسیر $P5, P3, P4, P7, P6$ فعال هستند در نتیجه $P5$ و $P6$ به شرط $P7$ مستقل نیستند.

(ب) درست. سه‌تایی $P3, P4, P2$ غیرفعال است در نتیجه مسیر بین $P1$ و $P2$ غیرفعال بوده و این دو به شرط $P5$ از هم مستقلند.

(ج) نادرست.
همه‌ی سه‌تایی‌ها در مسیر $P1, P3, P4, P2$ فعال هستند در نتیجه $P1$ و $P2$ به شرط $P8$ مستقل نیستند.

(د) درست. سه‌تایی $P4, P7, P8$ غیرفعال است در نتیجه مسیر بین $P5$ و $P8$ غیرفعال بوده و این دو به شرط $P7$ از هم مستقلند.

۴. (۱۴ نمره) یک مدل خطی به فرم زیر را در نظر بگیرید:

$$y(x_n, \mathbf{w}) = w_۰ + \sum_{i=۱}^{D} w_i x_{ni}$$

خطای آن را نیز به صورت زیر درنظر می‌گیریم:

$$E_D(\mathbf{w}) = \frac{۱}{۲} \sum_{n=۱}^{N} \{y(x_n, \mathbf{w}) - t_n\}^۲$$

حال فرض کنید که یک نویز گوسی $\epsilon_i \sim \mathcal{N}(۰, \sigma^۲ I)$ به هر ورودی $x_i$ اضافه شده است. ($\epsilon_i$ها به صورت i.i.d تولید شده‌اند.) اگر $\widetilde{E}_D(\mathbf{w})$ خطای مدل وقتی از $x_i + \epsilon$ استفاده می‌کنیم باشد، رابطه‌ی زیر را اثبات کنید.

$$\mathbb{E}[\widetilde{E}_D(\mathbf{w})] = E_D(\mathbf{w}) + \frac{N}{۲} \sum_{i=۱}^{D} w_i^۲ \sigma^۲$$

حل. داریم:

$$\begin{aligned} \widetilde{y}_n &= w_۰ + \sum_{i=۱}^{D} w_i (x_{ni} + \epsilon_{ni}) \\ &= y_n + \sum_{i=۱}^{D} w_i \epsilon_{ni} \end{aligned}$$

برای $\widetilde{E}_D(w)$ می‌توانیم بنویسیم:

$$\begin{aligned} \widetilde{E} &= \frac{۱}{۲} \sum_{n=۱}^{N} \{\widetilde{y}_n - t_n\}^۲ \\ &= \frac{۱}{۲} \sum_{n=۱}^{N} \{\widetilde{y}_n^۲ - ۲\widetilde{y}_n t_n + t_n^۲\} \\ &= \frac{۱}{۲} \sum_{n=۱}^{N} \Bigg\{ y_n^۲ + ۲y_n \sum_{i=۱}^{D} w_i \epsilon_{ni} + \left(\sum_{i=۱}^{D} w_i \epsilon_{ni}\right)^۲ \\ &\quad - ۲t_n y_n - ۲t_n \sum_{i=۱}^{D} w_i \epsilon_{ni} + t_n^۲ \Bigg\} \end{aligned}$$

اگر از رابطه‌ی بالا امید ریاضی بگیریم، جمله‌ی دوم و پنجم داخل سیگما با توجه به صفر بودن میانگین $\epsilon_i$ صفر می‌شوند.

اگر جمله‌ی سوم را باز کنیم خواهیم داشت:

$$\mathbb{E}\left[\left(\sum_{i=1}^{D} w_i\epsilon_{ni}\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{D}\sum_{j=1,j\neq i}^{D} w_i\epsilon_{ni}w_j\epsilon_{nj} + \sum_{i=1}^{D} w_i^2\epsilon_{ni}^2\right]$$

$$= \sum_{i=1}^{D}\sum_{j=1,j\neq i}^{D} w_iw_j\mathbb{E}[\epsilon_{ni}\epsilon_{nj}] + \sum_{i=1}^{D} w_i^2\mathbb{E}[\epsilon_{ni}^2]$$

$$= 0 + \sum_{i=1}^{D} w_i^2\sigma^2$$

پس:

$$\mathbb{E}[\widetilde{E}_D(w)] = \frac{1}{2}\sum_{n=1}^{N}\left\{y_n^2 + 0 + \sum_{i=1}^{D} w_i^2\sigma^2 - 2t_ny_n - 0 + t_n^2\right\}$$

$$= \frac{1}{2}\sum_{n=1}^{N}[y_n^2 - 2t_ny_n + t_n^2] + \frac{N}{2}\sum_{i=1}^{D} w_i^2\sigma^2$$

$$= E_D(w) + \frac{N}{2}\sum_{i=1}^{D} w_i^2\sigma^2$$

۵. (۱۰ نمره) با توجه به جدول زیر قصد داریم یک درخت تصمیم بسازیم که مقدار $Y$ را براساس $X_1$، $X_2$ و $X_3$ تعیین کند.

| $X_1$ | $X_2$ | $X_3$ | $Y$ |
|---|---|---|---|
| No | No | No | No |
| Yes | No | Yes | Yes |
| No | Yes | Yes | Yes |
| Yes | Yes | Yes | No |

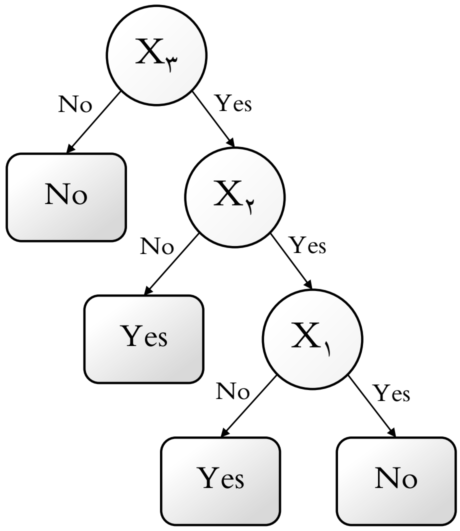
(آ) درخت تصمیم را رسم کنید. در هر مرحله ریشه را براساس Information Gain انتخاب کنید. (لزومی به محاسبه‌ی عددی IG نیست اما در صورت نیاز از آن استفاده کنید.)

(ب) آیا درخت به‌دست آمده (با انتخاب ریشه‌ها از طریق IG) بهینه است؟ اگر بهینه است علت آن را بیان کنید؛ در غیر این صورت درخت بهینه را رسم کنید. (منظور از بهینه، درختی با کوتاه‌ترین ارتفاع ممکن است که نمونه‌های سازگار را جداسازی کند.)

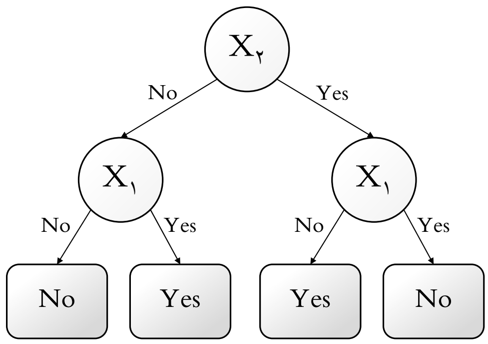
حل.

(آ) درخت در شکل ۳ رسم شده است.

(ب) با توجه به اینکه $Y = X_1 \oplus X_2$ درخت شکل ۴ ارتفاع کم‌تری دارد. به‌طور کلی لزوماً انتخاب ریشه با IG بهینه نیست؛ زیرا در این حالت ویژگی‌ها جداگانه بررسی می‌شوند و ارتباط توام آن‌ها بررسی نمی‌شود.




شکل ۳: درخت با کمک IG




شکل ۴: درخت بهینه

۶. (۱۵ نمره) می‌دانیم در شبکه‌های عصبی از توابع مختلفی برای فعال‌سازی استفاده می‌شود. در این سوال می‌خواهیم به کمک تابع پله، شبکه‌های عصبی ساده‌ای برای پیش‌بینی مقادیر توابع منطقی بدست آوریم. توجه داشته باشید در عبارات زیر، $\wedge$ نماد عملگر منطقی AND ، $\vee$ نماد عملگر منطقی OR و $\neg$ نماد عملگر منطقی NOT است.

در این سوال ورودی‌ها ($x_i$) صفر یا یک هستند و تابع فعال‌ساز هم به صورت زیر است:

$$f(x) = \begin{cases} 0 & -(w_0 + \sum_i w_i x_i) < 0 \\ 1 & -(w_0 + \sum_i w_i x_i) \geq 0 \end{cases}$$

در این فرمول $w_0$ نمایانگر بایاس بوده و بقیه $w_i$ ها وزن‌های نسبت داده شده به یال‌های درونی شبکه عصبی هستند.

(آ) یک شبکه عصبی با یک پرسپترون تشکیل دهید که بتواند تابع منطقی $y = x_1 \vee x_2$ را به درستی محاسبه کند.

(ب) یک شبکه عصبی با یک پرسپترون تشکیل دهید که بتواند تابع منطقی $y = x_1 \wedge x_2$ را به درستی محاسبه کند.

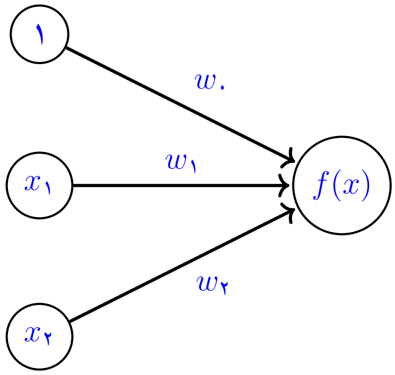
(ج) یک شبکه عصبی متشکل از یک لایه‌ی نهان تشکیل بدهید که تابع منطقی زیر را به درستی محاسبه کند.

$$y = (x_1 \wedge x_2 \wedge x_3) \vee (\neg x_1 \wedge x_3)$$

حل.

(آ) تابع OR در زمان‌هایی که حداقل یکی از ورودی‌ها ۱ است، برابر ۱ می‌شود. باید مقدار: $w_0 + w_1x_1 + w_2x_2$ را طوری تعیین کنیم که به ازای تمامی سه حالت $x_1 = 0, x_2 = 1 | x_1 = 1, x_2 = 0 | x_1 = 1, x_2 = 1$ مقدار آن بزرگتر مساوی ۰ باشد و در حالت $x_0 = 0, x_1 = 0$ کمتر ۰ باشد.
جواب‌های متفاوتی می‌توان بر وزن‌ها یافت که یکی از آن‌ها اعداد زیر هستند:

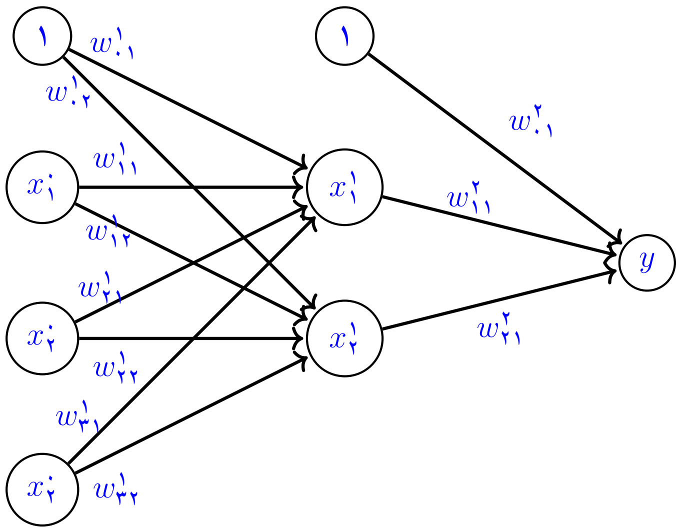
$$w_0 = 0/5, w_1 = -1, w_2 = -1$$




(ب) این قسمت مشابه قسمت قبلی حل می‌شود. یکی از جواب‌های قابل قبول این اعداد هستند:

$$w_0 = 1/5, w_1 = -1, w_2 = -1$$

(ج) در زیر ساختار شبکه و وزن‌های مربوط به آن برای ساخت عبارت خواسته شده قابل مشاهده است:




$$w^1_{01} = 2/1, w^1_{11} = -1, w^1_{21} = -1, w^1_{31} = -1$$

$$w^1_{02} = 1/5, w^1_{12} = 1, w^1_{22} = 0, w^1_{32} = -1$$

$$w^2_{01} = 0/5, w^2_{11} = -1, w^2_{21} = -1$$

توجه کنید که به نوعی در عبارت های متشکل از AND، هر جا خود متغیر ظاهر شده در وزن نظیر آن ۱ قرار داده‌ایم. هر جا نقیض عبارت ظاهر شده در وزن معادل آن ۱− و هر جا هم اصلا ظاهر نشده ۰ قرار داده‌ایم.

* با توجه به ذکر تابع پله در صورت سوال، در صورتی که به جای تابع داده شده $f$ که قرینه‌ی پله است از خود تابع پله استفاده شده باشد نیز مشکلی وجود ندارد. در این حالت تمامی وزن‌های گفته شده در این راه‌حل باید در منفی ضرب شود.

۷. (۱۴ نمره) می‌خواهیم مسئله‌ی پک‌من را با استفاده از مدل MDP حل کنیم. در مسئله‌ی پک‌من، یک موجود به نام پک‌من داریم که در یک جدول قرار دارد، می‌تواند به چهار جهت حرکت کند و هدف آن خوردن غذاهای موجود در صفحه‌ی بازی است. در مسئله‌ی مدل شده، حالت‌ها همان خانه‌های جدول و کنش‌ها حرکت به چهار جهت مختلف هستند. هر حرکتی که به سمت دیوار انجام شود، چه دیوارهای داخل محیط بازی و چه دیوارهای مرزی محیط بازی، بی‌تاثیر بوده و پک‌من در همان خانه‌ی قبلی خود باقی می‌ماند. همچنین فرض کنید در کل محیط بازی فقط یک غذا وجود دارد (که با $\circ$ در نقشه‌ی بازی نشان داده شده است) و پک‌من به محض رسیدن به آن یک امتیاز دریافت کرده و بازی تمام می‌شود.

- در این قسمت محیط بازی را مطابق شکل زیر در نظر بگیرید. (حروف موجود در خانه‌ها نام هر خانه هستند.)

| | | |
|---|---|---|
| A | B | C |
| D | E | F, $\circ$ |

ضریب تخفیف برابر ۰/۵ بوده و پاداش زنده‌بودن صفر است. با در نظر گرفتن این موارد به سوال‌های زیر پاسخ دهید:

(آ) سیاست بهینه را برای هر حالت مشخص کنید.

(ب) ارزش بهینه‌ی خانه‌ی A یا همان $V^*(A)$ چه قدر است؟

توجه کنید امتیاز موجود در یک خانه هنگام ورود به خانه‌ی دارای امتیاز محاسبه می‌شود. یعنی به طور مثال در شکل بالا، $\mathrm{R(E, راست, F) = 1}$ می‌باشد.

- حال محیط دیگری مانند شکل زیر را برای بازی در نظر بگیرید:

| | | | |
|---|---|---|---|
| A | B, $\circ$ | | |
| C | D, $*$ | E | F, $*$ |

در این محیط علاوه بر غذای اصلی ($\circ$) دو ماده‌ی غذایی دیگر ($*$) نیز در نقشه وجود دارند که خوردن آن‌ها ۵ امتیاز مثبت خواهد داشت. این امتیازها نیز هنگام ورود به خانه‌ی دارای غذا کسب می‌شوند. در این حالت همچنان غذای اصلی همان $\circ$ است که خوردن آن باعث تمام شدن بازی می‌شود و پاداش ۱ دارد. در این حالت به سوال‌های زیر پاسخ دهید:

(ج) اگر ضریب تخفیف برابر ۱ باشد و هزینه‌ی زنده‌ماندن برابر ۱−، سیاست بهینه را برای حالت‌های مختلف پیدا کنید.

(د) همچنان فرض کنید ضریب تخفیف برابر ۱ است، هزینه‌ی زنده‌ماندن در چه بازه‌ای می‌تواند باشد تا پک‌من با شروع از خانه‌ی $A$ دقیقا یک $*$ را بخورد؟

حل.

• (آ) سیاست بهینه به صورت زیر خواهد بود:

$$\begin{cases} \pi(A) = \text{راست یا پایین} \\ \pi(B) = \text{پایین یا راست} \\ \pi(C) = \text{پایین} \\ \pi(D) = \text{راست} \\ \pi(E) = \text{راست} \end{cases}$$

(ب) دو مسیر بهینه برای رسیدن به غذا از خانه A وجود دارد. یکی از آن‌ها به صورت زیر است:

A -> B -> C -> F

با توجه به این که ضریب تخفیف برابر ۰٫۵ است یعنی اگر در یک گام به غذا (F) برسیم، یک امتیاز، در دو گام ۰٫۵ امتیاز، و در سه گام ۰٫۲۵ امتیاز دریافت می‌کنیم. پس مقدار بهینه خانه‌ی A برابر ۰٫۲۵ خواهد بود.

• (ج) سیاست بهینه به شکل زیر خواهد بود:

| $s$ | $\pi(s)$ |
|---|---|
| A | پایین |
| C | راست |
| D و در F غذا است | راست |
| D و در F غذا نیست | بالا |
| E و در F غذا است | راست |
| E و در F غذا نیست | چپ |
| F | چپ |

مابقی حالت‌ها یا تکراری هستند و یا به آن‌ها نمی‌رسیم. در کل سه‌تایی (مکان، بودن غذا در F، بودن غذا در D) حالت سیستم را مشخص می‌کند.

(د) اگر $x$ هزینه زنده‌ماندن باشد، پاداش نخوردن هیچ کدام از $*$ها برابر $x + ۱$ خواهد بود. پاداش خوردن فقط یکی از $*$ها برابر $۳x + ۶$ خواهد بود. و پاداش خوردن دوتا $*$ برابر $۷x + ۱۱$ است. برای اینکه فقط یک $*$ خورده شود باید داشته باشیم:

$$\left.\begin{array}{ll} ۳x + ۶ > x + ۱ & \text{(خوردن یکی صرفه داشته باشد)} \\ ۳x + ۶ > ۷x + ۱۱ & \text{(خوردن دو تا صرفه نداشته باشد)} \end{array}\right\} \implies -۲٫۵ \leq x \leq -۱٫۲۵$$

۸. (۱۲ نمره) فرض کنید بازی‌ای فقط شامل دو حالت $A$ و $B$ بوده و کنش‌های قابل انجام از هر کدام از این حالت‌ها نیز حرکت به سمت بالا یا پایین است. عاملی در این محیط با استفاده از سیاست $\pi$ به انجام بازی پرداخته و دنباله‌ی حالت‌ها و پاداش‌های زیر را مشاهده کرده است:

| $t$ | $s_t$ | $a_t$ | $s_{t+1}$ | $r_t$ |
|---|---|---|---|---|
| ۰ | A | پایین | B | ۲ |
| ۱ | B | پایین | B | $-۴$ |
| ۲ | B | بالا | B | ۰ |
| ۳ | B | بالا | A | ۳ |
| ۴ | A | بالا | A | $-۱$ |

با درنظر گرفتن ضریب تخفیف برابر ۰٫۵ و $\alpha = ۰٫۵$ به سوال‌های زیر پاسخ دهید:

(آ) می‌خواهیم از روش Q-learning برای بروزرسانی Q-valueها استفاده کنیم. ابتدا رابطه‌ی بروزرسانی را برای Q-valueها بنویسید. سپس با درنظر گرفتن مقدار صفر برای مقدار اولیه‌ی Q-valueها مقدار $Q(A,\text{پایین})$ و $Q(B,\text{بالا})$ را بدست آورید.

(ب) در روش model-based ابتدا تابع گذار $T(s,a,s')$ و تابع پاداش $R(s,a,s')$ را تخمین می‌زنیم. برای این منظور با استفاده از دنباله‌ی حالت‌های مشاهده شده این توابع را به‌ازای مقادیر مختلف $s$، $a$ و $s'$ محاسبه کنید (در مجموع مقادیر هشت حالت را باید تخمین بزنید). اگر داده‌های لازم برای تخمین یکی از ورودی‌های هر کدام از توابع وجود ندارد به این موضوع اشاره کنید.

(ج) فرض کنید تجربه‌ی جدیدی از بازی کسب کرده‌ایم و از روی آن توابع $\hat{T}$ و $\hat{R}$ را تخمین زده‌ایم. نتیجه جدول زیر شده است:

| $s$ | $a$ | $s'$ | $\hat{T}(s,a,s')$ | $\hat{R}(s,a,s')$ |
|---|---|---|---|---|
| $A$ | بالا | $A$ | ۱ | ۱۰ |
| $A$ | پایین | $A$ | ۰/۵ | ۲ |
| $A$ | پایین | $B$ | ۰/۵ | ۲ |
| $B$ | بالا | $A$ | ۱ | $-$۵ |
| $B$ | پایین | $B$ | ۱ | ۸ |

با استفاده از این جدول سیاست بهینه $\hat{\pi}^*(s)$ و تابع ارزش بهینه $\hat{V}^*(s)$ را بدست آورید. (راهنمایی: برای هر $x$ حقیقی که $|x| < ۱$ داریم: $۱ + x + x^۲ + x^۳ + \cdots = \frac{۱}{۱-x}$)

حل.

(آ) رابطه‌ی بروزرسانی Q-valueها به صورت زیر است:

$$Q(s_t,a_t) \leftarrow (۱-\alpha)Q(s_t,a_t) + \alpha\left[r_t + \gamma \max_{a'} Q(s_{t+۱},a')\right]$$

با استفاده از این رابطه:

$$Q(A,\text{پایین}) \leftarrow (۱ - ۰/۵) \times ۰ + ۰/۵\,[۲ + ۰/۵ \times ۰] = ۱$$

$$Q(B,\text{پایین}) \leftarrow (۱ - ۰/۵) \times ۰ + ۰/۵\,[-۴ + ۰/۵ \times ۰] = -۲$$

$$Q(B,\text{بالا}) \leftarrow (۱ - ۰/۵) \times ۰ + ۰/۵\,[۰ + ۰/۵ \times ۰] = ۰$$

$$Q(B,\text{بالا}) \leftarrow (۱ - ۰/۵) \times ۰ + ۰/۵\,[۳ + ۰/۵ \times ۱] = \frac{۷}{۴}$$

$$Q(A,\text{بالا}) \leftarrow (۱ - ۰/۵) \times ۰ + ۰/۵\,[-۱ + ۰/۵ \times ۱] = \frac{-۱}{۴}$$

پس $Q(A,\text{پایین})$ برابر ۱ و $Q(B,\text{بالا})$ برابر ۱/۷۵ است.

(ب) با استفاده از شمارش تعداد حالت‌ها:

$$\begin{cases}\hat{T}(A,\text{بالا},A)=۱\\ \hat{T}(A,\text{بالا},B)=۰\\ \hat{T}(A,\text{پایین},A)=۰\\ \hat{T}(A,\text{پایین},B)=۱\\ \hat{T}(B,\text{بالا},A)=۰/۵\\ \hat{T}(B,\text{بالا},B)=۰/۵\\ \hat{T}(B,\text{پایین},A)=۰\\ \hat{T}(B,\text{پایین},B)=۱\end{cases}\qquad \begin{cases}\hat{R}(A,\text{بالا},A)=-۱\\ \hat{R}(A,\text{بالا},B)=na\\ \hat{R}(A,\text{پایین},A)=na\\ \hat{R}(A,\text{پایین},B)=۲\\ \hat{R}(B,\text{بالا},A)=۳\\ \hat{R}(B,\text{بالا},B)=۰\\ \hat{R}(B,\text{پایین},A)=na\\ \hat{R}(B,\text{پایین},B)=-۴\end{cases}$$

(ج) ابتدا سیاست بهینه را با توجه به مقدار پاداش‌های داده شده پیدا می‌کنیم:

$$\begin{cases}\hat{\pi}^*(A)=\text{بالا}\\ \hat{\pi}^*(B)=\text{پایین}\end{cases}$$

حال می‌توان مقادیر بهینه‌ی $V$ را با توجه به روابط بلمن حساب کرد:

$$\begin{aligned}\hat{V}^*(A)&=۱۰+\lambda*\hat{V}^*(A)\\ &=۱۰+۱۰\lambda+۱۰\lambda^۲+\ldots=\frac{۱۰}{۱-\lambda}=\frac{۱۰}{۰/۵}=۲۰\end{aligned}$$

$$\begin{aligned}\hat{V}^*(B)&=۸+\lambda*\hat{V}^*(B)\\ &=۸+۸\lambda+۸\lambda^۲+\ldots=\frac{۸}{۱-\lambda}=\frac{۸}{۰/۵}=۱۶\end{aligned}$$