به نام خدا

دانشکده مهندسی کامپیوتر

مبانی و کاربردهای هوش مصنوعی ترم پاییز ۱۴۰۰

پاسخ نامه تمرین چهارم

سوال ۱ (۲۵ نمره)

با در نظر داشتن جدول توزیع توأم مقابل، به سوالات زیر پاسخ دهید.

A	B	C	P
+	-	-	0.2
+	-	+	0.05
+	+	-	0.1
+	+	+	0.2
-	-	-	0.25
-	-	+	0.05
-	+	-	0.1
-	+	+	0.05

الف) رابطه میان توزیع توأم[1] و توزیع حاشیه‌ای[2] چیست؟ این موضوع را با حذف دو متغیر تصادفی دلخواه از توزیع توأم جدول نشان دهید.

توزیع توام در واقع یک جدول نشات گرفته شده از توزیع حاشیه ای است. با حذف یک یا چند متغیر تصادفی از یک جدول و جمع احتمالات متناظر با آنها می توان به توزیع حاشیه مد نظر رسید. برای مثال در جدول مقابل با حذف متغیر های B و C خواهیم داشت:

A	P
+	0.55
-	0.45

که یک توزیع حاشیه ای برای متغیر تصادفی A است

[1] Joint distribution
[2] Marginal distribution

ب) احتمالات شرطی[3] زیر را محاسبه کنید.

P(A+|B+) ***P(B-|C+)*** ***P(A+|B-,C+)***

$$P(A+|B+) = \frac{P(A+,B+)}{P(B+)} = \frac{0.3}{0.45} = \frac{2}{3}$$

$$P(B-|C+) = \frac{P(B-,C+)}{P(C+)} = \frac{0.1}{0.35} = \frac{2}{7}$$

$$P(A+|B-,C+) = \frac{P(A+,B-,C+)}{P(B-,C+)} = \frac{0.05}{0.1} = \frac{1}{2}$$

ج) توزیع شرطی[4] زیر را حساب کنید. (با Normalization trick)

P(A|C=+)

سطر های c+ را نگه می داریم و باقی را حذف می کنیم.

A	C	P
+	+	0.25
-	+	0.1

حال با جمع کل حالات C+ و تناسب مقادیر را از ۱ باز می نویسم

A	C	P
+	+	0.71
-	+	0.29

د) با روش inference by enumeration ابتدا هر یک از Query، Evidence و Hidden Variable را مشخص کرده و سپس مقادیر زیر را محاسبه کنید

$$P(A|B+,C-)$$
$$\text{evidence variables: B+, C}-$$
$$\text{query} = \text{A}$$
$$\text{hidden variables} = \text{None}$$

سطر های متناظر با شواهد مد نظر را نگه می داریم

[3] Conditional probability
[4] Conditional distribution

A	B	C	P
+	+	-	0.1
-	+	-	0.1

حال مشابه قسمت قبل متغیر های اضافی را حذف و نرمالایز می کنیم و به احتمال از ۱ می بریم

A	P
+	0.5
-	0.5

$P(A|B+)$

evidence variable: B +

query = A

hidden variable = C

مانند بخش قبل سطر های متناظر با شواهد مد نظر را نگه می داریم

A	B	P
+	+	0.3
-	+	0.15

سپس متغیر های اضافی را حذف و نرمالایز می کنیم و به احتمال از ۱ می بریم

A	P
+	0.67
-	0.33

۵) در حالت کلی، فرق میان احتمال یک رخداد[5] و یک رویداد[6] چیست؟ با یک مثال توضیح دهید.

رخداد یا خروجی یک آزمایش، یک انتساب به همه متغیر ها است که احتمال نهایی مشخصی دارد. اما رویداد یک مجموعه از رخداد های مختلف که احتمال آن برابر جمع احتمال رخدادهای تشکیل دهنده آن است. مثلا در پرتاب یک سکه شیر یا خط آمدن یک رخداد است که احتمال آن ۵۰٪ است. اما مثلا احتمال آمدن عدد زوج در پرتاب یک تاس، رویدادی شامل رخداد های ۲و۴و۶ است.

سوال ۲ (۵ نمره)

بر اساس جدول احتمالات توأم زیر، تعیین کنید که آیا متغیر های X و Y از هم دیگر مستقل هستند یا خیر؟

X	Y	P
+	-	0.2
+	+	0.5
-	-	0.1
-	+	0.2

توزیع حاشیه ای هر کدام از متغیر هارا محاسبه می کنیم:

X	P
+	0.7
-	0.3

Y	P
+	0.7
-	0.3

[5] Outcome
[6] Event

حال دو جدول حاصل را در هم ضرب می کنیم تا جدول توزیع توام اولیه را ایجاد کنیم:

X	Y	P
+	+	0.49
+	-	0.21
-	+	0.21
-	-	0.09

اگر دو متغیر مستقل می بودند، این جدول باید با جدول اول برابر می بود چرا که برای دو متغیر مستقل داریم:

$$P(x, y) = P(x)P(y)$$

سوال ۳ (۵ نمره)

فایده مستقل فرض کردن دو متغیر در چیست؟ به بیان دیگر، دلیل آنکه به دنبال یافتن استقلال (شرطی یا غیر شرطی) متغیر ها هستیم چیست؟

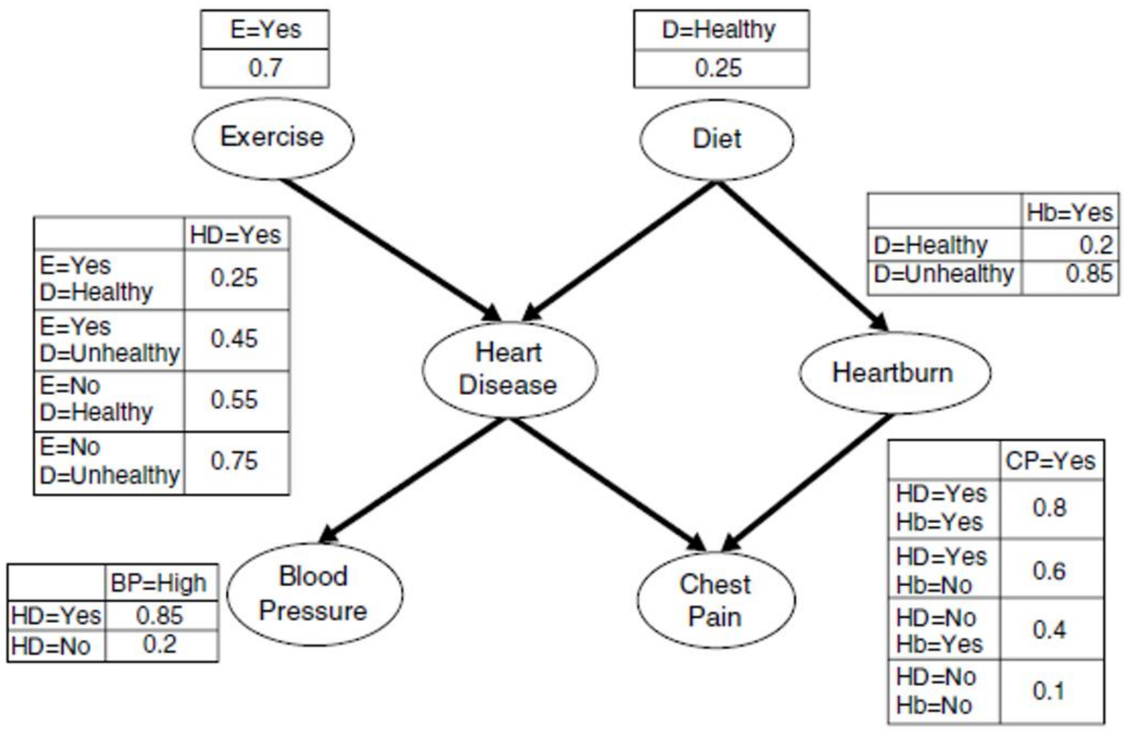
ساده تر کردن مدل سازی مسئله و کاهش محاسبات. اتفاقات دنیا عموما روابط پیچیده ای با هم دارند اما با مستقل کردن آنها تحت شرایط خاصی(و تحمل مقداری خطا) می توانیم مسائل را به نحو بهینه ای مدل سازی کرده و حل کنیم.

هدف استفاده از شبکه های بیزین چیست؟ به بیان دیگر، استفاده از آنها چه تاثیری روی روند محاسبات دارد؟

شبکه های بیزین مدل مختصر و کارایی در اختیار ما قرار می دهند تا ارتباطات پیچیده میان متغیر هارا به صورت مستقل شرطی بنویسیم و بعدا به راحتی ارتباط هر دو متغیر را حساب کنیم. به این صورت از حجم جداول عظیمی که در غیر آن باید ایجاد می کردیم به صورت چشمگیری کاسته میشود و فقط ضروری ترین اطلاعات نگه داشته می شوند و موارد مورد نیاز در لحظه محاسبه می شوند.

سوال ۴ (۵۰ نمره)

شبکه بیزین زیر را در نظر بگیرید و موارد خواسته شده را (با ذکر محاسبات) به دست آورید.



الف) تمام حالاتی که داشتن درد در قفسه سینه (Chest pain) مستقل از نوع رژیم غذایی فرد (Diet) خواهد بود را بنویسید.

بین دو گره CP و D دو مسیر (D > Hb > Cp) و (D > HD > CP) وجود دارد که طبق وجود دو مسیر casual chain با شواهد میانی و به طریق زیر مستقل از هم خواهند بود:

$D \prod CP \mid \{HD, Hb\}$

جواب کاملتر سوال (شامل همه حالات ممکن) گره های دیگر را نیز در نظر می گیرد. مشاهده BP بر روی مشاهده HD موثر است و می‌تواند طرق استقلال جدیدی ایجاد کنند:

$D \prod CP \mid \{E, Bp, Hb\}$ $D \prod CP \mid \{E, Bp, HD, Hb\}$ $D \prod CP \mid \{E, HD, Hb\}$

$D \prod CP \mid \{Bp, HD, Hb\}$ $D \prod CP \mid \{Bp, Hb\}$

ب) بر اساس الگوریتم D-separation برای هرکدام از موارد زیر علت استقلال یا عدم استقلال را ذکر کنید.
(علامت $\prod$ به معنی استقلال است)

$D \prod CP$

در مسیر های D>H>CP و D>HD>CP بدون مشاهده میانی تضمینی برای استقلال نیست

$D \prod CP / HD$

بدون مشاهده میانی Hb در مسیر D›HD›CP تضمینی برای استقلال نیست

$HD \prod Hb$

بدون مشاهده میانی D در مسیر HD›D‹Hb تضمینی برای استقلال نیست

$HD \prod Hb / D, CP$

به خاطر وجود مشاهد میانی CP در مسیر HD› CP› Hb تضمینی برای استقلال نیست

$HD \prod Hb / D$

هر دو مسیر HD›D‹Hb و HD›CP›Hb فعال هستند پس مستقل اند

$E \prod Hb$

دو مسیر کلی غیر فعال داریم. مسیر اول E›HD›D‹Hb و مسیر دوم E›HD›CP›Hb

در اولی دوتا سه‌تایی داریم که E›HD›D در آن غیر فعال است پس کل مسیر غیر فعال است. در دومی سه‌تایی HD›CP›Hb غیر فعال است پس مسیر دوم هم غیر فعال است. لذا مستقل اند

$E \prod Hb / CP$

دو مسیر کلی داریم که یکی از آنها فعال است. مسیر E›HD›CP›Hb که در آن سه‌تایی E›HD›CP و HD›CP›Hb هردو با فرض مشاهده CP فعال اند پس مسیر فعال است و لذا تضمینی برای استقلال نیست

$E \prod D$

دو مسیر کلی داریم شامل E›HD›D که غیر فعال است و E›HD›CP›Hb›D که سه تا سه‌تایی در آن است که دومین سه‌تایی آن یعنی HD›CP›Hb غیر فعال است لذا کل مسیر غیر فعال است. لذا دو متغیر مستقل اند.

$E \prod D / HD$

مسیر E›HD›D با مشاهده HD فعال است و لذا تضمینی برای استقلال نیست

ج) بر اساس اطلاعات موجود، احتمال اینکه یک نفر (دلخواه) بیماری قلبی[7] داشته باشد چقدر است؟

$$\begin{aligned} P(\texttt{HD}=\texttt{Yes}) &= \sum_{\alpha}\sum_{\beta} P(\texttt{HD}=\texttt{Yes}|E=\alpha, D=\beta)P(E=\alpha, D=\beta) \\ &= \sum_{\alpha}\sum_{\beta} P(\texttt{HD}=\texttt{Yes}|E=\alpha, D=\beta)P(E=\alpha)P(D=\beta) \\ &= 0.25\times 0.7\times 0.25 + 0.45\times 0.7\times 0.75 + 0.55\times 0.3\times 0.25 \\ &\quad + 0.75\times 0.3\times 0.75 \\ &= 0.49. \end{aligned}$$

د) اگر یک نفر فشار خون[8] بالا داشته باشد احتمال داشتن بیماری قبلی در او چقدر خواهد بود؟

$$\begin{aligned} P(\texttt{BP}=\texttt{High}) &= \sum_{\gamma} P(\texttt{BP}=\texttt{High}|\texttt{HD}=\gamma)P(\texttt{HD}=\gamma) \\ &= 0.85\times 0.49 + 0.2\times 0.51 = 0.5185. \end{aligned}$$

$$P(\texttt{HD}=\texttt{Yes}|\texttt{BP}=\texttt{High})$$

$$\begin{aligned} &= \frac{P(\texttt{BP}=\texttt{High}|\texttt{HD}=\texttt{Yes})P(\texttt{HD}=\texttt{Yes})}{P(\texttt{BP}=\texttt{High})} \\ &= \frac{0.85\times 0.49}{0.5185} = 0.8033. \end{aligned}$$

ه) اگر بدانیم که بیمار بخش قبل (ب) به صورت روزانه ورزش[9] می‌کند و رژیم سالمی[10] هم داشته دارد، احتمال داشتن بیماری قلبی در او چقدر خواهد بود؟

7

8 Blood pressure

9 Exercise

10 Diet

$$
\begin{aligned}
& P(\mathtt{HD}=\mathtt{Yes}|\mathtt{BP}=\mathtt{High}, D=\mathtt{Healthy}, E=\mathtt{Yes}) \\
= \quad & \left[\frac{P(\mathtt{BP}=\mathtt{High}|\mathtt{HD}=\mathtt{Yes}, D=\mathtt{Healthy}, E=\mathtt{Yes})}{P(\mathtt{BP}=\mathtt{High}|D=\mathtt{Healthy}, E=\mathtt{Yes})}\right] \\
& \qquad \times\ P(\mathtt{HD}=\mathtt{Yes}|D=\mathtt{Healthy}, E=\mathtt{Yes}) \\
= \quad & \frac{P(\mathtt{BP}=\mathtt{High}|\mathtt{HD}=\mathtt{Yes})P(\mathtt{HD}=\mathtt{Yes}|D=\mathtt{Healthy}, E=\mathtt{Yes})}{\sum_\gamma P(\mathtt{BP}=\mathtt{High}|\mathtt{HD}=\gamma)P(\mathtt{HD}=\gamma|D=\mathtt{Healthy}, E=\mathtt{Yes})} \\
= \quad & \frac{0.85\times 0.25}{0.85\times 0.25+0.2\times 0.75} \\
= \quad & 0.5862,
\end{aligned}
$$

و) فرض کنید که تعدادی نمونه از مریض‌ها بر اساس مدل شبکه بیز زیر در اختیار داریم. ابتدا با روش rejection-sampling و سپس با روش likelihood weighting، احتمال اینکه یک نفر ورزش انجام دهد به شرط آنکه هم فشار خون بالا و هم درد قفسه سینه داشته باشد را به دست آورید. برای روش دوم ۵ نمونه ایجاد کنید و برای تعیین متغیر های آزاد، از لیست دوم استفاده کنید. توضیح دهید که در این سوال استفاده از کدام روش موثرتر است؟

E+, HD-, BP-, CP-

E+, HD-, BP+, CP-

E+, HD+, BP+, CP-

E-, HD+, BP+, CP+

E+, HD-, BP-, CP+

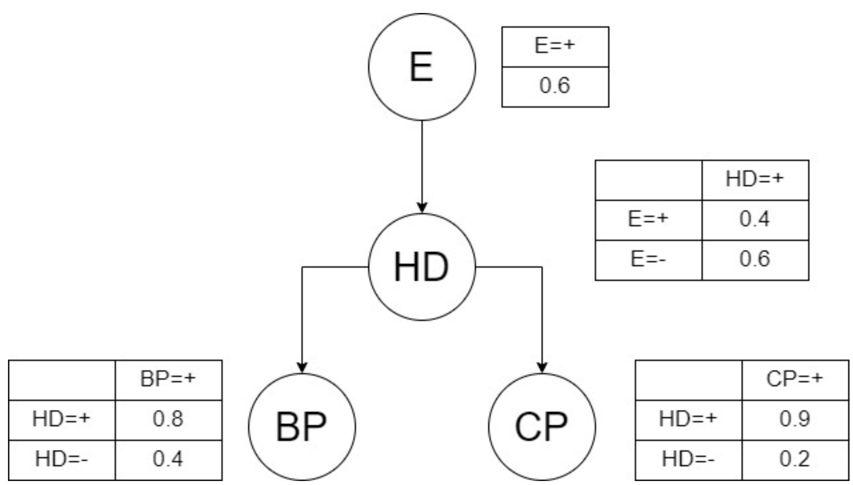
E+, HD-

E+, HD+

E-, HD+

E+, HD-

E-, HD+



Rejection-sampling:

فقط نمونه هایی که با شواهد مد نظر موافق هستند را نگه می داریم. لذا فقط یک نمونه داریم:

E-, HD+, BP+, CP+

در این نمونه هم فرد ورزش انجام نمی داده لذا احتمال مد نظر **صفر** است.

$$P(E+|BP+,CP+) = \frac{0}{1} = 0$$

Likelihood weighting:

دو مقدار ثابت داریم (با وزن ۱) و دو مقدار آزاد (که باید وزن دهی شوند)

$E+, HD-, BP+, CP+ \rightarrow\ weight: 1 \times 1 \times\ 0.4 \times 0.2 = 0.8$

$E+, HD+, BP+, CP+ \rightarrow\ weight: 1 \times 1 \times 0.8 \times 0.9 = 0.72$

$E-, HD+, BP+, CP+ \rightarrow\ weight: 1 \times 1 \times 0.8 \times 0.9 = 0.72$

$E+, HD-, BP+, CP+ \rightarrow\ weight: 1 \times 1 \times 0.4 \times 0.2 = 0.8$

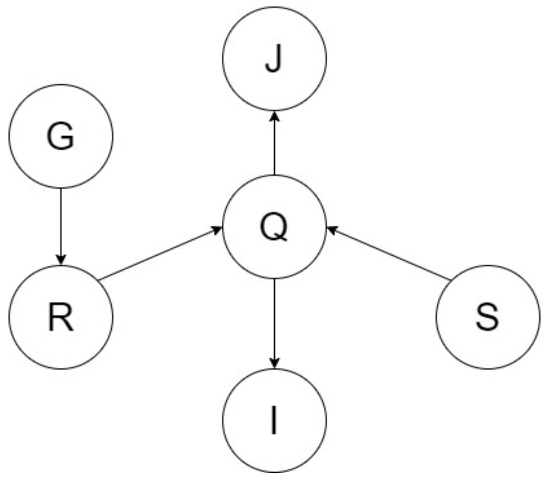
$E-, HD+, BP+, CP+ \rightarrow\ weight: 1 \times 1 \times 0.8 \times 0.9 = 0.72$

$$P(E+|BP+,CP+) = \frac{0.16\ +\ 0.72}{0.16\ +\ 0.72\ +\ 1.44} = \frac{0.88}{2.32}\ =\ 0.37$$

روش دوم (Likelihood weighting) مناسب تر است چرا که در روش اول، تعداد نمونه های کافی نداشتیم و احتمال نا دقیق صفر حاصل شد، با این حال با استفاده از Likelihood weighting توانستیم نمونه هایی با شرایط مد نظر ایجاد کنیم و احتمال واقعی تری به دست آوریم. به طور کلی روش Likelihood weighting در مواردی که شرایط نادر داریم که در نمونه ها زیاد دیده نمیشوند دقت بهتری دارد.

سوال ۵ (۱۵ نمره)

شبکه بیزین زیر را در نظر داشته باشید. با استفاده از روش حذف متغیر[11] مقدار $P(G \mid i, j)$ را محاسبه کنید. (ذکر مراحل محاسبه الزامی است)



مشاهدات J , I هستند و query مد نظر G می باشد. باقی متغیر ها آزاد هستند که باید حذف شوند.

$$\mathrm{P(G, i, j)} = \sum_{s,q,r} P(G,i,j,s,r,q) = \sum_{s,q,r} P(G)P(S)P(r|G)P(j|q)P(i|q)P(q|s,r)$$

$$= \sum_{q,r} P(G)P(r|G)P(j|q)P(i|q) \sum_{s} P(S)P(q|s,r) \ \rightarrow eliminate\ S$$

$$\rightarrow calculate\ factor \rightarrow f_1(q,r) = \sum_{s} P(S)P(q|s,r)$$

$$\sum_{q,r} P(G)P(r|G)P(j|q)P(i|q) \sum_{s} P(S)P(q|s,r)$$

$$= \sum_{q,r} P(G)P(r|G)P(j|q)P(i|q)\ f_1(q,r)$$

$$= \sum_{q} P(G)P(j|q)P(i|q) \ \sum_{r} P(r|G)f_1(q,r) \ \rightarrow eliminate\ R$$

$$\rightarrow calculate\ factor \rightarrow f_2(G,q) = \sum_{r} P(r|G)f_1(q,r)$$

---

[11] Variable elimination

$$\sum_{q} P(G)P(j|q)P(i|q) \sum_{r} P(r|G)f_1(q,r)$$

$$= \sum_{q,r} P(G)P(r|G)P(j|q)P(i|q)\, f_2(G,q)$$

$$= P(G)\sum_{q}(j|q)P(i|q)\, f_2(G,q) \rightarrow eliminate\ Q$$

$$\rightarrow calculate\ factor \rightarrow f_3(G,i,j) = \sum_{q}(j|q)P(i|q)\, f_2(G,q)$$

$$\rightarrow P(G)\sum_{q}(j|q)P(i|q)\, f_2(G,q) = \sum_{g} P(G)f_3(G,i,j)$$

$$\rightarrow P(G,i,j) = P(G)f_3(G,i,j)$$

$$Normalization: P(G|i,j) = \frac{P(G,i,j)}{P(i,j)} = \frac{P(G)f_3(G,i,j)}{\sum_g P(G,i,j)} = \frac{P(G)f_3(G,i,j)}{\sum_g P(G)f_3(G,i,j)}$$