دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

گروه مستقل مهندسی رباتیک

گزارش تمرین اول درس مدل‌های احتمالاتی گرافی

استاد درس:

دکتر نیک آبادی

تدریس‌یار:

مهندس طاهرخانی

نام دانشجو:

نوید خزاعی

۹۲۱۳۵۰۰۸

اردیبهشت ۹۴

فهرست مطالب

۱ بخش نظری

۱.۱ سوال اول

قضیه زیر را اثبات کنید:

Theorem 1. Let $\mathcal{G}$ be a Bayesian Network structure over a set of random variables $\mathcal{X}$ and let P be a joint distribution over $\mathcal{X}$. If P factorizes according to $\mathcal{G}$, then $\mathcal{G}$ is an I-map for P.

پاسخ:

با توجه به این که P روی گراف $\mathcal{G}$ فاکتورایز می‌شود، پس می‌دانیم که:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i)), \tag{۱}$$

که در آن $\mathrm{Pa}(X_i)$ والدین متغیر تصادفی i م هستند. برای آن که نشان‌دهیم گراف $\mathcal{G}$ یک *I-map* برای P است، باید نشان‌دهیم $I(\mathcal{G}) \subseteq I(\mathcal{P})$. با توجه به آن‌چه در درس آمده بود، اگر مجموعه روابط استقلال $I(\mathcal{G})$ را به‌گونه‌ای با استفاده از نتایج فرمول ۱ نشان دهیم، یعنی این روابط از روابط استقلال توزیع مورد نظر استخراج شده‌اند پس زیرمجموعه‌ی آن نیز هستند. برای این کار، با توجه به این که مجموعه روابط استقلال در گراف ایجاب می‌کند که:

$$\{X_i \perp \mathrm{ND}(X_i) \mid \mathrm{Pa}(X_i); i = 1, \ldots, n\} \tag{۲}$$

که در آن $\mathrm{ND}(X_i)$ مجموعه‌ی غیرنسل[1] متغیر (X_i) است. اگر بتوانیم نشان‌دهیم که :

$$P(X_i \mid \mathrm{ND}(X_i)) = P(X_i \mid \mathrm{Pa}(X_i)) \tag{۳}$$

آن‌گاه رابطه‌ی استقلال ۲ برقرار است. برای این‌کار $P(X_i \mid \mathrm{ND}(X_i))$ را محاسبه می‌کنیم (نسل X_i با $\mathrm{D}(X_i)$ نشان داده شده‌است) :

$$\begin{aligned} P(X_i \mid \mathrm{ND}(X_i)) &= \frac{P(X_i, \mathrm{ND}(X_i))}{P(\mathrm{ND}(X_i))} = \frac{\sum_{\mathrm{D}(X_i)} P(X_1, \ldots, X_n)}{\sum_{X_i, \mathrm{D}(X_i)} P(X_1, \ldots, X_n)} \\ &= \frac{\sum_{\mathrm{D}(X_i)} \prod_{j=1}^{n} P(X_j \mid \mathrm{Pa}(X_j))}{\sum_{X_i, \mathrm{D}(X_i)} \prod_{j=1}^{n} P(X_j \mid \mathrm{Pa}(X_j))} \end{aligned} \tag{۴}$$

[1] Non-descendant

در ۴ در محاسبه‌ی سیگما، مواردی هستند که جزو نسل X_i نیستند، پس می‌توان آن‌ها را از سیگمای روی نسل آن خارج نمود. همچنین آن‌چه باقی می‌ماند:

$$\sum_{\mathrm{D}(X_i)} \prod_{X_j \in \mathrm{D}(X_i)} P(X_j \mid \mathrm{Pa}(X_j)) = ۱$$

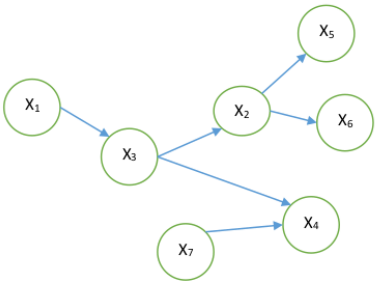
مشابه همین استدلال برای مخرج نیز پاسخ‌گو است، لذا خواهیم داشت:

$$\begin{aligned} &= \frac{\prod_{X_j \in (\mathrm{ND}(X_i) \cup X_i)} P(X_j \mid \mathrm{Pa}(X_j)) \times ۱}{\prod_{X_j \in \mathrm{ND}(X_i)} P(X_j \mid \mathrm{Pa}(X_j)) \times ۱} \\ &= P(X_i \mid \mathrm{Pa}(X_i)). \end{aligned} \tag{۵}$$

پس نشان‌دادیم ۳ و در نتیجه ۲ برقرار است، پس می‌توان نتیجه گرفت که $I(\mathcal{G}) \subseteq I(\mathcal{P})$، و حکم اثبات می‌شود.

۲.۱ سوال دوم

با توجه به شبکه‌ی شکل ۱ به سوالات پاسخ داده شده‌است.



شکل ۱: شکل سوال دوم

۱.۲.۱ توزیع توام

$$P(X_۱, \ldots, X_۷) = P(X_۱)P(X_۳ \mid X_۱)P(X_۲ \mid X_۳)P(X_۵ \mid X_۲)P(X_۶ \mid X_۲)P(X_۴ \mid X_۳, X_۷)P(X_۷)$$

۲.۲.۱ درستی و نادرستی

- $X_۱ \perp X_۵ \mid X_۲$: درست است. اگر $X_۲$ را بدانیم مسیر فعال بین $X_۱$ و $X_۵$ غیر فعال می‌شود و مستقل می‌شوند.

- $X_2 \perp X_7 \mid X_4$: نادرست است. در ساختار وی‌ـشکل بین X_7 و X_3 با دانستن X_4 وابستگی ایجاد می‌شود، و چون X_2 به پدرش X_3 وابسته است، با X_7 نیز در این شرایط وابسته است.

- $X_1 \perp X_5 \mid X_2$: درست است، چرا که دانستن X_3 ، استقلال بین X_2 و X_4 را ایجاد می‌کند و بنابر این از فرزندان X_2 یعنی X_5 نیز مستقل می‌شویم.

Markov Blanket for X_3 ۳.۲.۱

برای این کار ابتدا جهت‌ها را حذف می‌کنیم و سپس فساد[1] ها را حذف می‌کنیم. یعنی نباید گره‌ای باشد که والد یکی از فرزندان همین گره باشد، در صورت وجود به آن گره وصل می‌کنیم تا فساد از بین برود. به تمامی اتصالات موجود نیز وصل می‌کنیم، بنا بر این $MB(X_3) = \{X_1, X_2, X_4, X_7\}$ خواهد بود.

۲ بخش پیاده‌سازی

۱.۲ مقدمه

در این بخش باید با استفاده از پایگاه داده‌ی معرفی شده، چند مدل بیزین ارایه کنیم تا قادر به تشخیص بیماری قلبی باشد.

برای پیاده‌سازی این بخش، تصمیم به توسعه‌ی ابزار جدیدی برای کار با مدل‌های احتمالاتی گرافیکی گرفتیم، که کاملا با ابزارهای تحت خط فرمان سیستم‌عامل لینوکس و پوسته‌ی Bash کار می‌کند. این‌گونه ابزارها، معمولا برای استفاده در مواردی که فایل‌های حجیمی نیاز به پردازش دارند و می‌خواهیم درگیر مسایل مدیریت حافظه و Caching نشویم کاربرد دارند. همچنین، دید مبتنی بر فایل در بسیاری موارد به کمک توسعه‌دهنده می‌آید و توسعه‌ی ابزارها تحت قالب Shell Script، انعطاف بسیار زیادی را برای برنامه‌نویس فراهم می‌کند. این گونه ابزارها معمولا در فازهای یادگیری به دلیل نیاز به نوشتن مفرط در فایل کند هستند، به‌گونه‌ای که ابزار تولید شده در این تمرین عملا کارایی برای آموزش روی داده‌ها ندارد که در تمرینات آتی به رفع این ایرادها می‌پردازیم. در عوض، از آن‌جا که

[1] Immorality

به کل مساله‌ی آموزش و تست به دید مساله‌ی پردازش متن نگاه کرده‌ایم، در فاز تست سرعت پاسخ‌گویی بسیار بالا خواهد بود چرا که ابزارهای پروژه‌ی GNU مانند grep که برای جست‌وجو استفاده می‌شوند و یا awk، sed و cut که به دفعات بسیاری از آن‌ها استفاده نموده‌ایم، کارایی به شدت بالایی در پردازش فایل‌های عظیم دارند. به این ترتیب در پردازش پایگاه‌های داده‌ی بیش از حد بزرگ و زمانی که CPD ها بزرگ می‌شوند و نیاز به جست‌وجوی موثر داریم، کارا خواهند بود. همچنین انعطاف در تعیین نوع داده‌ای ویژگی‌ها و متغیرها از جمله مزیات دیگر این ابزار است و ورودی و خروجی‌های تولید شده به آسانی با دیگر ابزارهای تحت خط فرمان لینوکس قابل استفاده هستند.

در ادامه پس از توضیح چگونگی استفاده از پایگاه داده، به بررسی ابزار توسعه داده‌شده و پاسخ سوالات تمرین می‌پردازیم. ☹

۲.۲ پایگاه داده و پردازش ها

با دریافت پایگاه داده‌ی شهر کلیولوند[۱]

توضیحات	مقدار گسسته	ویژگی/نام متغیر	شماره ستون پایگاه
نشانه پایین دهه	۲،۳،۴،۵،۶،۷	سن	۱
	۰و۱	جنسیت/	۲

[۱] Cleveland