دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی
پاییز ۱۳۹۹
استاد: محمدحسین رهبان

تمرین پنجم، بخش اول — **مدل مارکوف و HMM** — مهلت ارسال: ۱۱ آذر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین (به استثنای هفته‌ی امتحان میانترم) تا سقف پنج روز و در مجموع ۱۵ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهندبود.
- 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 منبع خارج از کتاب و اسلایدهای درس، نام هم‌فکران و آدرس منابع مورد استفاده‌برای حل سوال مورد نظر را ذکرکنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.

سوالاتنظری (۵۰ + ۲۵ نمره)

۱. (۳۵ نمره) ماشینی داریم که رشته‌ایی به شکل خاص تولید میکند. این ماشین در هر مرحله یکی از دو حرف X و یا Y را بیرون میدهد و ساختار کلی رشته‌هایی که ایجاد میکند به شکل $X^{k1}Y^{k_2}X^{k_3}...$ میباشد، که k_i نشان‌دهنده‌ی طول رشته از جنس X یا Y است و k_i با احتمال برابر از مجموعه‌ی $\{1,2,3\}$ انتخاب میشود.
از طرفی برای یک پدیده میتوان چندین مدل مارکوف معادل تعریف کرد و از میان این مدل‌ها به مدلی که کمترین state را دارد مدل مارکوف مینیمال گفته میشود.

- (۲۵ نمره امتیازی) باتوجه به تعریف مینیمال بودن مدل مارکوف و طرز کارکرد ماشین توصیف شده، سعی کنید کارکرد ماشین را با یک HMM مدل کنید به طوری که مدل مارکوف بین state ها مینیمال باشد. سپس نمودار حالت بین state ها را بکشید و ماتریس transition و emission را مشخص کنید.
- (۱۰ نمره) حاال فرض کنید رشته‌ی خروجی دارای ساختار $X^{k_1}Y^{4-k_1}X^{k_2}Y^{4-k_2}...$ است. نمودار و ماتریس‌های خواسته شده برای قسمت قبل را برای این حالت بنویسید. لزومی به برقراری مینیمال بودن نیست.

۲. (۴۰ نمره) شما یک نمونه از Wall-E دارید که در محل جمع آوری زباله های دانشگاه مشغول فعالیت است. شما می خواهید محل قرارگیری Wall-E را با استفاده از HMM بدست آورید. محل جمع‌آوری زباله به شکل یک جدول $n \times n$ است و در هر گام زمانی $t = 1, 2, 3, ...$ ، Wall-E در خانه‌ی $X_t \in \{1, 2, ...n\}^2$ قرار دارد. نحوه‌ی حرکت Wall-E در هر گام به شکل زیر است:

- با احتمال $1 - \epsilon$ به صورت رندوم و با احتمال برابر به یکی از خانه‌های مجاور (حداکثر ۴ تا) می‌رود.
- به احتمال ϵ پرواز میکند و به یکی از خانه‌های جدول میرود. در این حالت می‌تواند به خانه‌ای که در آن قرار دارد هم برود.

همچنین سنسورهای Wall-E تنها میتوانند سطری که Wall-E در آن قرار دارد را به شما مخابره کنند.
فرض کنید که $n = 10, \epsilon = 0.5$
توجه کنید که تنها به E_t دسترسی دارید که همان سطری است که Wall-E در آن قرار دارد. و $E_1 = 1$ است.

- (۲۰ نمره) اگر $E_2 = 1$ باشد، احتمال اینکه در گام زمانی دوم Wall-E در خانه‌ی $X_2 = (2, 1)$ قرار داشته باشد را حساب کنید.
- (۲۰ نمره) اگر $E_2 = 4$ باشد، احتمال ایکه در گام زمانی دوم Wall-E در خانه‌ی $X_2 = (4, 4)$ قرار داشته باشد را حساب کنید.

سوالات عملی (۵۰ + ۲۵ نمره)

۱. (۵۰ نمره) سوال زیر را بخوانید و دو کد خواسته شده را به همراه جواب قسمت‌های نظری خواسته شده در سوال در یک فایل zip آپلود کنید.
یکی از موارد کاربر HMM در منطبق‌سازی رشته‌های DNA با الگوهای خاص است. یک مثال ساده از این کاربرد را بررسی میکنیم. همانطور که میدانید رشته‌های DNA از ۴ حرف $\{A, C, G, T\}$ تشکیل شده‌اند و رشته‌های خاصی میتوانند نشانگر ویژگی‌های خاصی باشند. گاهی در این رشته‌ها چندین رشته نشانگر یک خاصیت‌اند. به‌طور مثال رشته‌های ACAATG و AGAATC و ACACAGC و ACCGATC و TCAAT-GATC نشانگر یک خاصیت میباشند. این خاصیت را خاصیت X مینامیم. دانشمندان برای نشان دادن ترتیب حروف در این خاصیت از مدل زیر که ۵ رشته‌ی بالا را در خود جای داده‌است استفاده میکنند:

l_1	l_2	l_3	l_i	l_i	l_i	l_4	l_5	l_6
A	C	A	-	-	-	A	T	G
A	G	A	-	-	-	A	T	C
A	C	A	C	-	-	A	G	C
A	C	C	G	-	-	A	T	C
T	C	A	A	T	G	A	T	C

این مدل با این دید طراحی شده است که خاصیت X احتمالا به این شکل نمایان میشود که اگر ACAATG را حالت پایه در نظر بگیریم، ابتدا ۳ حرف اصلی از بخش اول این رشته با اندکی تغییر ظاهر میشود (مثلا همانطور که در جدول میبینید به‌جای ACA در یکی از حالت‌ها حروف ACC در ابتدا آمده است که یعنی حرف سوم با حالت پایه متفاوت شده است.) و سپس تعداد دلخواهی حروف دلخواه می‌آید (که در مدل با l_i مشخص شده است) و درنهایت ۳ حرف اصلی بخش دوم (باز هم با احتمال اندکی تغییر در ۳ حرف پایانی حالت پایه) ظاهر میشوند. حال که ۵ نمونه الگو از خاصیت X داریم، میخواهیم درصورت داشتن یک رشته‌ی s متشکل ازحروف A، C، G، T و با طول دلخواه k $(k > 0)$ مشخص کنیم که چقدر محتمل است این رشته با تمام یا پسوندی از الگویی مربوط به خاصیت X شروع شده باشد. برای ساده‌سازی محاسبات از مدل HMM استفاده میکنیم.

- (۱۵ نمره) الگوی خاصیت X را با یک HMM مدل کنید و ماتریس transition و emission آن را مشخص کنید. (راهنمایی: state ها و observation ها طبق جدول داده شده و دقیقا با همان نمادها مدل کنید و فقط یک حالت نهایی st به آن اضافه کنید که درصورت ورود به آن حالت، الگو تمام شده است و از آن جا به بعد حروف به صورت کاملا تصادفی و با احتمال برابر ادامه پیدا میکنند. یه این معنی که $P(st_{t+1}|st_t) = 1$)
- (۱۵ نمره) حال برای محاسبه‌ی کمیت مورد نظرمان یعنی میزان محتمل بودن اینکه رشته‌ی s با تمام یا پسوندی از الگوی خاصیت X شروع شده باشد میخواهیم از likelihood استفاده کنیم. چیزی که ما میخواهیم اندازه‌گیری کنیم درواقع $P(H|Y_1, Y_2, .., Y_k)$ است که Y_i ها همان observation ها میباشند و H مدل مارکوف‌ای است که شما در بخش قبل طراحی کرده‌اید. ولی به دلیل اینکه نمیتوانیم این احتمال را حساب کنیم از likelihood یعنی $P(Y_1, Y_2, ..., Y_k|H)$ استفاده میکنیم. پس لازم است فرض کنیم observation ها از مدل مارکوف طراحی شده می‌آیند و طبق آن $P(Y_1, Y_2, ..., Y_k)$ را محاسبه کنیم. ابتدا این احتمال را طبق آن چه درباره‌ی HMM آموخته‌اید ساده کنید و سپس برنامه‌ای بنویسید که این کمیت را برای یک رشته‌ی دلخواه ورودی حساب کند.

ـ ورودی: رشته‌ی s مثلا GATC
ـ خروجی: مقدار احتمال $P(Y_1, Y_2, ..., Y_k|H)$

- (۲۵ نمره امتیازی) برای این‌که عدد بدست آمده در قسمت قبل، قابل فهم باشد معمولا از likelihood نسبی استفاده میشود که score نامیده میشود و به شکل زیر تعریف میشود:

$$score = \frac{P(Y_1, ..., Y_k|H)}{(\frac{1}{4})^k}$$

که مخرج واضحا likelihood این است که رشته کاملا تصادفی باشد. حد پایین score را محاسبه کنید و استدلال خود را بیان کنید. سپس در کد قسمت قبل score را به عنوان خروجی دوم در خط بعد خروجی دهید.

- (۲۰ نمره) احتمالا اگر دو قسمت قبل را انجام داده باشید متوجه شده‌اید که روشی که استفاده کردیم برای رشته‌های طولانی میتواند خیلی زمانبر باشد. در نتیجه روش مورد استفاده بهینه نیست. برای حل این مشکل نگاه خود را به مسیله کمی تغییر میدهیم و از الگوریتم Viterbi استفاده میکنیم. با توجه به آن‌چه از این الگوریتم در درس معرفی شده است و با در نظر گرفتن مقادیر محاسبه شده در این الگوریتم استدلال کنید که این مقادیر چگونه میتوانند در حل مسیله‌ی پیش‌رو به ما کمک کنند تا یک معیار مناسب برای اندازه‌گیری میزان محتمل بودن شروع رشته‌ی s با پسوندی از خاصیت X بدست آوریم. توجه کنید که در این قسمت قرار نیست دقیقا معیار تعریف شده در موارد قبلی را بدست بیاورید بلکه به دنبال تعریف معیاری کمی هستید که با استفاده از آن بتوانید میزان محتمل بودن شروع یک رشته با پسوندی از خاصیت مورد نظر را در رشته‌های مختلف مقایسه کنید. سپس الگوریتم را پیاده کرده و کمیت معیار خود را خروجی دهید.
ـ ورودی: رشته‌ی s مثلا GATC
ـ خروجی: مقدار معیار ای که تعریف کرده‌اید.