دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

بهار ۱۴۰۱

استاد: محمدحسین رهبان
گردآورندگان: محمدمهدی اصمع - کیان باختری - علی عباسی

تمرین پنجم — Temporal Probability Models و مقدمه‌ای بر یادگیری ماشین — مهلت ارسال: ۱۰ خرداد

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۷ روز و در مجموع ۲۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر 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 نظری (۸۰ نمره)

۱. (۱۵ نمره)

(آ) فرض کنید که X یک متغیر تصادفی باشد که فقط دو مقدار ۰ و ۱ را بپذیرد. اگر p احتمال رخ دادن ۰ باشد و $H(X)$ را به صورت $h(p)$ بنویسسم، ثابت کنید $h(p)$ یک تابع مقعر است. آیا این گزاره برای متغیر تصادفی با n مقدار نیز درست است؟ برای جواب خود دلیل بیاورید.

(ب) با استفاده از گزاره بخش قبل، بیشینه $H(X)$ را برای یک متغیر تصادفی با n مقدار را به دست آورید.

(ج) آیا در حالت کلی هم می‌توان برای $H(X)$ حد بالا مشخص کرد؟

۲. (۱۵ نمره) یک ربات در صفحه‌ای ۵×۵ در حال حرکت است و یک سنسور برای ارسال موقعیت آن نصب شده‌است که مجموع عدد سطر و ستون موقعیت ربات را ارسال می‌کند. در هر حرکت این ربات به صورت یکنواخت با احتمال $۱ - \epsilon$ به یکی از همسایه‌هایش (خانه‌هایی که ضلع مشترک دارند) و با احتمال ϵ به یک خانه در سطر یا ستون خودش می‌رود (ربات در هر مرحله حتماً حرکت خواهد کرد). مقادیر احتمال زیر را محاسبه کنید.

(آ) $\mathbb{P}(x_۱ = ۳ | x_۲ + y_۲ = ۱۰)$

(ب) $\mathbb{P}(x_۱ + y_۱ = ۳ | x_۲ + y_۲ = ۷)$

۳. (۱۵ نمره) امید ریاضی تعداد دفعات پرتاب یک سکه تا برای اولین بار دنباله HTH را ببینیم، چند است؟

۴. (۲۰ نمره)

فرض کنید داده‌های زیر به شما داده شده‌است. هدف این است که مشخص کنیم فرد در آن روز تنیس بازی می‌کند یا خیر. با توجه به داده‌ها، به سوالات زیر پاسخ دهید:

(آ) درخت تصمیم را برای این داده‌ها ترسیم کنید.

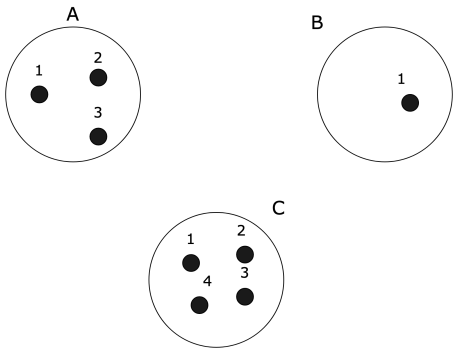
(ب) دقت درخت تصمیم برای این داده‌ها چقدر است؟

Outlook	Temperature	Humidity	Wind	Play tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

(ج) با استفاده از دو روش Naive Bayes و درخت تصمیم، مشخص کنید که این فرد در شرایط زیر تنیس بازی می‌کند یا خیر؟

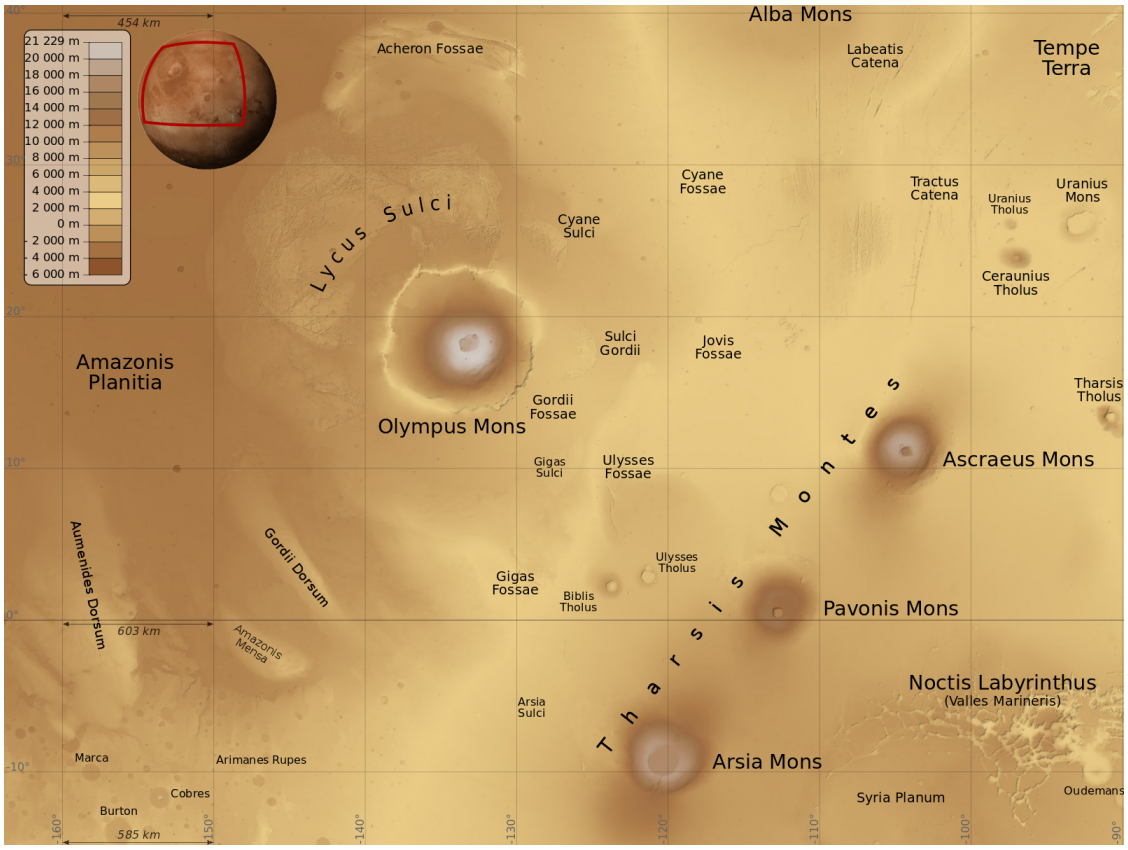
Outlook	Temperature	Humidity	Wind	Play tennis
Sunny	Cool	Normal	Strong	?
Rain	Hot	High	Weak	?

۵. (۱۵ نمره) ۳ شهر A، B و C به ترتیب ۳، ۱ و ۴ فرودگاه دارند. یک خلبان با شروع از A۱ در هر مرحله به صورت تصادفی به یک فرودگاه دیگر می‌رود یا در همان فرودگاه می‌ماند ولی به دلیل مشکل ایجاد شده در هواپیما، مرکز کنترل فقط از شماره فرودگاه مقصد اطلاع دارد. شرایط این هواپیما را به صورت یک HMM مدل کنید و نمودار حالت‌ها و ماتریس emission و transition را مشخص کنید.



سوالات عملی (۸۰ نمره)

۱. (۴۰ نمره) عکس زیر بخشی از مریخ را نشان می‌دهد. ناسا مریخ‌نوردی به فضا فرستاده است که طبق برنامه‌ریزی‌ها باید در نقطه‌ی مشخصی از این تصویر فرود آید. به علت خطاهای پیش‌بینی نشده، مریخ‌نورد در موقعیت مورد نظر فرود نمی‌آید و گم می‌شود. اما مهندسان می‌دانند که مریخ‌نورد جایی نامعلوم در همین منطقه به مریخ نشسته است. مریخ‌نورد مجهز به یک فرستنده‌ی رادیویی است که می‌تواند امواجی با طول موج بلند ارسال کند و بازتاب آن را از عوارض طبیعی اطراف (مانند کوه‌ها) دریافت کند. همان طور که در تصویر مشاهده می‌کنید، چهار کوه در این منطقه وجود دارد و مریخ‌نورد می‌تواند با ارسال سیگنال رادیویی، فاصله‌اش را با هر کدام از آن‌ها تخمین بزند.



شکل ۱: منطقه‌ی فرود مریخ‌نورد

مهندسان برای این که بتوانند موقعیت دقیق مریخ‌نورد را پیدا کنند، از particle filtering استفاده می‌کنند. آن‌ها

به مریخ‌نورد دستور می‌دهند که طی ۲۰ مرحله، ۲۰ گام بردارد (نقشه را دو بعدی و تخت فرض کنید). هر گام به اندازه‌ی δx در جهت مثبت محور x و به اندازه‌ی δy در جهت مثبت محور y است. مقادیر δx و δy به ترتیب از توزیع‌های گوسی با میانگین‌های ۲ و ۱ و واریانس ۱ می‌آیند. در هر مرحله (گام) مریخ‌نورد فاصله‌اش را با هرکدام از چهار کوه حاضر در نقشه (Olympus, Arsia, Pavonis, Ascraeus) تخمین می‌زند. این تخمین‌ها حاوی یک نویز گوسی با میانگین ۲ و واریانس ۱ هستند.

ورودی: در خط اول عبارت oly می‌آید و در ۲۰ خط بعدی فاصله‌ای که مریخ‌نورد در هر گام از کوه Olympus اندازه گرفته است می‌آید. سپس، عبارت ars می‌آید و در ۲۰ خط بعدی فواصلی که مریخ‌نورد در هر گام از کوه Arsia اندازه گرفته است می‌آید. در ادامه هم به همین ترتیب فواصل اندازه‌گیری شده از دو کوه دیگر می‌آیند (نمونه‌ی ورودی را ببینید).

خروجی: با استفاده از particle filtering مکان مریخ را در گام آخر تخمین بزنید. حداقل از ۱۰۰۰ دانه (particle) استفاده کنید و در هر مرحله هنگام resample کردن، نصف دانه‌ها که وزن کم‌تری دارند را کنار گذاشته و به همان تعداد، دانه‌ی جدید در اطراف دانه‌های حذف نشده تولید کنید. در نهایت وقتی به گام آخر رسیدید، میان مختصات دانه‌های باقی مانده میانگین وزن‌دار بگیرید. اکنون یک x و یک y به عنوان تخمین نهایی خود از مختصات مریخ‌نورد در گام آخر در اختیار دارید. این دو را با دقت یکان به بالا گرد کرده و به ترتیب در دو خط چاپ کنید. برای گرد کردن می‌توانید از این کد استفاده کنید:

$$x = int(np.ceil(x/۱۰) * ۱۰) \quad (۱)$$
$$y = int(np.ceil(y/۱۰) * ۱۰) \quad (۲)$$

مختصات کوه‌ها و نمونه‌ی ورودی و خروجی در این لینک قرار داده شده است. حتما پیش از شروع کد زدن، نمونه‌ها را ببینید تا با فرمت ورودی و خروجی آشنا شوید.

۲. (۴۰ نمره) در این سوال می‌خواهیم مدلی بر مبنای درخت تصمیم بسازیم که خوش‌خیمی یا بدخیمی توده‌ی سرطانی را بر اساس ویژگی‌های داده شده تشخیص بدهد.
ساختار کلی کلاس درخت تصمیم، توابع مورد نیاز آن و کارکرد هر کدام در فایل decision_tree.py مشخص شده است و تنها نیاز است که شما توابع آن را پیاده‌سازی کنید.
داده‌های آموزش به همراه برچسب [۱] آن‌ها (ستون target) در فایل breast_cancer.csv و داده‌های تست بدون بر چسب، در فایل test.csv قرار دارند که باید پس از آموزش مدلتان، برچسب‌هایی که برای این مجموعه داده پیش‌بینی می‌کند را در فایلی به نام output.csv ذخیره کنید.

- ساختار کلاس‌های داده شده، یک روش پیاده‌سازی پیشنهادی است و شما می‌توانید به دلخواه خود توابع یا کل روش پیاده‌سازی را تغییر دهید. اما لازم است که روش شما بر مبنای درخت تصمیم و بدون استفاده از کدهای آماده باشد.
- استفاده از کتاب‌خانه‌ی sklearn جز برای جدا کردن مجموعه داده‌ی آموزش و اعتبارسنجی [۲] مجاز نیست.
- برخلاف مثال‌هایی که در کلاس دیدید، مقادیر داده‌های ما حقیقی و پیوسته است؛ بنابراین split کردن آن‌ها بر اساس آستانه‌ای [۳] از یکی از فیچرها انجام می‌شود؛ به این صورت که داده‌هایی که این فیچرشان مقدار کمتر یا مساوی مقدار آستانه دارد در یک گروه، و دیگر داده‌ها در گروه دیگر قرار می‌گیرند. بنابراین شما باید در هر مرحله به کمک information gain بهترین ویژگی و بهترین آستانه‌ی آن را برای split کردن بیابید.
- max_depth و min_samples_split هایپرپارامترهای مدل شما هستند که نیاز است به کمک داده‌های اعتبارسنجی، آن‌ها را تنظیم کنید. در صورتی که همچنان کارکرد آن‌ها برایتان مورد سوال است، به مستندات DecisionTreeClassifier کتاب‌خانه‌ی sklearn مراجعه کنید.

---

[۱] Label
[۲] Validation
[۳] Threshold

- به تمام توابع و ماژول‌هایی که seed یا random_state می‌گیرند، seed بدهید و پس از پایان تمرین توجه کنید که با اجرای چندباره‌ی برنامه، فایل خروجی تغییری نکند. پس از تحویل تمرین، کد شما اجرا خواهد شد و در صورت مغایرت فایل خروجی کد و فایل بارگذاری شده، نمره‌ای به تمرینتان تعلق نخواهد گرفت.
- کد شما باید یک فایل csv با نام output.csv تولید کند که شامل تنها یک ستون است که در سطر اول آن نام ستون (target) و در سطرهای بعدی برچسب داده‌های تست قرار دارند.
- دو فایل decision_tree.py و output.csv را در یک فایل زیپ به فرمت DT-stdnum.zip آپلود کنید.