دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی

پاییز ۱۴۰۰

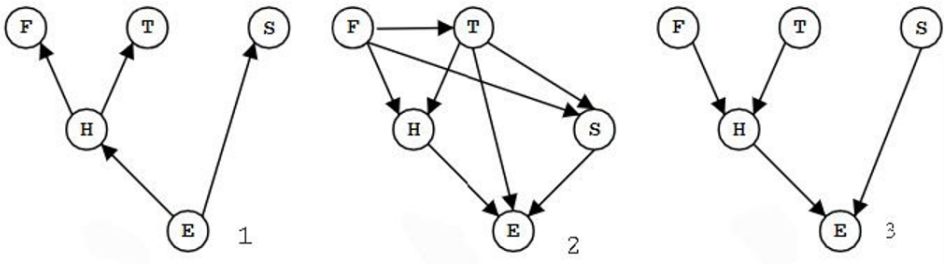
استاد: محمدحسین رهبان
گردآورندگان: پانیذ حلواچی، سایه جارالهی، امیرحسین باقری

تمرین چهارم | شبکه‌های بیزین | مهلت ارسال: ۲۲ اردیبهشت

- 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 (۷۰ نمره)

۱. (۱۰ نمره) می‌خواهیم عملکرد دانشجویان یک کلاس را در امتحان پیشبینی کنیم. می‌دانیم دانشجویانی عملکرد خوبی در امتحان دارند که به خوبی برایش مطالعه کرده باشند و همینطور در زمان امتحان سردرد نداشته باشند. از طرفی در نظر می‌گیریم سردرد ناشی از سینوزیت یا خستگی است. فرض کنید مطالعه، سینوزیت و خستگی دو به دو مستقل‌اند.



در نظر بگیرید که متغیر F نشان‌دهنده‌ی سینوزیت داشتن ، T نشاندهنده‌ی خسته بودن، H نشان‌دهنده‌ی سردرد داشتن، S نشان‌دهنده‌ی مطالعه و درنهایت E نشان‌دهنده‌ی امتحان را خوب دادن باشند. (متغیرهای معرفی شده باینری هستند.)

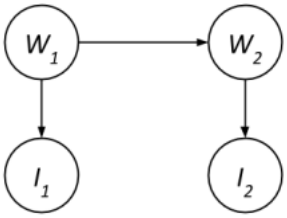
(آ) به نظر شما کدام یک از شبکه‌های بیزین زیر به خوبی مسئله شرح داده شده را توصیف می‌کند؟ دلیل رد هر کدام از شبکه‌های بیزین و آن‌چه در آنان نقض می‌شود را بیان کنید.

(ب) در واقعیت می‌دانیم که سینوزیت و خستگی روی مطالعه تاثیر می‌گذارند. شبکه بیزین مناسب را با درنظر گرفتن این نکته مجددا رسم نمایید.

(ج) میزان حافظه برای ذخیره جداول مربوط به شبکه بیزین قسمت ۱ و ۲ را محاسبه کنید. باتوجه به نتایج از نظر شما کدام شبکه بیزین را بهتر است در نظر گرفت؟ (میان بخش ۱ و ۲)

(د) بهار سینوزیت دارد. با استفاده از مدل بخش ۱ (مدلی که از میان سه مورد مناسب بود) احتمال آنکه امتحانش را خوب ندهد را براساس احتمال های شرطی و به ساده ترین شکل ممکن بنویسید.

۲. (۱۰ نمره) می‌خواهیم ارتباط میان درس خواندن و آب و هوا را بررسی کنیم! $W_۱$و $W_۲$ را آب و هوا در روز اول و دوم در نظر بگیرید که می‌توانند مقدار S (آفتابی) یا R (بارانی) داشته باشند. $I_۱$ و $I_۲$ را نیز نشان‌دهنده ی درس خواندن در روز اول و دوم در نظر بگیرید که مقادیر T (درس خواندن) و F (درس نخواندن) می‌تواند داشته باشد. می‌توانیم این مسئله را با شبکه بیزین و احتمالات زیر مدل کنیم.



W_1	$P(W_1)$
S	0.6
R	0.4

W_1	W_2	$P(W_2\|W_1)$
S	S	0.7
S	R	0.3
R	S	0.5
R	R	0.5

W	I	$P(I\|W)$
S	T	0.9
S	F	0.1
R	T	0.2
R	F	0.8

فرض کنید نمونه‌های زیر را از مدل تولید کرده‌ایم.$(W_۱, I_۱, W_۲, I_۲)$

(R,F,R,F),(R,F,R,F),(S,F,S,T),(S,T,S,T),(S,T,R,F),
(R,F,R,T),(S,T,S,T),(S,T,S,T),(S,T,R,F),(R,F,S,T)

(آ) نمونه‌هایی را که هنگام محاسبه $P(W_۲|I_۱ = T, I_۲ = F)$ در sampling rejection پذیرفته نمی‌شوند را خط بزنید.

(ب) حال روش Likelihood weighting را در نظر بگیرید. شش نمونه زیر را با دانستن $I_۱ = T$ و $I_۲ = F$ تولید کرده‌ایم.$(W_۱, I_۱, W_۲, I_۲)$

(S,T,R,F),(R,T,R,F),(S,T,R,F),(S,T,S,F),(S,T,S,F),(R,T,S,F)

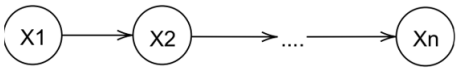
وزن نمونه‌ی (S،T،R،F) در Likelihood weighting چه قدر است؟

(ج) مقدار $P(W_۲|I_۱ = T, I_۲ = F)$ را با استفاده از Likelihood weighting تخمین بزنید.

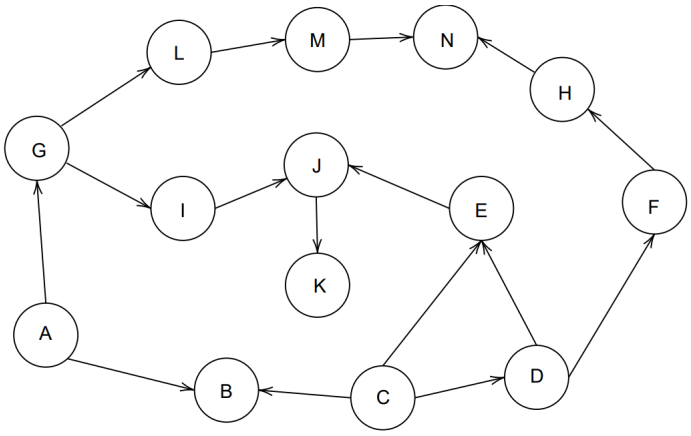
۳. (۱۰ نمره) پیچیدگی زمانی محاسبه‌ی احتمال $(P(X_n))$ در زنجیره‌ای از متغیرهای باینری را در دو حالت زیر به دست آورید (کافی است تعداد جمع و ضرب‌های لازم را محاسبه کنید):

(آ) انجام enumeration

(ب) انجام variable elimination

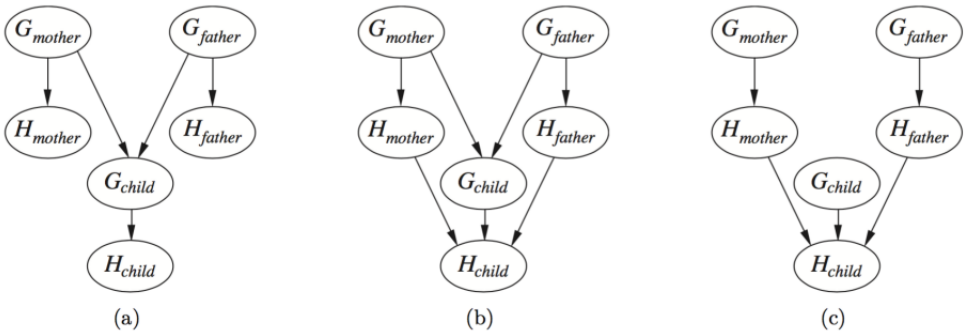


۴. (۱۲ نمره) با توجه به شبکه‌ی بیزین زیر درستی یا نادرستی عبارات زیر را با ذکر دلیل مشخص کنید.



- $I \perp E | B, N$
- $I \perp E | B, N, G$
- $I \perp L | K, N, G$
- $I \perp D | B, J, E$

۵. (۱۹ نمره) فرض کنید H_x یک متغیر تصادفی است که نشان‌دهنده‌ی راست دست بودن یا چپ دست بودن شخص x است که مقادیر left یا right را می‌تواند بگیرد. به منظور تشخیص راست دست یا چپ دست بودن فرد، فرض کنید که یک ژن به نام G_x وجود دارد که مقادیر L یا R را می‌تواند بگیرد و با احتمال s ، راست دست بودن یا چپ دست بودن شخص با ژنی که دارد یکسان می‌باشد. به‌علاوه، خود این ژن با احتمال‌های مساوی از هر یک از والدینش می‌تواند به ارث برسد و با احتمال اندک m نیز ممکن است دچار جهش شود.



(آ) کدام یک از شبکه‌های بیزین بالا با عبارت زیر در تناقض نیست؟

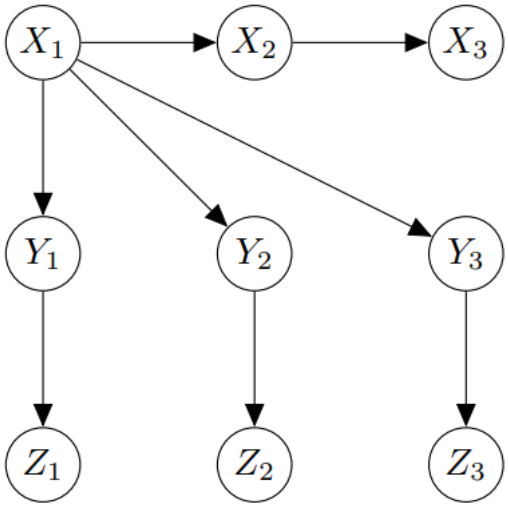
$$P(G_{father}, G_{mother}, G_{child}) = P(G_{father})P(G_{mother})P(G_{child})$$

(ب) کدام یک از شبکه‌های بیزین بالا به بهترین شکل مسئله شرح داده شده را توصیف می‌کند؟

(ج) جدول احتمالات شرطی را برای G_{child} در شبکه بیزین (a) برحسب s و m به دست آورید.

(د) فرض کنید $P(G_{father} = L) = P(G_{mother} = L) = q$ برقرار است. در این‌صورت عبارت $P(G_{child} = L)$ برحسب m و q و با دانستن مقدار گره‌های والدش در شبکه (a) به چه صورت خواهد بود؟

۶. (۹ نمره) شبکه‌ی بیزین زیر را در نظر بگیرید.



(آ) درست یا غلط بودن عبارات زیر را تعیین کنید.

- $P(X_1, X_2, X_3| + y_1) = P(X_1, X_2, X_3| - y_1)$
- $P(Z_3| + x_1, -y_3) = P(Z_3| - x_1, -y_3)$
- $P(Z_3| + x_1, -y_3) = P(Z_3| + x_1, +y_3)$
- $P(Y_1, Y_2, Y_3) = P(Y_1)P(Y_2)P(Y_3)$

(ب) برای یافتن احتمال $P(Z_۱| + x_۳, +z_۲, +z_۳)$:

بهترین ترتیب برای حذف متغیر variable elimination را تعیین کنید.

(منظور از بهترین ترتیب، ترتیبی است که با آن طول بزرگترین فاکتوری که در طول محاسبه به آن برمی‌خوریم نسبت به بقیه ترتیب‌ها کمترین باشد.)

(ج) آیا ترتیب حذف متغیرها تفاوتی در بدترین حالت زمان اجرای worst case running time شبکه بیزین Bayesian network inference ایجاد می‌کند؟

سوالات عملی (۷۰ نمره)

---

۱. (۳۵ نمره) در این تمرین به بررسی استقلال متغیرهای تصادفی و محاسبه احتمال درست بودن برخی از رئوس به شرط مشاهدات در یک شبکه بیز می‌پردازیم. الگوریتم d-separation یکی از الگوریتم‌هایی است که به دنبال مسیر فعال بین دو راس در شبکه بیز می‌گردد و درصورتی که حداقل یک مسیر فعال پیدا نشود، دو متغیر را مستقل از هم در نظر می‌گیرد. برای مشخص کردن استقلال بین دو متغیر این الگوریتم را پیاده‌سازی کنید. همچنین با استفاده از elimination variable احتمال درست بودن رئوس خواسته شده را مطابق cpt های داده شده به‌دست آورید.

به عنوان ورودی، شبکه بیز با استفاده از رئوس و یال‌هایش داده می‌شود. برای آسانی، هر یک از رئوس با یک عدد نمایش داده می‌شود که به ترتیب از یک شروع می‌شود. مقادیر مجاز برای هر یک از متغیرهای تصادفی از مجموعه $\{True, False\}$ است. همچنین رئوس evidence نیز همراه مقادیرشان به شما داده می‌شود. در انتها دو راس داده می‌شود که استقلال آن‌ها باید به کمک الگوریتم d-separation بررسی شود. همچنین برای هریک از دو راس، احتمال درست بودن آن به شرط evidence ها محاسبه و در خروجی اعلام شود.

فرمت ورودی برنامه:

در ابتدا عدد n که همان تعداد رئوس شبکه بیز است، داده می شود. رئوس شبکه به صورت $۱, ۲, ..., n$ خواهد بود.

$$۱ \leq n \leq ۲۰$$

در $2n$ خط بعدی ورودی، هر یک از دو خط متوالی به ترتیب مربوط به یکی از رئوس است. (دو خط اول مربوط به راس شماره ۱، دو خط دوم مربوط به راس شماره ۲ و ...) در خط اول پدران آن راس با یک اسپیس جدا شده اند.(درصورتی‌که هیچ پدری نداشته باشد این خط خالی است) در خط دوم cpt مربوط به آن راس قرار گرفته است.

در خط بعدی ورودی تمامی مشاهدات پیشین به‌صورت زیر داده می‌شود.

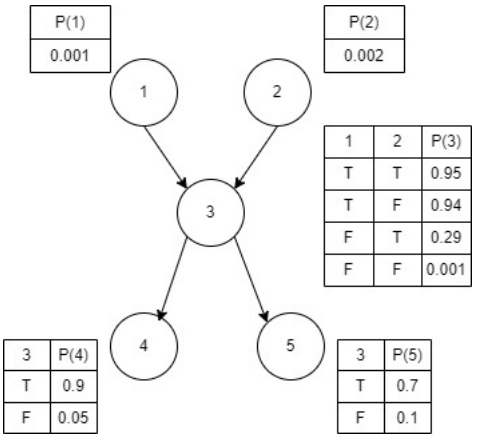
$$x_1 -> 0, x2 -> 1, x3 -> 0$$

و x_i ها شماره راس‌ها هستند و مقدار ۰ به معنای $False$ بودن آن متغیر تصادفی و ۱ بودن به معنای $True$ بودن آن است.

در خط انتهایی ورودی شماره دو راس $query$ که با یک اسپیس جدا شده‌اند داده می‌شود. ابتدا مستقل و یا وابسته بودن این دو راس را بررسی کنید. در صورتی که مستقل باشند عبارت $independent$ و در صورتی که وابسته باشند، عبارت $dependent$ را در خروجی چاپ کنید. پس از آن نیز به ازای هر یک از دو راس کوئری احتمال آنکه آن متغیر درست باشد را به ازای evidence های موجود به‌دست آورده و هریک را تا دو رقم اعشار گرد کرده و در یک خط مجزا چاپ کنید.

توجه: جدول احتمالات شرطی تنها دارای خانه‌هایی است که به ازای آن‌ها مقدار راس مورد نظر $True$ باشد. به دست آوردن مقادیر در حالتی که آن متغیر تصادفی مقدار $False$ داشته باشد بدیهی است. جهت توضیحات بیشتر به نمونه ورودی و گراف آن دقت کنید.

نمونه ورودی:



input:
5

0.001

0.002
1 2
0.95 0.94 0.29 0.001
3
0.9 0.05
3
0.7 0.1
1->1,2->1

4 5

output:

dependent

0.86

0.67

نکات:

شما باید فقط یک فایل پایتون آپلود کنید که ورودی را از کنسول بخواند و خروجی را چاپ کند.

نمره نهایی شما، پس از بررسی و تصحیح کد شما محاسبه می‌شود و نمره قابل مشاهده در کوئرا، نمره نهایی شما نیست.

۲. (۳۵ نمره) **بخش اول**

در این بخش از شما می‌خواهیم که یک سمپلینگ ساده را انجام دهید. ۳ توزیع به شما داده شده‌است و شما باید از این ۳ توزیع سمپل کنید. ۲ تا از این توزیع‌ها توزیع‌های پیوسته و یکی دیگر توزیع گسسته است.

نکته مهم آن است که شما حق استفاده از هیچ سمپلر آماده را ندارید و تنها حق استفاده از تابع تولید کننده عدد رندم از یک توزیع یونیفرم (بازه ها می‌توانند هر عدد دلخواهی باشند.) را دارید. همچنین دقت کنید که امکان استفاده از توابعی که کتابخانه‌هایی نظیر $scipy$ ارائه می‌دهند تا از یک توزیع احتمال $custom$ سمپل بگیرید را ندارید، و درصورت استفاده نمره ای از این بخش نمی‌گیرید.

برای محاسبه مقدار pdf هر توزیع می‌توانید از توابع آماده نامپای یا scipy استفاده کنید.

توزیع ها

$$۱) \frac{۳}{۱۰} gaussian(۴, ۲) + \frac{۳}{۱۰} gaussian(۳, ۲) + \frac{۴}{۱۰} exponential(۰/۰۱)$$

$$۲) \frac{۲}{۱۰} gaussian(۰, ۱۰) + \frac{۲}{۱۰} gaussian(۲۰, ۱۵) + \frac{۳}{۱۰} gaussian(-۱۰, ۸) + \frac{۳}{۱۰} gaussian(۵۰, ۲۵)$$

$$۳) \frac{۲}{۱۰} geometric(۰/۱) + \frac{۲}{۱۰} geometric(۰/۵) + \frac{۲}{۱۰} geometric(۰/۳) + \frac{۴}{۱۰} geometric(۰/۰۴)$$

توجه کنید که پارامتر دوم توزیع گوسی واریانس توزیع است نه انحراف معیار

تحویل دادنی های این بخش

برای هر تابع توزیع احتمال ۱ نمودار هیستوگرام از سمپل های آن و یک نمودار خود توزیع را رسم کنید. فرمت نام‌گذاری آن‌ها باید به صورت زیر باشد. همچنین یک فایل log.txt در هر خط آن میانگین و انحراف معیار را تا ۴ رقم اعشار گزارش کنید. (میانگین و انحراف معیار سمپل هایی که بدست آورده‌اید) تمامی این فایل‌ها را در پوشه‌ای به نام part1 قرار دهید. دقت کنید که این پوشه در کنار کد شما باید توسط کد شما ایجاد شود.

for each pdf 2 files:
pdf<num>.png
pdf<num>_sample.png
log.txt:

1 42.4419 79.6389
2 15.9503 24.8859
3 12.9775 18.802

دقت کنید که شماره‌ها را به ترتیب در ابتدای خط قرار دهید و سپس میانگین و بعد از آن انحراف معیار را با ۴ رقم اعشار گزارش دهید.

در یک فایل گزارش باید روند خود را توضیح دهید گزارش به صورت ژوپیتر قابل قبول نیست.

توضیح دهید برای سمپل کردن چه روشی را در پیش گرفتید و اگر نیاز به اثبات دارید آن‌را اثبات کنید.

بنابراین ۶ عکس نمودار و یک فایل لاگ و یک گزارش تحویلی‌های این بخش هستند. کد خودتان را نیز در بخش مربوطه اپلود کنید. دقت کنید که کدتان باید قابلیت ران شدن دوباره را داشته باشد

بخش دوم

حال به سراغ بخش اصلی تمرین یعنی سمپل کردن از یک شبکه بیزین می‌رویم. در این تمرین شما باید ۴ روش سمپلینگ توضیح داده شده در کلاس را پیاده‌سازی کنید. تمام ورودی‌های زیر باید از فایل خوانده شوند.

ورودی

ورودی شما یک گراف بیزین است که به صورت زیر مشخص میشود.

در خط اول تعداد راس های گراف می‌آید.

در خط بعد نام یک راس می‌آید و سپس جدول آن اگر راس پدر نداشته باشد آنگاه در یک خط احتمال ۱ بودن آن می‌آید. اگر پدر داشته باشد در یک خط نام پدر ها با فاصله از هم می‌آیند. سپس جدول آن در خطوط بعد می‌آید دقت کنید تنها احتمال ۱ بودن راس مربوطه ذکر می‌شود. به عنوان مثال داریم:

4
A
0.6
B
0.4
C
A B
0 0 0.2
0 1 0.3
1 0 0.2
1 1 0.9
D
C
0 0.1
1 0.3

دقت کنید که تضمین نمی‌شود که ورودی داده شده ترتیب توپولوژی را رعایت کند.

فرمت کوئری ها نیز بدین شکل است.

[[{"D": 1 }, {}],[{"A": 1 , "B" : 1}, {"C": 1 , "D" : 1}]]

که معادل یک لیست از ۲ کوئری به شکل زیر است

[P(D=1),P(A=1 , B=1 | C=1 , D=1)]

یک لیست از کوئری ها به شما داده می‌شود و شما باید به ازای هر کدام از آنها مقدار واقعی کوئری و اختلاف آنها با هر کدام از شیوه‌های سمپلینگ را بدست آورید.

در نهایت خروجی شما به شکل زیر است.

یک فایل با نام شماره گراف (در ادامه توضیح داده می‌شود) که درون آن در هر خط ابتدا مقدار واقعی کوئری نوشته می‌شود. پس از آن مقادیر قدرمطلق اختلاف هریک از شیوه‌های سمپلینگ و مقدار واقعی با یک فاصله نوشته می‌شود. ترتیب نوشته شدن این مقادیر در فایل به صورت زیر است.

real-value prior rejection likelihood-weighting Gibbs

یک نمودار که محور افقی آن نمایانگر شماره کوئری و محور عمودی آن نمایانگر اختلاف با مقدار واقعی برای ۴ نوع سمپلینگ است.

نام فایل تکس و فایل عکس برابر شماره گراف است.

یک نمونه از فایل های ورودی و وخروجی به شما داده شده است. به ساختار آن نگاه کنید.

در نهایت کد شما باید از درون پوشه inputs پوشه مربوط به هر گراف را بخواند. درون هر پوشه یک فایل input.txt قرار دارد که گراف مد نظر در آن است. و یک فایل q_input.txt که یک لیست از کوئری ها به شما داده می‌شود. و شما باید آنرا به فرمت خواسته شده خروجی دهید.

دقت کنید که برای محاسبه مقدار واقعی کوئری مجاز هستید از هر روشی استفاده کنید. برای سادگی می‌توانید کل جدول را تشکیل دهید و یا از کد سوال اول استفاده کنید اما دقت کنید که پیاده سازی روش های بهینه‌تر نمره اضافه ای نخواهند داشت.

تحویل دادنی های این بخش

در یک پوشه output که در کنار کد توسط کد شما تولید می‌شود باید فایل های تکست و عکس نمودار هر گراف را قرار دهید (در همان پوشه و نه داخل پوشه دیگری) اسم عکس و فایل تکست شماره گراف است. کد شما باید بتواند تعداد گراف های متفاوتی را هندل کند بنابراین تعداد را هاردکد نکنید. یک نمونه ورودی و خروجی به شما داده می شود تا با فرمت آن اشنا شوید.

نکات مهم

کد بخش اول با نام simple.py و کد بخش دوم با نام sample.py باید قرار داده شوند.

پوشه ورودی برای بخش دوم در اختیار شما قرار داده می‌شود اما پوشه مربوط به فایل های خروجی بخش یک و دو را باید کد شما تولید کند یعنی output , part1 توابع زیر ممکن است کمک کننده باشند.

json.load
map
bin
os.listdir
os.path.isfile

کتابخانه های مجاز

random
numpy as np
pandas as pd
matplotlib
sys
seaborn
scipy.stats import norm
scipy.stats import expon
all python built in libraries (The Python Standard Library)