دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

تمرین سری اول آمار و احتمال مهندسی

زمستان ۹۶

مقدمه

سلام!

این تمرین دارای دو بخش است. بخش اول سوالات حل کردنی و بخش دوم سوالاتی با عنوان آزمایش‌های کامپیوتری است. برای بخش اول یک PDF ازتون میخواهیم که توش سوالات حل شده باشه و اسم فایلش باید به فرم

PS_HW1_PART1_9231020.pdf

باشه. اگه توی برگه دوست دارید بنویسید، باز در آخر یه فایل pdf بسازید و مثلا یه سری فایل jpg آپلود نکنید.

برای بخش دوم هم، هم فایل‌های مربوط به برنامه‌تون رو می‌خواهیم و هم یک گزارش که اونم به صورت pdf باید باشه و فرمت اسمش هم به صورت PS_HW1_CE_9231020.pdf بگذارید.

تاکید می‌کنم که هم گزارش باید باشه هم فایل برنامه! حالا دقیق‌ترش رو توی بخش مربوط به خودش توضیح‌دادیم.

ما فایل های شما رو فقط و فقط از طریق مودل دریافت می‌کنیم و ارسال آن‌ها از طریق‌های دیگر به هیچ‌وجه مورد بررسی قرارنمی‌گیرد.

در ضمن توی کانال تلگرامی‌ای که مربوط به این درس هست هم می‌تونید عضو شید.

Link: https://t.me/PS9697

در صورتی که نیاز باشه خبری در مورد تمرین‌ها، کلاس تی‌ای و ... بدهیم سعی می‌کنیم هم از طریق این کانال این خبر رو انتقال بدهیم هم از طریق مودل.

اگر سوالی در رابطه با تمرین‌ها داشتید می‌تونید از طریق این ایمیل از ما بپرسید.

Email Address: autps9697@gmail.com

بخش اول، سوالات حل‌کردنی

سوال ۱

الف)

در کیسه ای ۳ مهره سفید و ۴ مهره سیاه وجود دارد. از آن کیسه یک مهره را به تصادف برمیدارم و خارج میکنیم. (رنگش رو نمی‌بینیم و میندازیم دور) سپس یک مهره دیگر بر می‌داریم. احتمال اینکه این مهره سفید باشد چقدر است؟

ب)

در کیسه ای ۷ مهره سفید و ۸ مهره سیاه داریم. ۱۰ مهره به تصادف برمیداریم و خارج می‌کنیم. (رنگش رو نمیبینیم و میندازیم دور) سپس یک مهره دیگر بر میداریم. احتمال اینکه این مهره سفید باشد چقدر است؟

سوال ۲

جعبه A شامل ۵ مهره سفید و ۵ مهره قرمز است. جعبه B شامل ۵ مهره سفید و ۲ مهره قرمز است. مهره‌ای به تصادف از جعبه A انتخاب کرده و درون جعبه B قرار می‌دهیم. سپس جعبه‌ای را به تصادف انتخاب کرده و دو مهره از آن به تصادف بیرون می‌آوریم. احتمال اینکه هردومهره سفید باشد چقدر است؟

سوال ۳

سه جعبه داریم. جعبه اول حاوی دو سکه طلا، جعبه دوم حاوی یک سکه طلا و یک سکه نقره و جعبه سوم حاوی دو سکه نقره است. بدون اینکه داخل جعبه‌ها را ببینیم، یک جعبه را انتخاب کرده و یک سکه از درون آن در می‌آوریم. اگر آن سکه طلا باشد، احتمال اینکه سکه دیگر آن جعبه نیز طلا باشد چقدر است؟

سوال ۴

سه سکه داریم با رنگ های آبی، قرمز و سیاه. که هریک با احتمال $\frac{3}{5}$ و مستقل از بقیه رو می‌آید. سکه‌ی اول در صورت رو آمدن ۱۰ امتیاز دارد و در صورت زیر آمدن ۲ امتیاز. سکه‌ی دوم در هر دو صورت ۴ امتیاز دارد و سکه‌ی سوم در صورت رو آمدن ۳ امتیاز و در صورت زیر آمدن ۲۰ امتیاز دارد. شما و حریفتان این بازی را انجام می‌دهید. بازی به این صورت است که نفر اول یک سکه را انتخاب می‌کند و نفر دوم یکی از سکه‌های باقی‌مانده را بر می‌دارد. سپس هر کدام سکه خود را پرتاب می‌کنند. هرکسی که امتیاز بیشتری بیاورد، ۱ میلیون تومان جایزه می‌گیرد. شما ترجیح می‌دهید نفر اول باشید یا دوم؟

(احتمال برنده شدن نفر اول و دوم در این بازی را بدست بیاورید)

سوال ۵

یکی از روش‌های اولیه در یادگیری ماشین استفاده از Naïve Bayes Classifier برای دسته‌بندی داده‌ها است. یکی از مثال‌های آن دسته‌بندی ایمیل‌ها به دو دسته‌ی Spam و Not Spam است. در این سوال قصد داریم به بررسی این روش بپردازیم.

در واقع سیستم ما یک ایمیل می‌گیرد و باید تشخیص بدهد که آیا آن ایمیل Spam است یا خیر. در این حالت باید با توجه به ویژگی‌های ایمیل، احتمال Spam بودن را بیان کند اگر این احتمال بیش از 0.5 بود، آن را Spam درنظر می‌گیریم.

یکی از ویژگی‌های یک ایمیل وجود کلمه 'free' در آن ایمیل است.

فرض کنید پیشامد X یعنی اینکه کلمه 'free' در یک ایمیل آمده باشد.

و پیشامد X~ یعنی کلمه 'free' در ایمیل نیامده باشد.

در این صورت، با دانستن اینکه پیشامد X و یا پیشامد X~ رخ داده است، می‌خواهیم حدس بزنیم که آیا ایمیل دریافتی Spam است یا نه.

برای محاسبه این احتمالات از قانون بیز استفاده می‌کنیم.

فرض کنید می‌دانیم که احتمال Spam بودن یک ایمیل بدون داشتن هیچ اطلاعی برابر با 0.8 است.

$$P(Spam) = 0.8$$

$$P(Not\ Spam) = 1 - 0.8 = 0.2$$

همچنین می‌دانیم که در در 0.9 ایمیل های Spam، کلمه free وجود دارد. و در 0.15 از ایمیل‌های سالم (غیر Spam) کلمه free وجود دارد. یعنی:

$$P(X|Spam) = 0.9, P(\sim X|Spam) = 0.1$$

$$P(X|Not\ Spam) = 0.15, P(\sim X|Not\ Spam) = 0.85$$

حال یک Email دریافت می‌کنیم که در آن Email کلمه 'free' وجود دارد.

در این صورت میخواهیم احتمال Spam بودن آن را تشخیص دهیم. ما باید دو احتمال را بدست آوریم و با هم مقایسه کنیم.

احتمال $P(Spam|X)$ و $P(Not\ Spam|X)$. (البته این دو در این حالت خاص مکمل همدیگر هستند.)

این احتمال را می‌توان به صورت زیر نوشت:

$$P(Spam|X) = \frac{P(Spam) * P(X|Spam)}{P(X)}$$

$$P(Not\ Spam|X) = \frac{P(Not\ Spam) * P(X|Not\ Spam)}{P(X)}$$

که $P(X)$ احتمال آمدن کلمه free به طور کلی در یک متن است. که البته محاسبه آن لزومی ندارد. زیرا میخواهیم دو احتمال بالا را باهم مقایسه کنیم و ببینیم کدام یک بزرگتر است. با توجه به یکسان بودن مخرج این دو احتمال، می‌توان از محاسبه آن صرف نظر کرد.

پس:

$$P(Spam|X) = \frac{1}{P(X)} * P(Spam) * P(X|Spam) = \frac{1}{P(X)} * 0.8 * 0.9 = \frac{1}{P(X)} * 0.72$$

$$P(Not\ Spam|X) = \frac{1}{P(X)} * P(Not\ Spam) * P(X|Not\ Spam) = \frac{1}{P(X)} * 0.2 * 0.15 = \frac{1}{P(X)} * 0.03$$

همانطور که می‌بینید، $P(Spam|X) > P(Not\ Spam|X)$ پس می‌توان گفت احتمال Spam بودن این ایمیل بیشتر است.

با توجه به اینکه

$$P(Spam|X) + P(Not\ Spam|X) = 1 \rightarrow P(X) = 0.75$$

البته نیازی به محاسبه اش نداشتیم.

حال می‌خواهیم یک ایمیل که در آن کلمه free وجود ندارد را بررسی کنیم و احتمال Spam بودنش را تشخیص دهیم.

در این حالت:

$$P(Spam|{\sim}X) = \frac{1}{P({\sim}X)} * P(Spam) * P({\sim}X|Spam) = \frac{1}{P({\sim}X)} * 0.8 * 0.1 = \frac{1}{P({\sim}X)} * 0.08$$

$$P(Not\ Spam|{\sim}X) = \frac{1}{P({\sim}X)} * P(Not\ Spam) * P({\sim}X|Not\ Spam) = \frac{1}{P({\sim}X)} * 0.2 * 0.85 = \frac{1}{P({\sim}X)} * 0.17$$

در این حالت $P(Spam|{\sim}X) < P(Not\ Spam|{\sim}X)$. پس ایمیل Not Spam است.

حال یک سوال پیش می‌آید. اینکه این احتمالات را از کجا بدست بیاوریم.

برای این کار باید با توجه به مجموعه داده‌هایی که توسط سرویس های ایمیل وجود دارند این کار را انجام داد.

در این مساله از شما میخواهیم که احتمال

فرض کنید با توجه به ایمیل‌های زیر می‌خواهیم احتمالات را بدست بیاوریم.

الف)

داده‌های مربوط به این مساله به این صورت است که در صورت رخدادن کلمات جدول در متن ایمیل، مقدار آن عدد ۱ و در غیر این صورت ۰ است.

Data	Free	Money	Credit	Inference	Description	Traditional	The	Class

1	1	1	0	0	1	0	1	Spam
2	0	1	1	0	0	1	1	Spam
3	1	0	0	0	0	0	1	Spam
4	1	0	1	1	0	0	1	Spam
5	1	1	1	0	0	1	1	Spam
6	1	1	1	0	1	0	1	Spam
7	1	0	0	0	0	0	1	Spam
8	1	0	1	0	1	1	1	Not Spam
9	0	0	0	1	0	1	1	Not Spam
10	0	1	1	1	1	0	1	Not Spam

جدول ۱داده های آموزش

Data	Free	Money	Credit	Inference	Description	Traditional	The	Spam Probability
11	1	1	1	0	0	0	1	
12	1	1	0	0	0	1	1	
13	0	0	0	1	1	1	1	
14	0	0	1	0	1	1	1	
15	1	0	0	1	1	1	1	

جدول ۲ داده تست

با توجه به جدول اطلاعات مربوط به ایمیل‌ها در مجموعه داده آموزش جدول ۱، احتمالات زیر را حساب کنید.

$$P(Spam)$$

$$P(Not\ Spam)$$

$$P(X|Spam), P(\sim X|Spam)$$

$$P(X|Not\ Spam), P(\sim X|Not\ Spam)$$

راهنمایی:

برای محاسبه $P(Spam)$ می‌توان تعداد Spam ها به کل را حساب کرد که برابر با 0.7 است.

برای محاسبه $P(X|Spam)$ به صورت زیر عمل کنید:

$$P(X|Spam) = \frac{\#(X, Spam)}{\#(Spam)}$$

و سپس احتمال Spam بودن و Not Spam بودن را برای مجموعه داده تست جدول ۲، حساب کنید.

ب)

در این قسمت، می‌خواهیم ویژگی‌های بیشتری را برای تصمیم‌گیری در نظر بگیریم. برای مثال فرض کنید

پیشامد Y به معنی رخداد کلمه Description در متن یک ایمیل باشد و Y~ نیز به معنی عدم رخداد آن کلمه.

در این صورت در حالتی که در یک متن کلمه free آمده باشد و کلمه Description نیامده باشد، احتمال Spam بودن آن متن چقدر است؟

$$P(Spam|X,\sim Y) = \frac{P(Spam) * P(X,\sim Y|Spam)}{P(X,\sim Y)} = \frac{1}{P(X,\sim Y)} * P(Spam) * P(X,\sim Y|Spam)$$

$$P(\sim Spam|X,\sim Y) = \frac{P(Not\ Spam) * P(X,\sim Y|Not\ Spam)}{P(X,\sim Y)}$$
$$= \frac{1}{P(X,\sim Y)} * P(\sim Spam) * P(X,\sim Y|NotSpam)$$

در این بخش نیازی به محاسبه دوباره $P(Spam)$ نداریم.

برای محاسبه $P(X,\sim Y|Not\ Spam)$ چه راهی به ذهنتان می‌رسد؟

یکی از راه‌ها شمردن تمام حالت‌ها است.

$$P(X,\sim Y|Not\ Spam) = \frac{\#(X,\sim Y, Not\ Spam)}{\#(Not\ Spam)}$$

اما در صورتی که فرض کنیم پیشامد های X و Y~ از هم مستقل هستند، می‌توانیم به صورت زیر عمل کنیم:

$$P(X,\sim Y|Not\ Spam) = P(X|Not\ Spam) * P(\sim Y|Not\ Spam) = \frac{\#(X, Not\ Spam)}{\#(Not\ Spam)} * \frac{\#(\sim Y, Not\ Spam)}{\#(Not\ Spam)}$$

روش Naïve Bayes از این فرض استفاده می‌کند که پیشامدها از هم مستقل هستند.

در این قسمت با استفاده از این دو ویژگی (free, description) دوباره مقادیر احتمال Spam بودن را برای جدول ۲ بدست بیاورید و گزارش کنید.

ج)

در این قسمت از تمام ویژگی‌ها برای تعیین Spam بودن یا نبودن یک میل استفاده کنید. و نتایج را برای جدول ۲ بدست بیاورید.

د)

در این قسمت مقادیر سه قسمت قبل را باهم مقایسه کنید. در این قسمت Label داده‌های تست نیز به شما داده شده است.

(یعنی مثلا یه چیزی شبیه جدول ۳ درست کنید و احتمال های بدست اومده رو کنار هم بذارید)

به نظر شما کدام یک از این سه مجموعه ویژگی‌ها مناسب‌تر است و به طور کلی چگونه می‌توان یک زیرمجموعه مناسب از ویژگی‌ها انتخاب کرد؟

Data	{free}	{free, description}	Complete set	Real Label (Spam)
11				1
12				1
13				0
14				0
15				0

جدول ۳ مقایسه حالت های مختلف

سوال ۶

الف)

در صورتی که A,B,C سه پیشامد دلخواه باشند، با استفاده از تعریف نشان‌دهید:

$$P(A,B,C) = P(A) * P(B|A) * P(C|A,B)$$

(البته با فرض اینکه $P(B|A), P(C|A,B)$ تعریف‌شده باشند.)

ب)

در این قسمت به حل یک مثال ساده از شبکه‌های بیز[۱] می‌پردازیم. فرض کنید می‌خواهیم با توجه به دانستن ویژگی‌های یک دانشجو، در مورد احتمال ریکام خوب گرفتن نظر بدهیم.

فرض کنید یک دانشجو هستید و می‌خواهید از یکی از استادان دانشگاه امیرکبیر recommendation letter بگیرید. اما نمی‌دانید که از استاد کدام درس با توجه به شرایط شما، recommendation letter بهتری برای شما می‌نویسد.

از این رو، می‌خواهید احتمال اینکه استاد برای شما recommendation letter خوب بنویسد را با توجه به داده‌های سال‌های گذشته پیش بینی کنید.

فرض کنید ما اطلاعات دانشجویان امیرکبیر در ۱۰ سال گذشته را جمع‌آوری کردیم. مانند جدول ۴:

Student ID	Intelligence	Exam Difficulty	Grade	Recommendation Letter for that course
8431001	Smart	Easy	Good	R3
8431001	Smart	Hard	Good	R3
8431002	Normal	Easy	Good	R3

[۱] Bayesian Networks

8431002	Normal	Hard	Bad	R1
…				
9331999	Smart	Easy	Good	R2

جدول ۴ اطلاعات دانشجویان در ۱۰ سال گذشته

توجه کنید که هر سطر مربوط به یک درس و یک دانشجو است. اطلاعات هر سطر به ترتیب، هوش فرد، میزان سختی درس، نمره آن درس و توصیه‌نامه‌ای که استاد آن درس بر اساس نمره آن درس داده است را نشان می‌دهد. برای مثال دانشجوی با شماره دانشجویی 8431001 در دو درس دو توصیه نامه خیلی خوب گرفته است و دانشجوی با شماره دانشجویی 8431002 یک توصیه نامه خیلی خوب و یک توصیه نامه بد گرفته است.

مقادیر متغیرها به صورت زیر است:

Recommendation Letter for that course	معنا
R3	خیلی خوب
R2	معمولی
R1	بد

Intelligence	Exam Difficulty	Grade
Normal=I0	Easy=D0	Good=G2
Smart=I1	Hard=D1	Normal=G1
		Bad=G0

حال می‌خواهیم حدس بزنیم برای یک دانشجو با اطلاعات (برای مثال) زیر:

Student ID	Intelligence	Exam Difficulty	Grade	Recommendation Letter for that course
9531020	Smart=I1	Easy=D0	Good=G3	?

جدول ۵ یک داده تست

احتمال اینکه پس از درخواست به استاد مربوطه، توصیه نامه خوب (R3) بگیرد را بدست آوریم.

در واقع هدف ما محاسبه مقدار زیر است:

$$P(R = R_3 | I = I_1, D = D_0, G = G_3) = \frac{P(R = R_3, I = I_1, D = D_0, G = G_3)}{P(I = I_1, D = D_0, G = G_3)}$$

حال به محاسبه آن می‌پردازیم:

با توجه به قسمت الف می‌دانیم که رابطه زیر برقرار است.

$$P(R,I,D,G) = P(D) * P(I|D) * P(G|I,D) * P(R|I,D,G)$$ رابطه ۱

همچنین می‌دانیم که میزان هوش یک فرد به میزان سختی امتحان ندارد. در نتیجه

$$P(I) = P(I|D)$$

همچنین با توجه به بررسی‌ها متوجه شدیم که استادها برای دادن برگه توصیه نامه، تنها به نمره فرد توجه می‌کنند. در نتیجه توصیه‌نامه از دو پیشامد D و G مستقل است.

$$P(R|I,D,G) = P(R|G)$$

در نتیجه رابطه ۱ به صورت زیر ساده می‌شود.

$$P(R,I,D,G) = P(D) * P(I) * P(G|I,D) * P(R|G)$$ رابطه ۲

برای محاسبه‌ی $P(I,D,G)$ نیز می‌توانید از قواعد بالا استفاده کنید. در نتیجه

$$P(I,D,G) = P(I) * P(D) * P(G|D,I)$$

جداول مربوط به احتمالات با توجه به داده‌های مربوط به ۱۰ سال بدست آمده است و به صورت زیر است:

Intelligence	Probability
I0	0.7
I1	0.3

Exam Difficulty	Probability
D0	0.6
D1	0.4

	G0 Probability	G1 Probability	G2 Probability
I0, D0	0.3	0.4	0.3
I0, D1	0.7	0.25	0.05
I1, D0	0.02	0.08	0.9
I1, D1	0.5	0.3	0.2

برای مثال با توجه به جدول بالا:

$$P(G = G_0|I = I_0, D = D_0) = 0.3$$

$$P(G = G_1|I = I_0, D = D_0) = 0.4$$

Grade	R1 Probability	R2 Probability	R3 Probability
G0	0.7	0.29	0.01
G1	0.2	0.6	0.2
G2	0.01	0.19	0.8

برای مثال با توجه به جدول بالا:

$$P(R = R_1|G = G_0) = 0.7$$

$$P(R = R_2|G = G_0) = 0.29$$

حال با توجه به جداول مربوط به احتمالات، احتمال توصیه نامه مختلف را برای داده جدول ۵ حساب کنید.

Student ID	Intelligence	Exam Difficulty	Grade	Probability of R1	Probability of R2	Probability of R3
9531020	Smart=I1	Easy=D0	Good=G3	?	?	?

جدول ۶ نتیجه خواسته شده

بخش دوم، سوالات آزمایش‌های کامپیوتری

در این بخش می‌توانید از زبان‌های python، matlab و R استفاده کنید.

فایل گزارش به ازای این دو سوال را در یک pdf به فرم PS_HW1_CE_9231020.pdf بنویسید.

فایل های غیرpdf مورد نیاز برای هربخش در خود سوال گفته شده‌است.

سوال اول

کیهان و ایمان داشتند برای امتحان میانترم آمارشون تمرین می‌کردند که به این سوال مواجه شدند.

"مهناز یک سکه‌ی متقارن را دوبار به بالا پرت می‌کند. می‌دانیم در حداقل یکی از این پرتاب‌ها شیر آمده است. احتمال اینکه سکه در هردو پرتاب شیر آمده باشد چقدر است؟"

کیهان به این سوال این پاسخ را می‌دهد.

"می‌دونیم که حداقل یکی از پرتاب‌ها شیر اومده. دونستن این موضوع برای اینکه در مورد سکه دیگه نظر بدیم تاثیری نداره. زیرا سکه دوم مستقلا از سکه اول پرتاب شده پس به احتمال ½ شیر و به احتمال ½ رو می‌آد.

اما ایمان پاسخ دیگری داشت:

"وقتی دوبار سکه رو بالا می‌ندازیم فضای رخدادهای ما به صورت {HH, HT, TH, TT} هست. (T: خط، H: شیر) احتمال این‌که حداقل یکی از سکه‌ها شیر اومده باشه ¾ هست چون سه پیش‌آمد از فضای اولیه رو پوشش میده. {HT, TH, TT}. احتمال این‌که هردو سکه شیر بیان توی مسئله‌ی ما برابر ¼ هست چون فقط یک حالت از فضای پیش‌آمدهای ما رو شامل می‌شه.

بنابراین جواب سوال برابر ⅓ هست."

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{3/4} = \frac{1}{3}$$

شما با کدام جواب موافقید؟ برای فهمیدن جواب درست، می‌خواهیم از آزمایش کامپیوتری استفاده‌کنیم.

آزمایش کامپیوتری ما شامل دو قسمت است.

آزمایش اول

فایل برنامه آزمایش اول را به صورت HW1_CE_1_a نامگذاری کنید.
(برای مثال در صورتی که از زبان پایتون استفاده می‌کنید: HW1_CE_1_a.py)

ابتدا چند متغیر به صورت زیر تعریف کنید:

```
number_of_experiment = 1000
exp1_n = 0
exp1_f = 0
a = b = None
```

این کار را تا زمانی که number_of_experiment صفر نشده‌است، انجام دهید:

به طور تصادفی (به احتمال 0.5 مقدار ۱ و به احتمال 0.5 مقدار صفر بدهید) یک مقدار به متغیر a بدهید. و سپس به طور تصادفی یک مقدار به متغیر b بدهید.

در صورتی که متغیرa=0 بود:

به ازای این مرحله هیچکاری انجام ندهید.

در غیر این صورت (در صورتی که a=1 بود):

number_of_experiment را یک واحد کم کرده ، exp1_n را یک واحد زیاد بکنید.

سپس به متغیر b نگاه کنید.

در صورتی که b=1 بود:

مقدار exp1_f را یک واحد زیاد کنید.

در غیر این صورت (در صورتی که b=0 بود):

هیچ مقداری را تغییر ندهید.

در نهایت نسبت $\frac{exp1-f}{exp1-n}$ را چاپ کنید. (برنامه رو اجرا کنیم این مقادیر رو چاپ کنه)

این آزمایش را به ازای number_of_experiment={10, 100, 1000, 10000} انجام داده و جدول زیر را پر کنید.

Number of Experiment	$\frac{exp1-f}{exp1-n}$
10	
100	
1000	
10000	

آزمایش دوم

فایل برنامه آزمایش اول را به صورت HW1_CE_1_b نامگذاری کنید.
(برای مثال در صورتی که از زبان پایتون استفاده می‌کنید: HW1_CE_1_b.py)
این آزمایش بسیار شبیه به آزمایش اول است، اما در قسمتی با آن متفاوت است.

ابتدا چند متغیر به صورت زیر تعریف کنید:

```
number_of_experiment = 1000
exp2_n = 0
exp2_f = 0
a = b = a_prime = b_prime = None
```

این کار را تا زمانی که number_of_experiment صفر نشده‌است، انجام دهید:

به طور تصادفی (به احتمال 0.5 مقدار ۱ و به احتمال 0.5 مقدار صفر بدهید) یک مقدار به متغیر a بدهید. و سپس به طور تصادفی یک مقدار به متغیر b بدهید.

به احتمال 0.5 متغیر های a_prime, b_prime را اینگونه مقدار دهی کنید:

a_prime=a, b_prime=b

و به احتمال 0.5 متغیرهای a_prime, b_prime را اینگونه مقدار دهی کنید:

a_prime=b, b_prime=a

در صورتی که متغیر a_prime=0 بود:

به ازای این مرحله هیچکاری انجام ندهید.

در غیر این صورت (در صورتی که c=1 بود):

number_of_experiment را یک واحد کم کرده ، exp2_n را یک واحد زیاد بکنید.

سپس به متغیر b_prime نگاه کنید.

در صورتی که b_prime=1 بود:

مقدار exp2_f را یک واحد زیاد کنید.

در غیر این صورت (در صورتی که b_prime=0 بود):

هیچ مقداری را تغییر ندهید.

در نهایت نسبت $\frac{exp2-f}{exp2-n}$ را چاپ کنید. (برنامه رو اجرا کنیم این مقادیر رو چاپ کنه)

این آزمایش را به ازای number_of_experiment={10, 100, 1000, 10000} انجام داده و جدول زیر را پر کنید.

Number of Experiment	$\frac{exp2-f}{exp2-n}$
10	
100	

1000	
10000	

با توجه به مقادیر بدست آمده در دو آزمایش بالا، بگویید نظر کیهان (احتمال ½) درست است یا نظر ایمان (احتمال 1/3)

سوال دوم

فایل برنامه آزمایش اول را به صورت HW1_CE_SPAM نامگذاری کنید.
(برای مثال در صورتی که از زبان پایتون استفاده می‌کنید: HW1_CE_SPAM.py)
علاوه بر این فایل، یک فایل با فرمت csv نیز برای این سوال از شما خواسته شده است که باید به فرم SPM_9231020.csv باشد.

در این سوال می‌خواهیم با توجه به داده های مربوط به Spam detection، تشخیص دهیم یک داده آیا Spam است یا خیر.

مساله ما همان مساله ای است که در سوال ۵ بخش حل کردنی توضیحات آن را گفتیم. البته با داده های بیشتر! شما در آن سوال به صورت دستی، تعداد رخداد را شمردید و احتمالات را بدست آوردید. حال در این بخش می‌خواهیم همانکار را به صورت برنامه نویسی انجام بدهید. و پس از آن، خروجی را در یک فایل ذخیره سازی کنید.

در این تمرین ۳ فایل به شما داده می‌شود. train.csv و test.csv و answers.csv.

شما باید با خواندن فایل train، احتمالات مربوط به

$$\forall 1 \leq i \leq 58: P(feature_i = 1|Spam), P(feature_i = 1|NotSpam)$$

را به ازای تمام ۵۸ ویژگی با استفاده از داده‌ها یاد بگیرید.

سپس در قسمت test، به ازای هر داده بگویید آیا این داده Spam است یا نه. (در صورتی که Spam بود عدد ۱ و در غیر این صورت عدد صفر را قرار دهید) و یک فایل خروجی مانند فایل answers به فرم SPM_9231020.csv تحویل دهید.

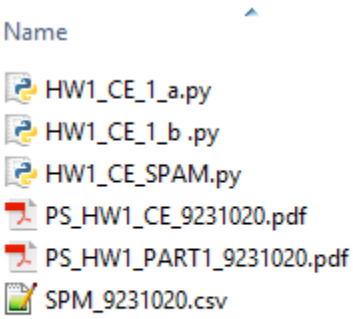
برای مثال در زبان پایتون این برنامه ورودی را به درستی دریافت می‌کند و فرمت خروجی اش درست است. اما نمره‌ای ندارد!

```
import numpy as np
import pandas as pd
train  = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
#Do something
cheat = pd.read_csv("answers.csv")
output = pd.DataFrame(cheat)
output.to_csv("SPM_9231020", index=False)
```

توجه کنید که ممکن است کد شما را با داده های جدید آموزش بدهیم و خروجی برنامه شما را با جواب واقعی مقایسه کنیم!

فایل‌های خواسته شده از شما

فایل های خروجی شما باید اینطوری باشد:



و همچین فایلی رو zip کرده و به فرم PS_HW1_9231020.zip در مودل آپلود نمایید.