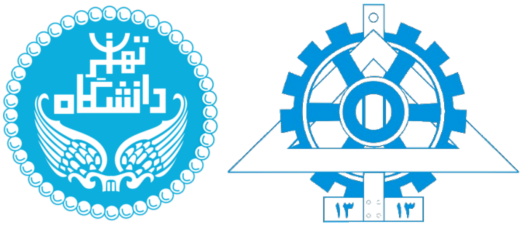

# دانشگاه تهران، دانشکده مهندسی برق و کامپیوتر
# آمار و احتمال مهندسی

تمرین کامپیوتری صفر ـ پایتون و قانون بیز

طراح: علی مهاجری

سوپروایزر: علی محمدی

تاریخ تحویل: ۱۴۰۲/۷/۳۰

## ۱. قانون بیز و پردازش متن — ۱۴۰ نمره

در پروژه صفر قصد داریم که با زبان برنامه‌نویسی پایتون آشنا شده و با استفاده از کتابخانه‌های کاربردی آن، قسمت‌های مختلف یک پروژه را با استفاده از مفاهیم پایه‌ی آمار و احتمال انجام دهیم. به همین منظور، برای مرور و یا آشنا شدن با پایتون و توابع و کتابخانه‌های آن قبل از ادامه دادن صورت پروژه، فایل Python_Intro.rar را مطالعه کنید.

پیش از شروع صورت پروژه، ابتدا به مرور تئوری قانون بیز می‌پردازیم.

### قانون بیز — Bayes' Theorem

قانون بیز برای محاسبه احتمال وقوع یک رویداد با توجه به دانش ما در مورد رویدادهای قبلی و شرایط موجود استفاده می‌شود. به عبارت دیگر، این قانون برای برآورد احتمال وقوع یک رویداد جدید با توجه به دانش قبلی و شرایط فعلی استفاده می‌شود:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|x) \rightarrow Posterior, \quad P(x|c) \rightarrow Likelihood, \quad P(c) \rightarrow Prior, \quad P(x) \rightarrow Evidence$$

در این رابطه عبارت $P(c|x)$ احتمال وقوع رویداد مورد نظر با در نظر گرفتن شرایط خاص را بیان می کند. یا با ادبیات دیگر، این عبارت بیان می‌کند که با فرض مشاهده شدن یک داده (datapoint)، احتمال تعلق داده به کلاس یا برچسب $c$ چقدر است.

عبارت $P(c)$ دانش ما در مورد رویداد های قبلی را نشان می‌دهد. یعنی طبق اطلاعات موجود چه احتمالی وجود دارد که یک رویداد اتفاق بیفتد. یعنی احتمال رخ دادن کلاس $c$ بدون در نظر گرفتن داده‌ها چقدر است.

عبارت $P(x|c)$ بیان می کند با فرض وقوع شرط، چه احتمالی وجود دارد که رویداد اتفاق بیفتد. به عبارت دیگر احتمال اینکه یک داده از کلاس $c$ رخ بدهد چقدر است.

عبارت $P(x)$ نشان دهنده احتمال وقوع شرط است. یعنی چقدر احتمال دارد که یک داده بدون در نظر گرفتن یک کلاس خاص رخ بدهد. (یادآوری: این قسمت در فرمول بیز صرفا برای normalization و تصحیح مقادیر به بازه‌ی ۰ و ۱ استفاده می‌شود)

حال فرض کنید می‌خواهید احتمال وقوع یک سری مشخص از داده‌ها را در یک کلاس مشخص به دست آورید. با فرض **مستقل بودن رخداد هر یک از داده‌ها** می‌توان قانون بیز را به صورت زیر نوشت :

$$P(c|X) = \frac{P(x_۱|c) \times P(x_۲|c) \times \ldots \times P(x_n|c) \times P(c)}{P(X)}$$

که در این رابطه $X$ بردار شروط است که شامل شرط های $x_i|i = ۱, ..., n$ است.

نکته دیگر در این رابطه این است که نیازی نیست برای طبقه‌بندی هر بار رابطه $P(x)$ را محاسبه کنیم زیرا همانطور که اشاره شد، این عبارت فقط برای نرمالایز کردن عبارت است تا خروجی احتمال قانون بیز بین صفر و یک قرار بگیرد. در نتیجه با فرض استقلال شروط می توان قانون بیز را برای طبقه بندی به صورت زیر نوشت :

$$P(c|X) \propto P(x_۱|c) \times P(x_۲|c) \times \ldots \times P(x_n|c) \times P(c)$$

**برای مطالعه بیشتر:**

در شرایط کلی :

$$P(c|X) \neq P(x_۱|c) \times P(x_۲|c) \times \ldots \times P(x_n|c) \times P(c)$$

زیرا با فرض وابستگی شروط، termهای دیگری نیز در فرمول بیز تولید می‌شوند. یعنی نمی‌توان عبارت بالا را به صورت ضرب تک تک شرط‌ها در کلاس نوشت. شرطی که باعث می شود این برابری برقرار شود استقلال شروط است. در این حالت استقلال، طبقه‌بند بیز را Naïve Bayes می‌گوییم.

قانون Naïve Bayes یک تعمیم از قانون بیز است که در آن، به جای یک رویداد، مجموعه‌ای از رویداد های مستقل در نظر گرفته می‌شود. به عبارت دیگر، این قانون به ما اجازه می‌دهد تا احتمال وقوع یک مجموعه از رویدادهای مستقل را با توجه به دانش ما در مورد رویدادهای قبلی و شرایط فعلی برآورد کنیم.

در حالت‌هایی که شروط از یکدیگر مستقل باشند استفاده از Naïve Bayes بسیار کمک کننده است زیرا حجم محاسبات را کاهش می‌دهد و تصمیم‌گیری سریع‌تر و با محاسبات کمتر قابل انجام خواهد بود.

## مثال:

در دوران کرونا آزمایشی طراحی شد که نتایج آن در مطالعات بالینی در بین صد بیمار کرونا و صد فرد سالم به صورت زیر است :

| | بیمار | سالم |
|---|---|---|
| تشخیص کرونا | ۹۵ | ۱۰ |
| تشخیص غیر کرونا | ۵ | ۹۰ |

فرض کنید در یک جامعه، ۶۰ درصد افراد به کرونا مبتلا شده‌اند.

حال اگر فردی توسط آزمایش مبتلا به کرونا تشخیص داده شود، احتمال اینکه در اصل، سالم بوده باشد چقدر است ؟

حال گام به گام مسئله را حل می‌کنیم:

**مرحله اول :** احتمال وقوع رخداد بدون شرط (Prior)

احتمال اینکه فردی از این جامعه مبتلا به کرونا نباشد: در ۴۰ درصد جامعه

**مرحله دوم :** احتمال وقوع شرط در صورت وقوع رخداد (likelihood)

احتمال اینکه فرد آزمایش تشخیص کرونا مثبت داشته باشد به شرط اینکه سالم باشد: از هر ۱۰۰ نفر سالم ۱۰ نفر آزمایش تشخیص کرونا مثبت دارند

**مرحله سوم :** احتمال وقوع شرط (Evidence)

احتمال اینکه تست کرونا مثبت باشد. این احتمال دو قسمت دارد.

افرادی که سالم هستند و تست مثبت دارند: از هر ۱۰۰ فرد سالم ۱۰ نفر تست مثبت دارند.

افرادی که بیمار هستند و تست مثبت دارند : از هر ۱۰۰ نفر بیمار ۹۵ نفر تست مثبت دارند.

در نتیجه احتمال به بالا به صورت زیر محاسبه می‌شود:

$$P(\text{ سالم بودن } | \text{ آزمایش مثبت کرونا }) = \frac{P(\text{ سالم بودن }) P(\text{ آزمایش مثبت کرونا } | \text{ سالم بودن })}{P(\text{ آزمایش مثبت کرونا })}$$

$$P(\text{ سالم بودن } | \text{ آزمایش مثبت کرونا }) = \frac{۰/۱ \times ۰/۴}{۰/۴ \times ۰/۱ + ۰/۶ \times ۰/۹۵} = \frac{۴}{۶۱}$$

## تعریف مسئله

در این مسئله مجموعه داده تعدادی کتاب در فرمت csv در اختیار شما قرار گرفته است. در این داده نام کتاب، توضیحات مربوط به کتاب و همینطور دسته‌بندی (برچسب) کتاب مشخص شده است. در این مجموعه داده تعداد ۶ دسته وجود دارند که به صورت زیر می‌باشند:

مدیریت کسب و کار، رمان، کلیات اسلام، داستان کودک و نوجوانان، جامعه‌شناسی، داستان کوتاه

| | title | description | categories |
|---|---|---|---|
| 0 | ساختار نظریه‌های جامعه‌شناسی ایران | ...ساختار نظریه‌های جامعه‌شناسی ایران» نوشته ابو» | جامعه‌شناسی |
| 1 | جامعه و فرهنگ کانادا | ...جامعه و فرهنگ کانادا» از مجموعه کتاب‌های «جام» | جامعه‌شناسی |
| 2 | پرسش از موعود | ...پرسش‌های مختلفی درباره زندگی و شخصیت امام مهدی | کلیات اسلام |
| 3 | موج، دریا | ...موج دریا» به قلم مهری ماهوتی (-۱۳۴۰) و تصویرگ» | داستان کودک و نوجوانان |
| 4 | پرسش از غرب | ...پرسش از غرب» به قلم دکتر اسماعیل شفیعی سروستا» | جامعه‌شناسی |

شکل ۱: قسمتی از دیتاست

دو فایل در اختیار شما قرار می گیرد. فایل اول با نام books_train.csv که اطلاعات اولیه را در مورد چند کتاب به شما می دهد. فایل دوم با نام books_test.csv در اختیار شما قرار می گیرد که حاوی توضیحات در مورد کتاب است و شما باید با استفاده از توضیحات کتاب مشخص کنید کتاب از چه موضوعی است.

## فاز اول پروژه : پیش‌پردازش داده (۳۰ نمره)

در فاز اول باید اطلاعات متنی داخل مجموعه داده را پیش‌پردازش کنیم. برای این کار می‌توانید از کتابخانه هضم استفاده کنید یا خودتان موارد مورد نیازتان را پیاده سازی کنید.

در این مرحله باید سعی کنید اطلاعات فایل‌ها را به نحوی مدیریت کنید که به بهترین حالت در پروژه استفاده کنید. به طور مثال، یکی از پیشنهادهای اولیه در این مرحله می‌تواند حذف علائم نگارشی و همچنین اعداد از عنوان و توضیحات هر کتاب باشد. زیرا این علائم اطلاعات خاصی در مورد برچسب کتاب به ما نخواهند داد و قابل حذف هستند.

دقت کنید که مرحله پیش‌پردازش باید روی داده‌های هر دو فایل انجام شود. نکته دیگری که باید به آن توجه کنید این است که لزوما اجرای هر نوع پیش پردازشی باعث بالا رفتن دقت پیش‌بینی شما نخواهد شد.

## فاز دوم : حل مسئله (۷۰ نمره)

در این مسئله می‌خواهیم با استفاده از قاعده بیز بر اساس نام و توضیحات موجود برای هر کتاب، تشخیص دهیم که این کتاب در کدام دسته موضوعی قرار می‌گیرد.

در این مسئله از مفهوم Bag of Words استفاده می‌کنیم. همانطور که از نام این روش مشخص است، فرض می‌کنیم مجموعه‌ای از کلمات داریم که بدون توجه به دستور زبان کنار هم قرار گرفته‌اند. به عنوان مثال به دو جمله زیر دقت کنید:

جمله‌ی ۱: من از غذای این رستوران خوشم آمد.

جمله‌ی ۲: غذای رستوران خیلی خوب بود ولی رفتار پرسنل نه.

حال هر واژه یکتا را در نظر می گیریم و تعداد پیشامدهای آن در هر جمله را مشخص می‌کنیم:

| | من | از | غذای | این | رستوران | خوشم | آمد | خیلی | خوب | بود | ولی | رفتار | پرسنل | نه |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| جمله اول | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ | ۰ |
| جمله دوم | ۰ | ۰ | ۱ | ۰ | ۱ | ۰ | ۰ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ | ۱ |

همانطور که در بالا مشاهده می شود یک BoW تشکیل شد که نشان می‌دهد هر واژه در جمله وجود دارد یا خیر. حال فرض کنید جمله اول متعلق به کلاس اول و جمله دوم متعلق به کلاس دوم باشد. اگر تعداد زیادی نمونه از این جملات متعلق به کلاس‌ها یا برچسب‌های مختلف را داشته باشیم، می‌توانیم ماتریس BoW را طوری تشکیل دهیم که بعدا بتوانیم از آن برای پیش‌بینی کلاس یا برچسب جمله‌ها یا نمونه‌های جدید استفاده کنیم.

با توجه به توضیحات BoW ، در فایل books_train.csv هر کلمه را مستقل از جایگاه و ترتیب آن در جمله در نظر گرفته و BoW را تشکیل دهید. در این پروژه BoW بر اساس تعداد تکرار کلمات یکتا بر اساس دسته‌بندی کتاب‌ها مشخص می‌شود. یعنی در نهایت ابعاد ماتریس BoW حاصل به صورت تعداد کلمات یکتا × تعداد موضوعات خواهد بود. از ماتریس به دست آمده در قسمت بعد برای محاسبه احتمال پیشین (Prior) استفاده می‌شود.

حال با استفاده از داده‌های موجود در فایل books_test.csv کلمات هر کتاب را بررسی می کنیم و با توجه به ماتریس BoW که در مرحله قبل پیدا کردید و با استفاده از قاعده بیز احتمال اینکه این کتاب به چه دسته‌ای تعلق داشته باشد را محاسبه می‌کنید.

با استفاده از قاعده بیز برای این مسئله داریم:

$$P(c|X) = \frac{P(X|c)P(c)}{P(X)}$$

۱ـ X : کلماتی که در متن وجود دارد.

۲ـ C : دسته بندی کتاب.

۳ـ $P(X|c)$ : احتمال دیده شدن کلمات موجود در متن (X) در دسته‌بندی (C).

۴ـ $P(c|X)$ : احتمال اینکه کتاب متعلق به دسته‌بندی (C) باشد با فرض اینکه کلمات (X) در متن توضیحات کتاب باشد.

۵ـ $P(c)$ : احتمال اینکه کتابی با دسته‌بندی (C) باشد.

**پرسش:**

۱ـ اگر در توضیحات موجود درباره یک کتاب، با کلمه ای مواجه شوید که در BoW وجود نداشته باشد چه باید کرد؟ احتمال صفر باید در نظر گرفت یا باید آن کلمه را در نظر نگرفت؟

راهنمایی: در مورد روش Additive Smoothing تحقیق کنید. این روش را در پروژه خود پیاده سازی کنید.

۲ـ فرض کنید در شرایطی متن توضیحات یک کتاب طولانی باشد در این صورت با ضرب شدن احتمال‌های هر کلمه چه اتفاقی می‌افتد؟ پیشنهاد شما برای رفع این مشکل چیست؟

راهنمایی:

$$P(Y|X) \propto P(Y) \cdot \prod_{i=1}^{n} P(X_i|Y)$$

$$\log(P(Y|X)) \propto \log(P(Y)) + \sum_{i=1}^{n} \log(P(X_i|Y))$$

# فاز سوم : امتیازی (۴۰ نمره)

**قسمت اول : (۱۵ نمره)**

احتمالا پس از بررسی داده‌های کتاب‌ها به این نتیجه رسیده‌اید که کلماتی وجود دارند که کاملا مشابه یکدیگر هستند ولی به خاطر اتصال ضمایر ، پیشوند و پسوندهای متفاوت در BoW هر کدام جداگانه یک کلمه یکتا در نظر گرفته شده‌اند. حتی ممکن است به کلماتی برخورده باشید که از یک ریشه هستند ولی به شکل‌های متفاوت در متن آمده‌اند. این کلمات معنی‌های خیلی نزدیکی به یکدیگر دارند و اختصاص کلمات یکتا به هر کدام دقت پیش‌بینی را تحت تاثیر قرار می‌دهد.

به همین منظور یکی از راه‌های افزایش دقت در این پروژه تغییر روند حل مسئله است. در این قسمت انتظار می‌رود با استفاده از ریشه کلمات BoW را تشکیل دهید و سپس با استفاده از قانون بیز ژانر کتاب ها را پیش‌بینی کنید.

در نهایت بررسی کنید که این تغییر چقدر باعث بهتر شدن پیش‌بینی شده است.

**راهنمایی:**

یکی از کتابخانه‌های مناسب برای پردازش متون فارسی، کتابخانه هضم است. توابع این کتابخانه در این قسمت کاربردی خواهند بود. در مورد lemmatization و stemming تحقیق کرده و از با استفاده از کتابخانه‌ی هضم سعی کنید دقت پیش‌بینی را افزایش دهید.

**قسمت دوم : (۱۵ نمره)**

یکی دیگر از مشکلاتی که باعث می‌شود دقت پیش‌بینی ما کاهش پیدا کند، وجود کلماتی است که در هر متنی ممکن است وجود داشته باشند. کلماتی نظیر حروف اضافه، ضمایر موصولی، ضمایر ملکی و ...

به عبارت دیگر، برخی از کلمات به طور مکرر در تمام جملات از همه‌ی کلاس‌ها و برچسب‌های مختلف تکرار می‌شوند. یعنی با اینکه احتمال وقوع آنها بالاست، اطلاعاتی در مورد برچسب آن جمله به ما اضافه نمی‌کنند. در نتیجه برای افزایش دقت پیش‌بینی یکی از راه‌حل‌ها می‌تواند حذف این کلمات از BoW باشد.

این راهکار را پیاده‌سازی کرده و نتیجه را با قسمت‌های قبل مقایسه کنید.

**قسمت سوم : (۱۰ نمره)**

میزان افزایش دقت ترکیب دو راهکار قسمت های اول و دوم را بررسی و گزارش کنید.

# نحوه‌ی تحویل

فایل یا فایل‌های py. یا ipython. حاوی کدها و فایل PDF گزارش را در یک فایل زیپ با نام CA#0-STDNo.zip قرار داده و روی سایت درس بارگزاری کنید. دقت کنید که مراحل کدزنی و نتایج شما باید به طور کامل در گزارش بیان شده باشد.

. **موفق باشید**