

به نام خدا



دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر



هوش مصنوعی

پروژه سوم - Naive Bayes Classifier

مهلت تحویل: ۲۰ اردیبهشت ۱۴۰۰

طراحان: ژيوار صورتی، محسن فیاض، امیرعلی عطایی

بهار ۱۴۰۰

مقدمه:

در این پروژه قصد داریم با استفاده از Naive Bayes Classifier به تجزیه و تحلیل کتاب‌ها و موضوعات آن‌ها بپردازیم و سعی کنیم با استفاده از توضیحاتی که در مورد متن کتاب و همین‌طور نام کتاب داریم، موضوع آن را پیش‌بینی کنیم.

معرفی مجموعه داده:

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 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

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

دو فایل در اختیار شما قرار گرفته است که یکی برای آموزش و دیگری برای ارزیابی مدل شما است. فایل مربوط به آموزش مدل به عنوان books_train.csv و همین‌طور فایلی که مربوط به ارزیابی مدل شما است با نام books_test.csv در اختیار شما قرار گرفته است. دقت داشته باشید که تعداد سطرها به ازای هر موضوع دسته‌بندی در هر فایل به صورت متوازن قرار داده شده است و نیازی به یکسان کردن تعداد کتاب‌ها از موضوعات متفاوت که به نام resampling شناخته می‌شود نیست. البته برای مطالعه بیشتر می‌توانید این موضوع را نیز در نظر بگیرید. این کار برای از بین بردن bias موجود در داده‌هایی که تعداد کلاس‌های خروجی آن‌ها با هم برابر نیست استفاده می‌شود.

فاز اول: پیش‌پردازش داده

در فاز اول باید اطلاعات متنی داخل مجموعه داده را برای تحلیل های بعدی پیش‌پردازش کنیم. برای این کار می‌توانید از کتابخانه‌ی **هضم**¹ استفاده کنید یا خودتان موارد مورد نیازتان را پیاده سازی کنید. شما باید متن و تیتراژهایی که موجود است را تا حد ممکن Normalize کنید. (روش های ممکن، شامل حذف کلمات پرتکرار یا همان stop words، تبدیل کلمات به ریشه آنها و ... است.)

دقت کنید که این کار هم روی داده های train و هم روی داده های test باید انجام شود و لزوماً اجرای هر نوع پیش‌پردازی باعث بالا رفتن دقت مدل شما نخواهد شد. روش های متفاوت را با استفاده از کتابخانه یا بدون آن امتحان کنید و ترکیب هر کدام از آنها که به مدل شما بیشتر کمک می‌کند را اجرا کنید.

البته به جز موارد توضیح داده شده می‌توانید تنها به حذف ایست واژه ها و کاراکترهای بی اهمیت مانند \n و \r بسنده کنید. اما لازم است تا تاثیر انواع دیگر پیش پردازش ها را نیز مشاهده کنید و در گزارش خود توضیحی در مورد آنها ارائه دهید.

1) در گزارش کار خود، جایگزین کردن کلمات با روش stemming یا lemmatization را توضیح دهید.

فاز دوم: فرآیند مسئله

در این مسئله می‌خواهیم با استفاده از Naive Bayes بر اساس نام و توضیح موجود برای هر کتاب، تشخیص دهیم که این کتاب، در کدام یک از موضوعات مربوطه جای می‌گیرد. در این مسئله از مدل bag of words استفاده می‌کنیم. به این صورت که هر کلمه را مستقل از جایگاه و ترتیب آن در جمله در نظر می‌گیریم. feature های این مسئله را تعداد هر کلمه در کلاس مربوطه در نظر بگیرید. یعنی هر چه تعداد یک کلمه در یک کلاس بیشتر باشد، احتمال اینکه آن کلمه متعلق به آن کلاس باشد بیشتر است. برای حل این مسئله به صورت کلی از naive bayes استفاده می‌کنیم که مفهوم پشت آن با توجه به مفاهیم احتمالی زیر قابل بحث است.

¹ <https://github.com/sobhc/hazm>

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

Likelihood
Class Prior Probability
Posterior Probability
Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

2) در گزارش کار خود، توضیح دهید که هر کدام از (evidence, likelihood, prior, posterior) بیانگر چه

مفهومی در این مسئله هستند و چگونه محاسبه می‌شوند.

دقت کنید که نیازی نیست عبارت Evidence در مخرج کسر به صورت مستقیم محاسبه شود.

فرآیند کلی‌ای که باید انجام دهید به این شکل است که در ابتدا برای متن‌هایی که در اختیار دارید تعداد هر کلمه را به تفکیک کلاس آن پیدا کنید. با این کار به نوعی مدل خود را train کرده‌اید. حال برای بررسی یک متن جدید از naive bayes استفاده کنید و با استفاده از احتمال قبلی که در مورد هر کلاس داشته‌اید و همینطور استفاده از کلمات موجود در متن و احتمال دیده شدن آن‌ها در آن کلاس احتمال اینکه متن برای کلاس بخصوصی باشد را بیابید.

برای محاسبات، می‌توانید دو ستون نام و توضیحات را یکی کنید که در آن صورت یک ستون برای داده خواهید داشت و یک ستون برای دسته‌بندی موضوع. همینطور راه دیگری که وجود دارد این است که می‌توانید یکی از دو ستون نام یا توضیحات را کاملاً نادیده بگیرید. دو راه گفته شده تنها مثال‌هایی برای ایده دادن به شما هستند و شما می‌توانید کارهای دیگری نیز برای یکی کردن داده‌های در دست و همینطور استفاده از آن‌ها به شکل‌های متفاوت انجام دهید. در این مورد می‌توانید با عنوان مهندسی ویژگی‌ها² اطلاعات بیشتری نیز کسب کنید و در این پروژه استفاده کنید. فقط توجه کنید که در بخش ارزیابی، باید دقت شما روی داده‌ی تست از حداقل گفته شده بیشتر باشد.

² Feature Engineering

Bigrams

نکته‌ای که در مورد فرآیند ابتدایی naive bayes در قسمت قبل وجود دارد این است که در این مدل، وجود هر کلمه را به تنهایی و بدون توجه به ترتیب کلمات و همینطور دیگر نکات مربوط به بافت³ متن در نظر می‌گیریم، در حالی که همانطور که مشهود است نکات گفته شده می‌توانند تاثیر گذار باشند.

در مورد مشکلی که در این قسمت بیان شد، می‌توان گفت اشکال در فرآیندی است که با استفاده از آن token ها را از متن داده شده بیرون می‌کشیم و هر کلمه را به خودی خود بررسی می‌کنیم. در مقابل این کار می‌توان هر دو کلمه که پشت هم آمده‌اند را یک token در نظر گرفت. به بیان دیگر به جای استفاده از unigram هایی که در قسمت قبل در نظر گرفتیم، در این قسمت از bigram ها استفاده می‌کنیم.

(3) دو جمله مثال بزنید که یک کلمه یکسان در آنها دو معنی متفاوت داشته باشد. استفاده از bigram ها

چطور به مشخص شدن معنی آن کلمه کمک می‌کند؟ آیا bigram برای مشخص کردن معنی کلمه در

مثال شما کافیت یا نیاز به n-gram طولانی‌تری هست؟

اختیاری: استفاده از ترکیب bigram با unigram را در مدل خود اعمال کنید و آن را روی داده‌های خود train کنید و نتیجه بدست آمده را گزارش کنید. دقت کنید طراحی خود را به شکلی انجام دهید که استفاده از هر دو مدل تغییرات زیادی را در کد شما ایجاد نکند به گونه‌ای که اگر خواستید token های ۳ کلمه‌ای و یا حتی ۴ کلمه‌ای را نیز در نظر بگیرید، تغییرات زیادی نیاز نباشند.

³ Context

Additive Smoothing

مشکلی که ممکن است در بدست آوردن موضوعات به آن برخورد کنید این است که در کتاب‌هایی که مربوط به موضوع مشخصی هستند، کلمه‌ای وجود داشته باشد که در کتاب‌های دسته‌ای دیگر نباشد و بالعکس، یا حتی به کلمه‌ای در کتاب جدیدی که می‌خواهیم بررسی کنیم برخورد کنیم که در هیچ کدام از کتاب‌های دیده شده در داده train وجود نداشته باشد.

مشکلی که در حالت گفته شده ایجاد خواهد شد به این مسئله خواهد برگشت که اگر به عنوان مثال کلمه‌ی "جامعه" تنها در کتاب‌های مربوط به جامعه‌شناسی باشد ولی در کتاب‌های مربوط به موضوعات دیگر نباشد، مدل ایجاد شده با قطعیت تشخیص می‌دهد که هر کتابی که در متن آن کلمه "جامعه" وجود دارد مربوط به کتاب‌های موضوع جامعه‌شناسی است در حالیکه نتیجه‌گیری انجام شده لزوماً درست نمی‌باشد.

- 4) در گزارش خود با در نظر داشتن naive bayes توضیح دهید چرا این اتفاق رخ می‌دهد.
- 5) درباره روش Additive Smoothing تحقیق کنید و با پیاده‌سازی آن در پروژه، این مشکل را برطرف کنید. در گزارش خود این روش را توضیح دهید و بگویید دقیقاً چطور به حل این مشکل کمک می‌کند.
- (در بخش ارزیابی، تفاوتی که استفاده از این روش بر دقت می‌گذارد را باید گزارش کنید)

بررسی صحت

- 6) با توجه به تعداد کلمات دیده شده مربوط به هر موضوع، شش عدد bar plot رسم کنید که نشان دهد در کتاب‌های هر موضوع چه کلماتی بیشترین تکرار را دارند. (6 نمودار که در هر کدام حداقل 5 کلمه با بیشترین تکرار در آن دسته آمده است.)
- شرط گفته شده مربوط به بالا بودن تعداد تکرار، تنها یکی از راه‌هایی است که می‌توان کلماتی با بیشترین تاثیر در هر موضوع را شناسایی کرد. تنها نکته‌ای که خوب است در نظر گرفته شود این است که بعضی کلمات در تمامی موضوعات تعداد تکرار بالایی دارند که در نتیجه آن باعث می‌شود تاثیری در شناسایی موضوع برای کتاب نداشته باشند. در این مورد، حذف این کلمات از دایره کلمات می‌تواند گزینه خوبی باشد.

فاز سوم: ارزیابی

برای ارزیابی مدل خود باید از 4 معیار زیر استفاده کنید.

$$Accuracy = \frac{Correct\ Detected}{Total}$$

$$Precision = \frac{Correct\ Detected\ Class}{All\ Detected\ Class\ (Including\ Wrong\ Ones)}$$

$$Recall = \frac{Correct\ Detected\ Class}{Total\ Class}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Correct Detected Class: تعداد کتاب‌هایی که به درستی در موضوع مورد نظر تشخیص داده شده‌اند.

All Detected Class: تعداد تمام کتاب‌هایی که در موضوع مورد نظر تشخیص داده شده. (حتی اگر به اشتباه)

Total Class: تعداد تمام کتاب‌هایی که در مجموعه داده تست در آن موضوع خاص بودند.

به جای Class می‌توانید هرکدام از دسته‌بندی‌های موضوعی موجود را بگذارید.

(7) در گزارش کار خود توضیح دهید که چرا مقدار Precision و Recall هر کدام به تنهایی برای ارزیابی مدل کافی نیست؟ برای هر کدام مدلی را مثال بزنید که در آن، این معیار مقدار بالایی دارد ولی مدل خوب کار نمی‌کند.

(8) در گزارش کار خود توضیح دهید معیار F1 از چه نوع میانگین‌گیری بین precision و recall استفاده می‌کند؟ تفاوت آن نسبت به میانگین‌گیری عادی چیست و در اینجا چرا اهمیت دارد؟

(9) با توجه به اینکه مسئله ما بیشتر از ۲ کلاس دارد در مورد multi-class metrics تحقیق کنید. در گزارش کار خود، سه حالت میانگین‌گیری macro و micro و weighted را شرح دهید. برای تحقیق می‌توانید از این [سایت](https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-cbe8b2c2ca1)⁴ استفاده کنید.

⁴ <https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-cbe8b2c2ca1>

مدل خود را که با استفاده از naive bayes و براساس داده‌ی train ساخته‌اید، روی داده‌ی تست که در اختیارتان قرار دارد اجرا کنید و برای هر کدام از سطرهای آن تشخیص مدل‌تان را بدست آورید. سپس براساس آن، معیارهای بالا را برای هر کلاس به صورت تنها و سپس با استفاده از سه نوع میانگین‌گیری گفته شده برای تمام کلاس‌ها محاسبه کنید. (از کتابخانه‌ها برای محاسبه معیارها نیاید استفاده شود اما برای مطمئن شدن از محاسباتتان می‌توانید از توابعی مثل [classification_report](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html)⁵ استفاده کنید.)

مقدار accuracy و Macro F1 در حالت a باید به ترتیب بیش از 75 و 75 باشند.

(10) در گزارش خود، معیارها را برای حالت‌های زیر به دست آورید

(a) نتایج با استفاده از Additive Smoothing

(b) نتایج بدون استفاده از Additive Smoothing

(11) در گزارش خود، مقادیر بدست آمده در بخش قبل را تحلیل کنید.

(12) در گزارش خود 5 مورد از کتاب‌هایی که در داده‌ی تست هستند و مدل شما موضوع اشتباهی برای آن‌ها

تشخیص داده است بیاورید. به نظر شما چه بخش یا بخش‌هایی از راه حلی که پیش گرفتیم باعث شده این

موارد اشتباه تشخیص داده شوند؟

⁵ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

نکات پایانی:

- دقت کنید که هدف پروژه تحلیل نتایج است بنابراین از ابزارهای تحلیل داده بطور مثال نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید.
- نتایج و گزارش خود را در یک فایل فشرده با عنوان `AI_CA3_<#SID>.zip` تحویل دهید. محتویات پوشه باید شامل فایل `jupyter-notebook`، خروجی `html` و فایل های مورد نیاز برای اجرای آن باشد. توضیح و نمایش خروجی های خواسته شده بخشی از نمره این تمرین را تشکیل می دهد. از نمایش درست خروجی های مورد نیاز در فایل `html` مطمئن شوید.
- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت توسط ایمیل با طراحان در ارتباط باشید.
- هدف از تمرین، یادگیری شماسست. لطفا تمرین را خودتان انجام دهید.