

گزارش پروژه پایانی درس هوش مصنوعی

دسته بندی متون با استفاده از دسته بند بیز

و مدلهای Unigram و Bigram

حسین محمدی

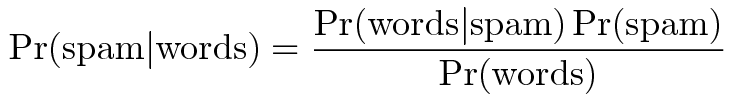
9533081

بهمن 1398

در این پروژه قصد داریم با استفاده از داده های آموزشی (HAM-Train.txt) که در آن هر متن به همراه کلاسش آمده ، کلاس متون داده های آزمایشی (HAM-Test.txt) را حدس زده و تشخیص دهیم.

برای این کار از قانون بیز استفاده میکنیم. در ادامه توضیح این قانون را میبینیم.

در حوزه یادگیری ماشین، تکنیک و روش (Naive Bayes Classifiers) با بکارگیری قضیه بیز و فرض استقلال بین متغیرها، به عنوان عضوی از خانواده «دسته‌بندهای برمبنای احتمال» (Probabilistic Classifiers) قرار می‌گیرد. در سال‌های ۱۹۶۰ تحقیق و بررسی‌های زیادی پیرامون بیز ساده بخصوص در زمینه «بازیابی متن» (Text Retrieval) صورت گرفت و حتی امروز هم به عنوان ابزاری برای «دسته‌بندی متن» (Text Categorization) برای حل مسائلی مانند تشخیص «هرزنامه‌ها» (Spam Mails) به کار می‌رود. معمولا این کار به کمک برآورد تابع احتمال و از طریق فراوانی یا فراوانی نسبی کلمات در اسناد متنی صورت می‌گیرد.

برای مثال ما به دنبال محاسبه مقدار زیر هستیم. 

به این ترتیب به کمک حداکثرسازی تابع درستنمایی (Likelihood maximization)، برآورد پارامترهای مدل میسر می‌شود. در حوزه آمار و دانش رایانه، مدل بیز ساده با نام‌های دیگری نظیر «بیز ساده» (Simple Bayes) و «بیز مستقل» (Independence Bayes) نیز شناخته می‌شود که در بسیاری از حوزه‌های دیگر نیز کاربرد دارد.

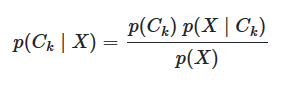
بسیاری از روش‌های به کار گرفته شده در یادگیری ماشین، از تکنیک‌های آماری بهره می‌برند. دسته بند بیز ساده نیز همانطور که از اسمش بر می‌آید از این قاعده مستنثنی نیست. هر چند که تکنیک دسته‌بند بیز ساده از قضیه بیز به منظور تفکیک احتمالات استفاده می‌کند ولی نمی‌توان آن را یک «استنباط بیزی» (Bayesian Inference) در نظر گرفت.

**دسته‌ بند بیز ساده**

اغلب به عنوان یک راهکار ساده برای دسته‌بندی و تعیین روشی برای تشخیص برچسب اشیاء یا نقاط از تکنیک دسته‌بند بیز استفاده می‌شود. برای به کارگیری دسته‌بند بیز ساده، الگوریتم یکتایی وجود ندارد در عوض خانواده‌ای از الگوریتم‌ها موجود است که با فرض استقلال ویژگی‌ها یا متغیرها نسبت به یکدیگر عمل می‌کنند.

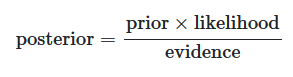
بیز ساده را می‌توان یک مدل برمبنای احتمال شرطی در نظر گرفت. فرض کنید X=(x1,…,xn) برداری از n ویژگی‌ را بیان کند که به صورت متغیرهای مستقل هستند. به این ترتیب می‌توان احتمال رخداد  یعنی p(Ck∣x1,…,xn) را به عنوان یکی از حالت‌های کلاس رخدادهای مختلف به ازاء k

های متفاوت، به شکل زیر نمایش داد.



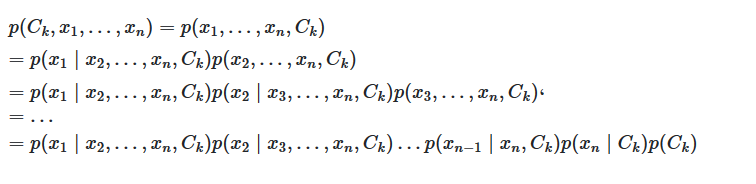
رابطه ۱

همانطور که دیده می شود رابطه بالا همان قضیه بیز است. به عنوان یادآوری قضیه بیز را براساس احتمالات پیشامدهای «پیشین» (Prior)، «پسین» (Posterior)، «درستنمایی» (Likelihood) و «شواهد» (Evidence) در رابطه زیر بازنویسی می‌کنیم.

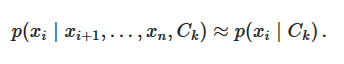


به این ترتیب برای محاسبه احتمال p(∣x1,…,xn)

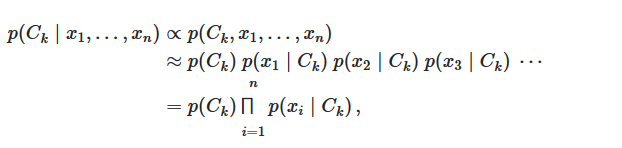
کافی است از «احتمال توام» (Joint Probability) کمک بگیریم و به کمک احتمال شرطی با توجه به استقلال متغیرها، آن را ساده کنیم.



با فرض استقلال مولفه‌ها یا ویژگی‌های xiها از یکدیگر می‌توان احتمالات را به شکل ساده‌تری نوشت. کافی است رابطه زیر را در نظر بگیریم.



به این ترتیب احتمال توام را به صورت حاصلضرب احتمال شرطی می‌توان نوشت.

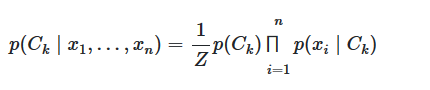
رابطه ۲

**نکته**: در رابطه ۱ مخرج کسر در همه محاسبات یکسان و ثابت است در نتیجه می‌توان احتمال شرطی را متناسب با احتمال توام در نظر گرفت. در رابطه بالا این تناسب را با علامت ∝نشان داده‌ایم.

با توجه به نکته گفته شده، و رابطه ۲ می‌توانیم احتمال شرطی معرفی شده در رابطه ۱ را به صورت زیر بدست آوریم. در نتیجه احتمال تعلق یک مشاهده به دسته یا گروه Cـk

با توجه به مشاهدات X

مطابق با رابطه زیر مشخص خواهد شد.



توجه داشته باشید که در اینجا احتمال شواهد (مشاهدات) به صورت 

در نظر گرفته شده است. واضح است که Z به شواهد و مشاهدات x1,…,xn وابسته است.

**ساخت دسته‌بند براساس مدل احتمالاتی**

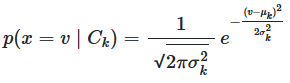
در قسمت قبل با نحوه محاسبه مدل احتمالاتی بیز آشنا شدید. اما در این بخش به کمک «قواعد تصمیم» (Decision Rule)، دسته‌بند بیز را ایجاد و کامل می‌کنیم. یکی از اساسی‌ترین قواعد تصمیم، انتخاب فرضیه محتمل‌تر است. به این ترتیب از بین تصمیمات مختلف، آن کاری را انجام می‌دهیم که براساس شواهد جمع‌آوری شده، بیشترین احتمال رخداد را دارد. این قاعده را «حداکثر پسین» (Maximum Posterior) یا به اختصار MP می‌نامند. به این ترتیب دسته بند بیز را می‌توان به صورت تابعی از تصمیمات  در نظر گرفت که بوسیله تابع  ‌ تخمین زده می‌شود. حداکثرسازی این تابع را به صورت زیر نشان می‌دهیم.



در نتیجه با توجه به توزیع‌های مختلفی که ممکن است نمونه تصادفی داشته باشد (نمونه از آن جامعه آمده باشد) یعنی  می‌توان پارامترها را محاسبه و یا برآورد کرد.

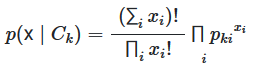
**دسته بند بیز ساده گاوسی (Gaussian Naive Bayes)**

اگر مشاهدات و داده‌ها از نوع پیوسته باشند، از مدل احتمالی با توزیع گاوسی یا نرمال برای متغیرهای مربوط به شواهد می‌توانید استفاده کنید. در این حالت هر دسته یا گروه دارای توزیع گاوسی است. به این ترتیب اگر kدسته یا کلاس داشته باشیم می‌توانیم برای هر دسته میانگین و واریانس را محاسبه کرده و پارامترهای توزیع نرمال را برای آن‌ها برآورد کنیم. فرض کنید که  میانگین و  واریانس دسته kام یعنی  باشد. همچنین  را مشاهدات حاصل از متغیرهای تصادفی X در نظر بگیرید. از آنجایی که توزیع X در هر دسته گاوسی (نرمال) فرض شده است، خواهیم داشت:

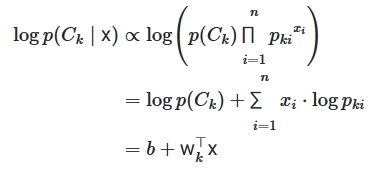


**دسته بند بیز ساده چندجمله‌ای (Multinomial Naive Bayes)**

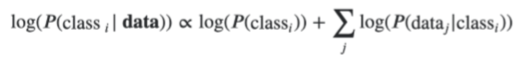
بیز ساده چندجمله‌ای، به عنوان یک دسته‌بند متنی بسیار به کار می‌آید. در این حالت برحسب مدل احتمالی یا توزیع چند جمله ای ، برداری از nویژگی برای یک مشاهده به صورت X=(x1,…,xn) با احتمالات  (p1,…,pn) در نظر گرفته می‌شود. مشخص است که در این حالت بردار X بیانگر تعداد مشاهداتی است که ویژگی خاصی را دارا هستند. به این ترتیب تابع درستنمایی در چنین مدلی به شکل زیر نوشته می‌شود.



اگر مدل بیز ساده را براساس لگاریتم تابع درستنمایی بنویسیم، به صورت یک دسته‌بند خطی درخواهد آمد.



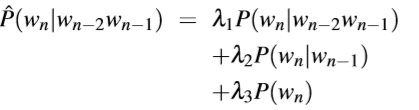
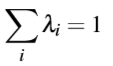
که در مسئله ما فرمول بالا به شکل زیر در می آید و از این روش استفاده میکنیم.



که dataj در مدل یونیگرام یک کلمه و در مدل بیگرام یک زوج کلمه پشت سرهم خواهد بود.

برای افزایش دقت و همچنین در نظر گرفتن کلمه یا زوج کلمه هایی که در داده آموزش ما نیست از روش هموار سازی backoff استفاده میکنیم.

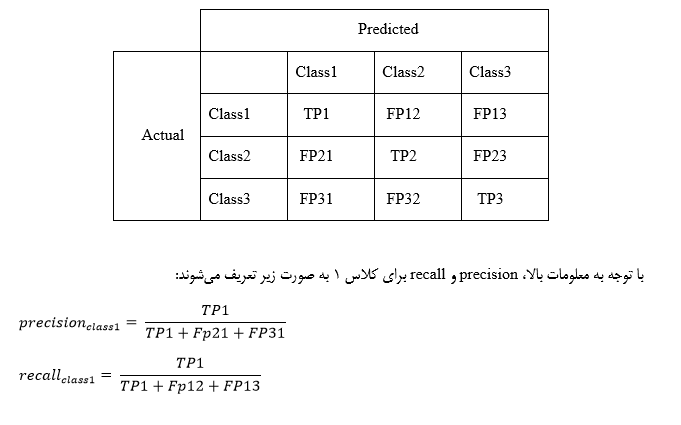
شایان ذکر است ما فقط از بیگرام همراه با بک آف یونیگرام استفاده میکنیم. یعنی یونیگرام تنها برای استفاده در هموارسازی استفاده میشود.بک آف برای مدل تریگرام به شکل زیر است.

حال به ازای لامبداهای گوناگون دقت خروجی متفاوت خواهد بود. بدین منظور مقادیر گوناگون را بررسی میکنیم.فرمول بک آف پیاده سازی شده به شکل زیر است.

prob = Lambda1 \* unigram\_prob + Lambda2 \* bigram\_prob

روش محسابه Recall و Precision و F-measure برای هر کلاس به شکل زیر است.





حال مقادیر گوناگون

Lambda1 = 0.15  
Lambda2 = 0.85

Total Accuracy is: 92.55813953488372

Recall Precision F-measure

0.962 0.9398 0.9507 اقتصاد

0.92 0.8638 0.891 سیاسی

0.7931 1.0 0.8846 ادب و هنر

0.8466 0.8734 0.8597 اجتماعی

0.9868 0.9911 0.9889 ورزش

Lambda1 = 0.35  
Lambda2 = 0.65

Total Accuracy is: 92.55813953488372

Recall Precision F-measure

0.9668 0.94 0.9532 اقتصاد

0.92 0.8558 0.8867 سیاسی

0.8275 1.0 0.9056 ادب و هنر

0.8282 0.8823 0.8543 اجتماعی

0.9868 0.9911 0.9889 ورزش

Lambda1 = 0.8  
Lambda2 = 0.2

Total Accuracy is: 91.86046511627907

Recall Precision F-measure

0.9573 0.9266 0.9416 اقتصاد

0.91 0.8465 0.8771 سیاسی

0.8793 0.9622 0.9188 ادب و هنر

0.7975 0.8783 0.8359 اجتماعی

0.9868 0.9955 0.9911 ورزش

Lambda1 = 0.3  
Lambda2 = 0.7

Total Accuracy is: 92.67441860465117

Recall Precision F-measure

0.9668 0.94 0.9532 اقتصاد

0.92 0.8598 0.8888 سیاسی

0.8275 1.0 0.9056 ادب و هنر

0.8343 0.8831 0.858 اجتماعی

0.9868 0.9911 0.9889 ورزش

مشاهده ماکسیمم دقت در حالت زیر بدست می آید.

Lambda1 = 0.3  
Lambda2 = 0.7