گزارش پروژه پایانی درس هوش مصنوعی
دسته بندی متون با استفاده از دسته بند بیز
و مدلهای Unigram و Bigram

حسین محمدی
۹۵۳۳۰۸۱

بهمن ۱۳۹۸

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.

$$\Pr(\text{spam}|\text{words}) = \frac{\Pr(\text{words}|\text{spam})\Pr(\text{spam})}{\Pr(\text{words})}$$

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 مستثنی 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 nویژگی را بیان کند که به صورت متغیرهای مستقل هستند. به این ترتیب می‌توان احتمال رخداد C_k یعنی p(Ck|x1,...,xn) را به عنوان یکی از حالت‌های کلاس رخدادهای مختلف به ازاء k های متفاوت، به شکل زیر نمایش داد.

$$p(C_k \mid X) = \frac{p(C_k)\, p(X \mid C_k)}{p(X)}$$

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.

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

به این ترتیب برای محاسبه احتمال$p(C_k|x1,...,xn)$

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

$$\begin{aligned}
&p(C_k, x_1, \ldots, x_n) = p(x_1, \ldots, x_n, C_k) \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k) p(x_2, \ldots, x_n, C_k) \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k) p(x_2 \mid x_3, \ldots, x_n, C_k) p(x_3, \ldots, x_n, C_k) \\
&= \ldots \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k) p(x_2 \mid x_3, \ldots, x_n, C_k) \ldots p(x_{n-1} \mid x_n, C_k) p(x_n \mid C_k) p(C_k)
\end{aligned}$$

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

$$p(x_i \mid x_{i+1}, \ldots, x_n, C_k) \approx p(x_i \mid C_k)\,.$$

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

$$\begin{aligned} p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &\approx p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\ &= p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,, \end{aligned}$$

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 k-C

با توجه به مشاهدات X

مطابق با رابطه زیر مشخص خواهد شد.

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

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

$$Z = p(\mathrm{x}) = \sum_{k} p(C_k)\, p(\mathrm{x} \mid C_k)$$

در نظر گرفته شده است. واضح است که Zبه 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 C_k در نظر گرفت که بوسیله تابع $\hat{y}$ تخمین زده می‌شود. حداکثرسازی این تابع را به صورت زیر نشان می‌دهیم.

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\; p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

در نتیجه با توجه به توزیع‌های مختلفی که ممکن است نمونه تصادفی داشته باشد (نمونه از آن جامعه آمده باشد) یعنی $p(x_i|C_k)$ 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 اگرk دسته 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 μ_k میانگین و σ_k^2 واریانس دسته kام یعنی C_k باشد. همچنین ν 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:

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

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.

$$p(\mathrm{x} \mid C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i}$$

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

$$\begin{aligned} \log p(C_k \mid \mathrm{x}) &\propto \log\left(p(C_k) \prod_{i=1}^{n} {p_{ki}}^{x_i}\right) \\ &= \log p(C_k) + \sum_{i=1}^{n} x_i \cdot \log p_{ki} \\ &= b + \mathrm{w}_k^\top \mathrm{x} \end{aligned} \qquad w_{ki} = \log p_{ki} \text{ و } b = \log p(C_k)$$

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

$$\log(P(\text{class}_i \mid \mathbf{data})) \propto \log(P(\text{class}_i)) + \sum_j \log(P(\text{data}_j \mid \text{class}_i))$$

که data_j 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.

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1 P(w_n|w_{n-2}w_{n-1}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n)$$

$$\sum_i \lambda_i = 1$$

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.

		Predicted		
		Class1	Class2	Class3
Actual	Class1	TP1	FP12	FP13
	Class2	FP21	TP2	FP23
	Class3	FP31	FP32	TP3

با توجه به معلومات بالا، precision و recall برای کلاس ۱ به صورت زیر تعریف می‌شوند:

$$precision_{class1} = \frac{TP1}{TP1 + Fp21 + FP31}$$

$$recall_{class1} = \frac{TP1}{TP1 + Fp12 + FP13}$$

$$.F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

حال مقادیر گوناگون

```
Lambda1 = 0.15
Lambda2 = 0.85
```

Total Accuracy is: 92.55813953488372

	Recall	Precision	F-measure
اقتصاد	۰/۹۵۰۷	۰/۹۳۹۸	۰/۹۶۲
سیاسی	۰/۸۹۱	۰/۸۶۳۸	۰/۹۲
ادب و هنر	۰/۸۸۴۶	۱/۰	۰/۷۹۳۱
اجتماعی	۰/۸۵۹۷	۰/۸۷۳۴	۰/۸۴۶۶
ورزش	۰/۹۸۸۹	۰/۹۹۱۱	۰/۹۸۶۸

```
Lambda1 = 0.35
Lambda2 = 0.65
```

Total Accuracy is: 92.55813953488372

	Recall	Precision	F-measure
اقتصاد	۰/۹۵۳۲	۰/۹۴	۰/۹۶۶۸
سیاسی	۰/۸۸۶۷	۰/۸۵۵۸	۰/۹۲
ادب و هنر	۰/۹۰۵۶	۱/۰	۰/۸۲۷۵
اجتماعی	۰/۸۵۴۳	۰/۸۸۲۳	۰/۸۲۸۲
ورزش	۰/۹۸۸۹	۰/۹۹۱۱	۰/۹۸۶۸

```
Lambda1 = 0.8
Lambda2 = 0.2
```

Total Accuracy is: 91.86046511627907

	Recall	Precision	F-measure
اقتصاد	۰/۹۴۱۶	۰/۹۲۶۶	۰/۹۵۷۳
سیاسی	۰/۸۷۷۱	۰/۸۴۶۵	۰/۹۱
ادب و هنر	۰/۹۱۸۸	۰/۹۶۲۲	۰/۸۷۹۳
اجتماعی	۰/۸۳۵۹	۰/۸۷۸۳	۰/۷۹۷۵
ورزش	۰/۹۹۱۱	۰/۹۹۵۵	۰/۹۸۶۸

```
Lambda1 = 0.3
Lambda2 = 0.7
```

Total Accuracy is: 92.67441860465117

	Recall	Precision	F-measure
اقتصاد	۰/۹۵۳۲	۰/۹۴	۰/۹۶۶۸
سیاسی	۰/۸۸۸۸	۰/۸۵۹۸	۰/۹۲
ادب و هنر	۰/۹۰۵۶	۱/۰	۰/۸۲۷۵
اجتماعی	۰/۸۵۸	۰/۸۸۳۱	۰/۸۳۴۳
ورزش	۰/۹۸۸۹	۰/۹۹۱۱	۰/۹۸۶۸

مشاهده ماکسیمم دقت در حالت زیر بدست می آید.

```
Lambda1 = 0.3
Lambda2 = 0.7
```