کلاس حل تمرین آمار و احتمال

جلسه اول :۲۸و۲۹ مهر ۱۳۹۷

Machine learning: یکی از شاخه های هوش مصنوعی است که با استفاده از تکنیک‌های آماری، سیستم‌های کامپیوتری را قادر می‌سازد تا بتوانند از روی داده‌ها بدون برنامه ریزی صریح یاد بگیرد.

یکی از taskهای machine learning از منظر کابرد classification است. در classification داده‌های ورودی به دو یا تعداد بیشتری کلاس تقسیم می‌شوند. یادگیرنده‌ی ما باید مدلی را تولید کند تا یک داده‌ی ورودی که تا به حال ندیده است را به کلاس مربوطه ("در مواردی کلاس‌های مربوط") نسبت دهد.

غالبا این عمل با استفاده از یادگیری همراه معلم انجام می‌شود.

یک مثال از کاربردهای classification ، spam filtering است که ورودی سیستم ایمیل هاست و کلاس‌های خروجی spam یا not spam است.

فرض کنید ایمیل جدید آمده وظیفه ما پیش بینی کلاس spam یا not spam است که از ایمیل‌های قبلی (داده های آموزشی یا training set)استفاده می‌کنیم.

spam/not spam=Y	ویژگی‌های متن ایمیل= x

مثال دیگر فرض کنید یک سری عکس از سگ و گربه همراه با برچسب به شما داده می‌شود و پس از آن یک عکس جدید که تا به حال ندیده‌ایم می آید شما باید تعیین کنید متعلق به کلاس سگ است یا گربه

در این جا بردار x ما پیکسل‌های تصویر و y برچسب this is a cat است.

به مثال spam filtering برگردیم :
پس از مشاهده و یادگیری داده‌های آموزشی حال ایمیل جدید آمده است. میخواهیم spam بودن یا نبودن آن را مشخص کنیم. چه احتمالی را باید حساب کنیم ؟

$P(Y=spam|x_{new})$ و $P(Y=not\ spam|x_{new})$ و مقایسه‌ی آنها

در این جا ما می‌خواهیم از الگوریتم Naive Bayes classifier استفاده کنیم. الگوریتم Naive Bayes از جمله روش‌های generative است یعنی احتمال‌های فوق را با اساس قانون بیز و با استفاده از دو احتمال $p(x|y)$ و $p(y)$ که piror نیز نامیده می‌شود به دست می‌اورد. طبق قانون بیز داشتیم :

$$P(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

در نتیجه برای به دست آوردن کلاس هدف در مدل generative مان باید y را بیابیم که مقدار زیر را ماکزیمم می‌کند.

$$\arg\max_y p(y|x) = \arg\max_y \frac{p(x|y)p(y)}{p(x)}$$

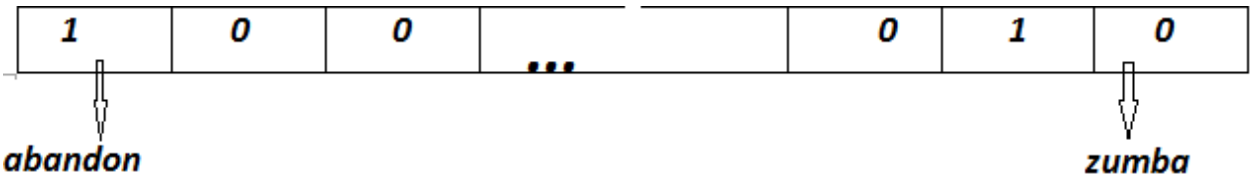
با دقت به رابطه بالا در می‌یابیم که p(x) در مخرج برای یک ایمیل جدید یکسان است و برای مقایسه احتمالات نیازی به محاسبه آن نیست :

$$\arg\max_y p(y|x) = \arg\max_y \frac{p(x|y)p(y)}{p(x)} = \arg\max_y p(x|y)p(y).$$

تا این جا ایده کلی مدل‌های generative در الگوریتم‌های classification گفته شد. در ادامه به طور خاص به الگوریتم Naive Bayes classifier می‌پردازیم . این الگوریتم علی‌رغم طراحی ساده و فرض‌های آسان گرفته شده در بسیاری از موارد حتی از الگوریتم‌های پیچیده در این زمینه عملکرد بهتری دارد . اگر داده‌های ما گسسته باشند آن‌گاه Naive Bayes classifier بر روی آن قابل اجرا خواهد بود.

ساخت یک spam filtering

هر ایمیل را با یک بردار ویژگی به طول دیکشنری نمایش می‌دهیم به عنوان مثال :



هر کلمه‌ای که در متن ایمیل حداقل یک‌بار آمده باشد در بردار ویژگی بعد متناظر با آن مقدار ۱ و در غیر این‌صورت ۰ قرار دارد. فرض کنید دیکشنری ما شامل ۵۰۰۰۰ کلمه باشد در این‌صورت ابعاد برداروِیژگی هر یک از داده‌های ما ۱*۵۰۰۰۰ بوده و فضای حالت ممکن $۲^{۵۰۰۰۰}$ خواهد بود.

برای مدل کردن p(x|y) یک فرض قوی (strong assumption) داریم و آن مستقل شرطی بود x_iها به شرط y است.

x_i's are conditionally independent given y

این فرض را فرض بیز ساده (Naïve Bayes Assumtion) و الگوریتم حاصل را (Naïve Bayes Classifier) می‌نامند.

به عنوان مثال فرض کنید می‌دانیم ایمیل جدید اسپم است (y=1) در این صورت اطلاع از مقدار x_1 (این که کلمه abandon در ایمیل وجو دارد یا خیر) هیچ چیز به دانش ما در باره x_{50000} (وجود یا عدم وجود کلمه Zumba در ایمیل) اضافه نخواهد کرد.

توجه کنید در بسیاری از مسایل دنیا واقعی از جمله همین spam filtering این شرط در واقعیت بر قرار نباشد ولی همچنان الگوریتم فوق می‌تواند جواب های مناسب و قابل قبولی را برگرداند.

دقت نمایید فرض فوق شرط مستقل شرطی بودن را داشته و و الزماً x_i از یکدیگر مستقل نیستند. رابطه‌ی $p(x_{50000})=p(x_{50000}|x_1)$ الزاما برقرار نیست.

طبق فرض Naïve Bayes برای محاسبه احتمال $P(x|y)$ خواهیم داشت :

$$
\begin{aligned}
& p(x_1, \ldots, x_{50000}|y) \\
&= p(x_1|y)p(x_2|y, x_1)p(x_3|y, x_1, x_2) \cdots p(x_{50000}|y, x_1, \ldots, x_{49999}) \\
&= p(x_1|y)p(x_2|y)p(x_3|y) \cdots p(x_{50000}|y) \\
&= \prod_{i=1}^{n} p(x_i|y)
\end{aligned}
$$

در رابطه بالا $P(x_1|y=1)$ برای مساله spam filtering چیست ؟ چگونه محاسبه می‌شود.؟
احتمال وجود کلمه abandon در ایمیل جدید به شرط این که بدانیم ایمیل مذکور از کلاس spam است .
برای محاسبه احتمال فوق به داده‌های آموزشی یا همان تمامی ایمیل‌هایی که تا کنون دیده ایمیل مراجعه میکنیم . در این بین نسبت تعداد ایمیل‌هایی که spam بوده و کلمه abandon در آن‌ها حضور داشته به تعداد کل ایمیل‌های spam را به دست می‌آوریم.

$P(y=1)$ چگونه محاسبه می‌شود؟

الگوریتم Naïve Bayes توضیح داده شده در خیلی از مسائل به خوبی عمل می‌کند ولی یک تغییر کوچک می‌تواند عملکرد آن را بهتر نماید.
فرض کنید یک ایمیل جدید به دست ما رسیده که می‌خواهیم آن را کلاس‌بندی کنیم. ایمیل شامل کلمه‌ی جدید مثلا Bernoulli است. با این که کلمه Bernoulli در دیکشنری ما وجود داشته و فرضا اندیس ۱۰۰ام است ولی در گذشته هیچگاه رخ نداده است. در این صورت محاسبه دو احتمال
$P(y=spam|x_{new})$ و $P(y=not\ spam|x_{new})$ چگونه خواهد بود؟
گفتیم به دلیل یکسان بود x از محاسبه $P(x)$ در رابطه زیر صرف نظر میکنیم .

$$
\arg\max_y p(y|x) = \arg\max_y \frac{p(x|y)p(y)}{p(x)}
$$

در این صورت طبق فرض Naïve Bayes چون $P(x_{100}=1|y=spam)$ و $P(x_{100}=1|y=not\ spam)$ هر دو صفر است هر عبارت یکسان و برابر صفر می شود زیرا یک جمله صفر در ضرب عبارات داریم .مقایسه امکان پذیر نیست و تاثیر کلمات دیگر موجود در ایمیل جدید هم از بین رفته اند.
راه‌حل نوعی هموار سازی است در واقع ما به جای شمارش صفر مقدار کوچک و ثابتی اضافه می‌کنیم تا اثر صفر از بین رود .
می‌توانید این گونه در نظر بگیرید که همیشه یک رکورد بیشتر شمارش میکنیم . در این صورت دیگر به مشکل احتمال صفر بر نخواهیم خورد .