



شماره تمرین:بخش دوم تمرین شماره 2 درس داده کاوی

بخش 1:

اگر فضای نمونه ای ما توسط B1,B2,B3,...,Bn افراز شده باشد بطوریکه احتمال هر کدام از آنها بزرگتر از O باشد آنگاه برای هر پیشامد A می توانیم بنویسیم :

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{P(A)}$$

زيرا بر اساس رابطه احتمال شرطى مى دانيم:

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A|B_j)P(B_j)}{P(A)}$$

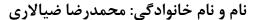
كه نام اين قضيه ، قضيه بيز مي باشد .

: Naive Bayes Classifiers

بیز ساده را میتوان یک مدل برمبنای احتمال شرطی در نظر گرفت. فرض کنید (x1,...,xn)برداری از رویش کنید (x1,...,xn)پرداری از Ck ویژگی را بیان کند که به صورت متغیرهای مستقل هستند. به این ترتیب میتوان احتمال رخداد کلاس رخدادهای مختلف به ازاء الهای متفاوت، به شکل زیر امایش داد:

$$p(C_k \mid X) = rac{p(C_k) \; p(X \mid C_k)}{p(X)}$$

$$\mathsf{P}(\mathsf{X} | \mathsf{Ck}) = \prod_{i=1}^n p(x_i \mid C_k)$$
 : چون مولفه های X از یکدیگر مستقل هستند داریم که





شماره تمرین:بخش دوم تمرین شماره 2 درس داده کاوی

در نتیجه احتمال تعلق یک مشاهده به دسته یا گروه Ck با توجه به مشاهدات Xمطابق با رابطه زیر مشخص خواهد شد.

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

$$Z = p(\mathsf{x}) = \sum\limits_k p(C_k) \ p(\mathsf{x} \mid C_k)$$
 که

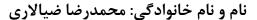
دسته بند بیز ساده گاوسی(Gaussian Naive Bayes):

اگر مشاهدات و دادهها از نوع پیوسته باشند، از مدل احتمالی با توزیع گاوسی یا نرمال برای متغیرهای مربوط به شواهد می توانید استفاده کنید. در این حالت هر دسته یا گروه دارای توزیع گاوسی است. به این ترتیب اگر k دسته یا کلاس داشته باشیم می توانیم برای هر دسته میانگین و واریانس را محاسبه کرده و پارامترهای توزیع نرمال را برای آنها برآورد کنیم. فرض کنید که μ میانگین و $\sigma^2 k$ واریانس دسته μ ام یعنی μ اباشد. همچنین ν را مشاهدات حاصل از متغیرهای تصادفی μ در نظر بگیرید. از آنجایی که توزیع μ در هر دسته گاوسی (نرمال) فرض شده است، خواهیم داشت:

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

دسته بند بيز ساده چندجملهای(Multinomial Naive Bayes) :

بیز ساده چندجملهای، به عنوان یک دستهبند متنی بسیار به کار می آید. در این حالت برحسب مدل احتمالی یا توزیع چند جملهای، برداری از nویژگی برای یک مشاهده به صورت $X=(x_1,...,x_n)$ احتمالات $X=(x_1,...,x_n)$





شماره تمرین:بخش دوم تمرین شماره 2 درس داده کاوی

در نظر گرفته می شود. مشخص است که در این حالت بردار Xبیانگر تعداد مشاهداتی است که ویژگی خاصی را دارا هستند. به این ترتیب تابع درستنمایی در چنین مدلی به شکل زیر نوشته می شود.

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

دسته بند بیز ساده برنولی(Bernoulli Naive Bayes):

در این قسمت به بررسی توزیع برنولی و دستهبندی بیز خواهیم پرداخت. به شکلی این نوع از دستهبند بیز بیشترین کاربرد را در دستهبندی متنهای کوتاه داشته، به همین دلیل محبوبیت بیشتری نیز دارد. در این مدل در حالت چند متغیره، فرض بر این است که وجود یا ناموجود بودن یک ویژگی در نظر گرفته شود. برای مثال با توجه به یک لغتنامه مربوط به اصطلاحات ورزشی، متن دلخواهی مورد تجزیه و تحلیل قرار می گیرد و بررسی میشود که آیا کلمات مربوط به لغتنامه ورزشی در متن وجود دارند یا خیر. به این ترتیب مدل تابع درستنمایی متن براساس کلاس های مختلف Ckبه شکل زیر نوشته می شود.

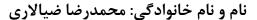
$$p(\mathsf{x}\mid C_k) = \prod\limits_{i=1}^n p_{ki}^{x_i} (1-p_{ki})^{(1-x_i)}$$

كاربردهاي الگوريتم دسته بندي بيز ساده Naive Bayes

از الگوریتم دسته بندی بیز ساده Naive Bayes میتوان استفاده های بی شماری کرد. در اینجا برخی از برنامه های رایج Naive Bayes برای کارهای روزمره وجود دارد:

طبقه بندی اسناد:

این الگوریتم می تواند به شما کمک کند تا تعیین کنید که یک سند مشخص به کدام دسته تعلق دارد. می تواند برای طبقه بندی متون به زبانها، ژانرها یا موضوعات مختلف (از طریق کلمات کلیدی) مورد استفاده قرار گیرد. فیلتر هرزنامه Naive Bayes به راحتی هرزنامه ها را با استفاده از کلمات کلیدی مرتب می کند به همین





شماره تمرین:بخش دوم تمرین شماره 2 درس داده کاوی

دلیل از آنها برای حل مسائلی مانند spam-filtering استفاده می شود. الگوریتم باید برای تشخیص چنین احتمالاتی آموزش ببیند و سپس بتواند آنها را به طور موثر دسته بندی کند.

تحليل احساسات:

بر اساس آنچه احساسات کلمات در متن بیان میشود، Naive Bayesمی تواند احتمال مثبت یا منفی بودن آن را محاسبه کند. به عنوان مثال در بررسی مشتری، "خوب "یا "ارزان" معمولاً به معنای رضایت مشتری است. با این حال، Naive Bayesبه کنایه و کلماتی که به معنای متفاوتی از خودشان به کار میرود حساس نیست.

طبقه بندی تصویر:

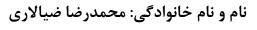
طبقه بندی تصویر برای اهداف شخصی و تحقیقاتی، ساخت طبقه بند Naive Bayesian آسان است، می توان آن را آموزش داد تا از طریق یادگیری ماشین تحت نظارت، ارقام دست نوشته را تشخیص دهد یا تصاویر را در دسته ها قرار دهد.

پیش بینی در لحظه:

Naive Bayesیک طبقه بندی یادگیری مطمئن و سریع است. بنابراین، میتوان از آن برای پیشبینی در زمان واقعی استفاده کرد. این کاربرد بیشتر مربوط به طبقه بیز ساده گوسی می باشد .

پیش بینی چند کلاسه:

این الگوریتم به دلیل ویژگی پیشبینی چند کلاس نیز شناخته شده است. در اینجا میتوانیم احتمال چندین کلاس متغیر هدف را پیش بینی کنیم.





شماره تمرین:بخش دوم تمرین شماره 2 درس داده کاوی

بخش 2,3:

در این بخش پس از بررسی غیر تهی بودن داده ها توابع برای محاسبه دسته بند بیزی ساده گوسی را پیاده سازی کردیم و به بررسی معیار های سنجش دقت پرداختیم . نتایج زیر حاصل شد :

f1_score 0.7516960651289011 precision_score 0.756043956043956 recall_score 0.7510775862068966

بخش 4 و5:

در ادامه با استفتده از پکیج ها دسته بند بیزی ساده گوسی را به کار بردیم و نتایج مشابه حاصل شد . که این اتفاق منطقی است زیرا داده های آموزشی و تست و فرمول های محاسبه یکسان هستند .

f1_score 0.7516960651289011 precision_score 0.756043956043956 recall_score 0.7510775862068966