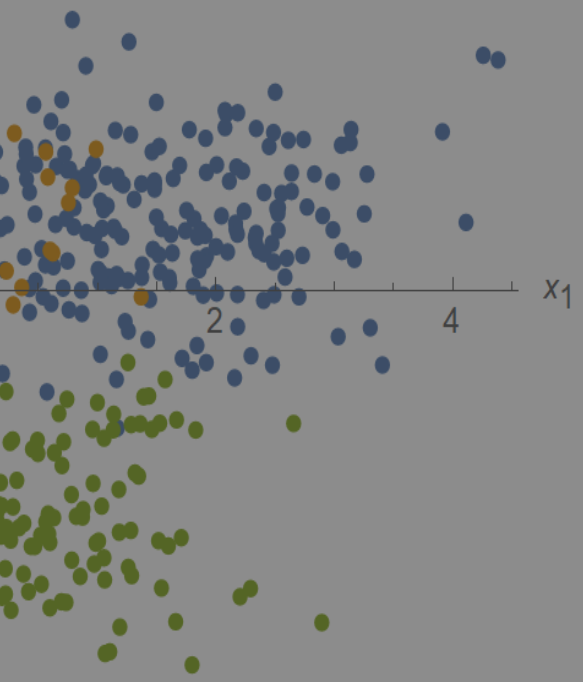
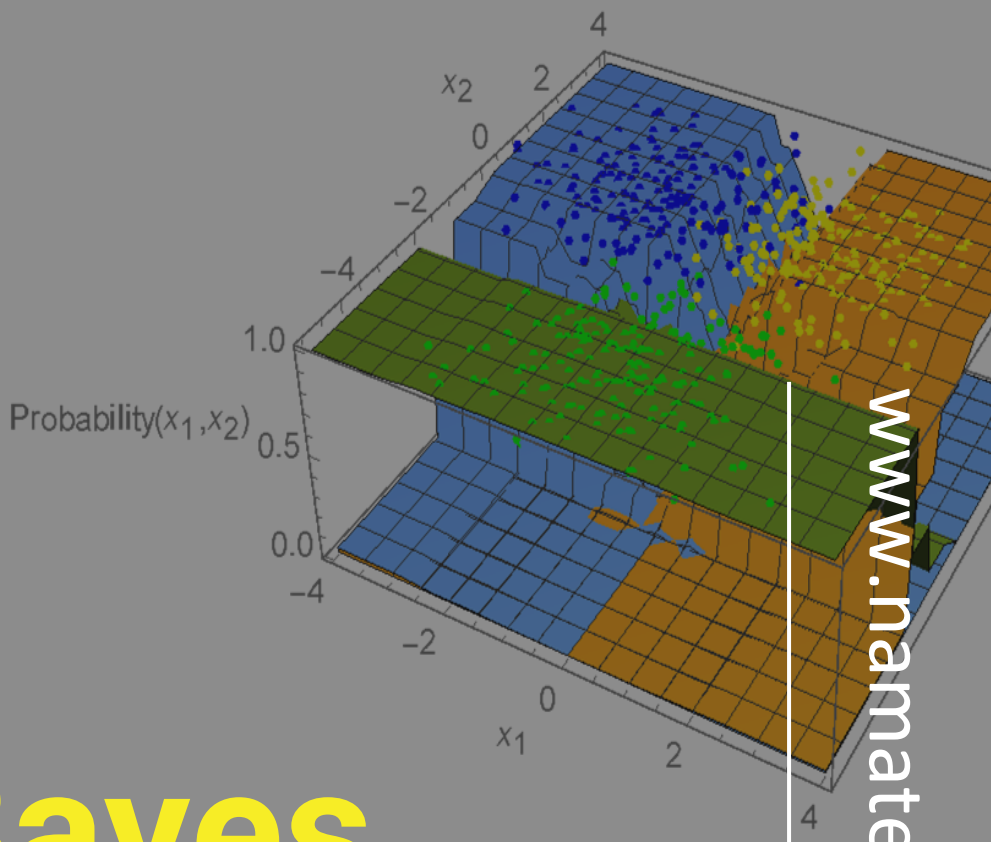
Namatek
True Education





Naive Bayes Algorithm

www.namatek.com

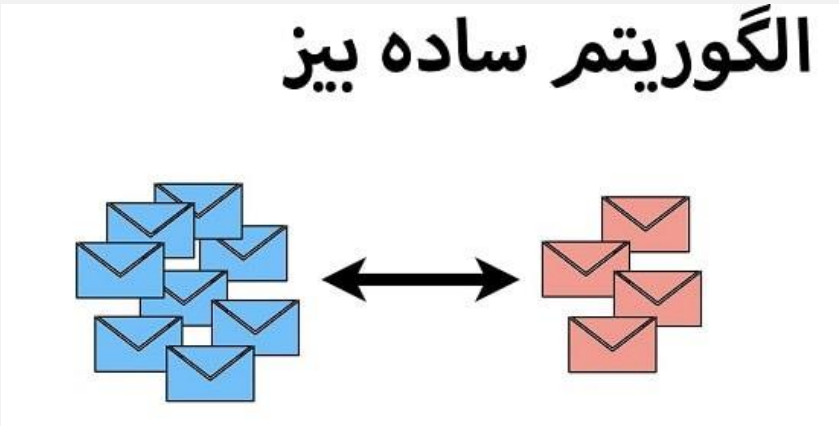
الگوریتم بیز ساده

فهرست مطالب

۱. الگوریتم بیز ساده چیست؟
۲. طبقه بندی کننده الگوریتم بیز ساده چیست؟
۳. چگونگی کارکرد الگوریتم بیز ساده
۴. کاربرد الگوریتم بیز ساده چیست؟
۵. مزایا و معایب الگوریتم بیز ساده
۶. بهبود کارکرد الگوریتم بیز ساده

در دنیای بررسی و تحلیل داده های کلان، نیاز به استفاده از ابزارهای مختلف برای دسته بندی کردن داده ها به چشم می خورد که یکی از مهم ترین آن ها الگوریتم بیز ساده است. این الگوریتم یکی از ده ها روش طبقه بندی داده ها به صورت خودکار و با استفاده از یادگیری ماشینی است که در علومی مانند داده کاوی استفاده می شود. در این مقاله به بررسی الگوریتم بیز ساده، انواع طبقه بندی کننده های این الگوریتم، کاربردهای آن و نکات مهمی در مورد الگوریتم بیز ساده می پردازیم.

الگوریتم بیز ساده چیست؟



الگوریتم بیز ساده بخشی از خانواده الگوریتم های یادگیری مولد است. به این معنا که به دنبال مدل سازی توزیع ورودی های یک کلاس یا دسته خاص است؛ اما برخلاف طبقه بندی کننده های دیگر مانند رگرسیون یا لجستیک، این موضوع که کدامیک از ویژگی های موجود برای تمایز بین کلاس ها مهمتر است را نمی آموزد.

الگوریتم بیز ساده (Naive Bayes Algorithm)، احتمال هر آبجکت (Object)، ویژگی های آن و این که به کدام گروه تعلق دارد را بیان می کند و عمدتاً برای حل مسائل مربوط به طبقه بندی احتمالی استفاده می شود.

برای درک بهتر الگوریتم بیز ساده را با یک مثال عملی توضیح می دهیم.
فرض کنید پروژه علم داده ای دارید که در آن وضعیت زیر حاکم است:
شما در حال کار روی یک مشکل طبقه بندی هستید و مجموعه ای از فرضیه ها و ویژگی هایی را ایجاد کرده اید و اهمیت متغیرها را مورد بحث قرار داده اید.
در عرض یک ساعت، ذینفعان می خواهند اولین برش مدل را ببینند. در این وضعیت، چه خواهید کرد؟ شما صدها هزار نقطه داده و چندین متغیر در مجموعه داده های آموزشی خود دارید.
در چنین شرایطی اگر انتخاب با ما باشد، از الگوریتم بیز ساده استفاده می کنیم؛ زیرا که نسبت به سایر طبقه بندی ها، سریع تر عمل می کند. الگوریتم بیز ساده یک الگوریتم یادگیری نظارت شده، بر اساس قضیه بیز است و عمدتاً در طبقه بندی متن که شامل مجموعه داده های آموزشی با ابعاد بالا است، استفاده می شود. طبقه بندی کننده بیز ساده یکی از ساده ترین و مؤثرترین الگوریتم های طبقه بندی است که به ساخت مدل های یادگیری ماشینی سریع، کمک می کند تا بتوانند، پیش بینی های سریعی را انجام دهند. طبقه بندی احتمالی به این معنا است که طبقه بندی بر اساس پیش بینی احتمال یک آبجکت صورت می پذیرد.
برخی از نمونه های محبوب الگوریتم بیز ساده عبارت اند از:

- فیلتر کردن هرزنامه
- تجزیه و تحلیل احساسات
- طبقه بندی مقالات

علت نام گذاری الگوریتم بیز ساده

الگوریتم بیز ساده، از دو کلمه Naive و Bayes تشکیل شده است.

علت نام گذاری آن را می توان به صورت زیر توصیف کرد:

- **Naive:** به معنای ساده یا ساده لوح است؛ زیرا در این حالت فرض بر آن است که وقوع یک ویژگی مستقل از وقوع سایر ویژگی ها است. مانند این که یک میوه را بر اساس رنگ، مزه و شکل آن تشخیص دهند. یک میوه قرمز رنگ، کروی شکل با مزه ای شیرین به عنوان سیب شناخته می شود. از این رو، هر ویژگی به صورتی جداگانه به تشخیص این که میوه مذبور یک سیب است، بدون توجه به وابستگی آن ها به یکدیگر کمک می کند.
- **Bayes:** بیز نامیده می شود، زیرا به اصل قضیه بیز اشاره و بستگی دارد.

1) قضیه بیز

قضیه بیز که با عنوان قانون بیز هم شناخته می شود؛ به منظور تعیین احتمال یک فرضیه با دانش قبلی استفاده می شود و بستگی به احتمال شرطی دارد. فرمول قضیه بیز به صورت زیر است:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

که در آن:

- **P (A | B):** احتمال پسین است. احتمال فرضیه B روی رویداد مشاهده شده A است.

- **P (B | A):** احتمال شواهد با توجه به این که احتمال یک فرضیه درست است.
- **P (A):** احتمال قبلی است. احتمال فرضیه قبل از مشاهده شواهد را بیان می کند.
- **P (B):** احتمال حاشیه ای است و احتمال شواهد را بیان می کند.

نمونه ای از الگوریتم بیز ساده

به عنوان مثال، به همان مثال سیب بر می گردیم؛ اگر میوه ای قرمز، گرد و عرضی در حدود ۷ یا ۸ سانتیمتر مشاهده کنیم، ممکن است آن را سیب بنامیم. حتی اگر این مسائل به یکدیگر مرتبط باشند، هر یک به ما کمک می کند تا تصمیم بگیریم که شیء مورد نظر احتمالاً سیب است؛ به این دلیل که به آن بیز یا ساده می گوییم. ساخت یک مدل الگوریتم بیز ساده، آسان است و به ویژه برای مجموعه ای از داده های بسیار بزرگ مفید است. در کنار سادگی، این مدل از روش های طبقه بندی پیچیده نیز بهتر عمل می کند. قضیه بیز راهی برای محاسبه احتمال پسین P (c | x) از P (c) ، P (x)و P (x | c) ارائه می دهد؛ به معادله زیر توجه کنید:

$$P(C|x) = \frac{P(x|C)\,P(C)}{P(x)}$$

که در آن:

- **P (c |x):** احتمال پسین طبقه بندی C یا هدف (Target) است با پیش بینی کننده X یا ویژگی ها (Attributes) است.
- **P (c):** احتمال پسین طبقه بندی است.

- **P (x | c):** احتمال طبقه بندی کننده پیش بینی کننده را بیان می کند.
- **P (x):** احتمال پسین پیش بینی کننده است.

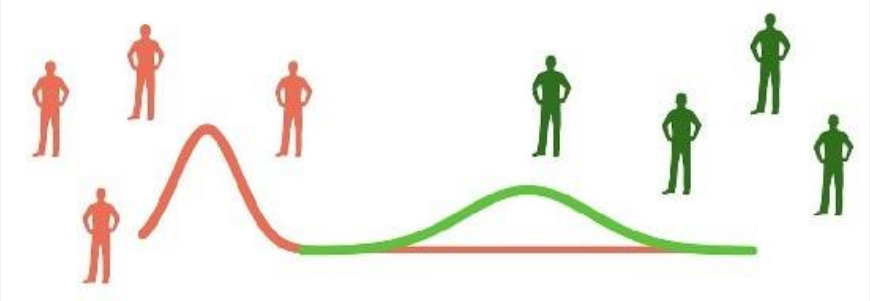
طبقه بندی کننده الگوریتم بیز ساده چیست؟

بیز یک الگوریتم یادگیری ماشینی نظارت شده است که در آن، طبقه بندی کننده به منظور کارهای طبقه بندی مانند طبقه بندی متن استفاده می شود. این نوع الگوریتم ها از اصول احتمال برای انجام وظایف طبقه بندی استفاده می کنند.

انواع طبقه بندی کننده های الگوریتم بیز ساده

فقط یک نوع طبقه بندی ساده بیز وجود ندارد. محبوب ترین انواع طبقه بندی کننده های الگوریتم بیز ساده، بر اساس توزیع مقادیر ویژگی با یکدیگر متفاوت هستند. در ادامه با چند نمونه از انواع طبقه بندی کننده های الگوریتم بیز ساده، آشنا خواهیم شد.

1) Gaussian Naïve Bayes (Gaussian NB)



که با نام الگوریتم بیز گاوسی نیز شناخته می شود و از طبقه بندی کننده های ساده بیز است که با توزیع های گاوسی (یعنی توزیع های نرمال) و متغیرهای پیوسته استفاده می شود. این مدل با یافتن میانگین و انحراف معیار در هر کلاس، تطابق داده می شود.

2) Multinomial NB

که با نام الگوریتم چند وجهی ای ساده بیز شناخته می شود. در این نوع طبقه بندی کننده بیز ساده، فرض می شود که ویژگی ها از توزیع های چند جمله ای هستند. این نوع از طبقه بندی کننده ها، هنگام استفاده از داده های گسسته مانند شمارش فرکانس، مفید است و معمولاً در پردازش زبان طبیعی مانند طبقه بندی هرزنامه، استفاده می شود.

3) Bernoulli Naïve Bayes (Bernoulli NB)

که با نام الگوریتم برنولی ساده بیز شناخته می شود. نوع دیگری از طبقه بندی الگوریتم بیز ساده است که همراه با متغیرهای بولی (Boolean) استفاده می شود. متغیرهای بولی متغیرهای دو مقداری مانند صحیح و غلط (True & False) یا صفر و یک هستند.

4) Optimal Naïve Bayes

که با نام الگوریتم بیز ساده بهینه نیز شناخته می شود. کلاسی را انتخاب می کند که بیشترین احتمال وقوع را دارد. اما از تمامی احتمالات عبور می کند که سبب می شود بسیار کند عمل کند و زمان بر باشد.

چگونگی کارکرد الگوریتم بیز ساده

با طی مراحل زیر می توان یک الگوریتم بیز ساده ایجاد کرد:

۱. مجموعه داده ها را به جدول فرکانس تبدیل کنید.

۲. احتمالات را پیدا کنید و یک جدول احتمال ایجاد کنید.

۳. از معادله بیزی ساده به منظور محاسبه احتمال پسین استفاده کنید.

کاربرد الگوریتم بیز ساده چیست؟

الگوریتم بیز ساده به همراه تعدادی از الگوریتم های دیگر، متعلق به خانواده الگوریتم های داده کاوی هستند که حجم زیادی از داده ها را به اطلاعات مفیدی تبدیل می کنند. برخی از کاربردهای الگوریتم بیز ساده به صورت زیر است.

فیلتر کردن هرزنامه

طبقه بندی هرزنامه یکی از محبوب ترین برنامه های کاربردی الگوریتم بیز ساده است.

طبقه بندی اسناد

طبقه بندی اسناد و متن دست در دست هم دارند. یکی دیگر از موارد استفاده رایج از الگوریتم بیز ساده، طبقه بندی محتوا است. طبقه بندی محتوای یک وب سایت رسانه خبری را تصور کنید؛ تمامی طبقه بندی های محتوا را می توان بر اساس هر مقاله در سایت و بر طبق یک طبقه بندی موضوعی خاص، دسته بندی کرد.

تجزیه و تحلیل احساسات

این کاربرد، شکل دیگری از طبقه بندی متن است؛ تحلیل احساسات معمولاً در بازاریابی و به منظور درک بهتر و کمی کردن نظرات و نگرش ها در مورد محصولات و مارک های خاص استفاده می شود.

پیش بینی وضعیت های ذهنی

با استفاده از داده های بیز ساده برای پیش بینی حالات شناختی مختلف در بین انسان ها و کمک به درک بهتر حالات شناختی مورد استفاده قرار می

گیرد. هدف استفاده از الگوریتم بیز ساده در این مورد، تحقیقات پنهان به ویژه در میان بیماران آسیب دیده مغزی است.

تشخیص چهره

به عنوان یک طبقه بندی به منظور شناسایی چهره ها یا سایر ویژگی های آن مانند دهان، چشم و مواردی از این قبیل استفاده می شود.

پیش بینی آب و هوا

از الگوریتم بیز ساده در پیش بینی خوب یا بد بودن آب و هوا نیز استفاده می شود.

تشخیص پزشکی

پزشکان می توانند با استفاده از اطلاعات طبقه بندی کننده، بیماری را تشخیص دهند. متخصصان به منظور نشان دادن این موضوع که آیا بیمار در معرض خطر بیماری ها و شرایط خاص مانند بیماری های قلبی، سرطان و سایر بیماری ها است یا خیر، از این الگوریتم استفاده می کنند.

پیش بینی کننده بی درنگ

طبقه بندی ساده بیز، یک طبقه بندی کننده مشتاق یادگیری است که بسیار سریع عمل می کند؛ بنابراین، می توان از آن به منظور پیش بینی در زمان های واقعی استفاده کرد.

پیش بینی چند کلاسه

این الگوریتم برای ویژگی پیش بینی چند کلاسه نیز به خوبی شناخته شده است. در اینجا می توانیم احتمال چند کلاس متغیر هدف را پیش بینی کنیم.

سیستم توصیه

طبقه بندی کننده ساده بیز و فیلتر مشارکتی، با هم یک سیستم توصیه می سازند که از تکنیک های یادگیری ماشین و داده کاوی به منظور فیلتر کردن اطلاعات نادیده گرفته شده، استفاده می کند و پیش بینی می کند که آیا کاربر نیاز به منبع خاصی دارد یا خیر.

مزایا و معایب الگوریتم بیز ساده

الگوریتم بیز ساده دارای مزایا و معایبی است که در ادامه به بررسی هر یک از آن ها خواهیم پرداخت.

مزایای الگوریتم بیز ساده

از جمله مزایای الگوریتم بیز ساده، می توان به موارد زیر اشاره کرد:

- **پیچیدگی کمتر:** الگوریتم بیز ساده در مقایسه با سایر طبقه بندی کننده ها، به عنوان طبقه بندی کننده ساده تری در نظر گرفته می شود؛ زیرا تخمین پارامترها در آن ساده تر است. در نتیجه، یکی از اولین الگوریتم هایی است که در دوره های علم داده و یادگیری ماشین آموزش داده می شود.
- **مقیاس خوب:** الگوریتم بیز ساده، در مقایسه با رگرسیون لجستیک، یک طبقه بندی کننده سریع و کارآمد است. همچنین تا زمانی که فرض استقلال شرطی هم وجود داشته باشد، نسبتاً دقیق عمل می کند. این الگوریتم نیاز به ذخیره سازی کمی دارد.
- **مدیریت داده هایی با ابعاد بالا:** در موارد خاصی مانند طبقه بندی اسناد می تواند ابعاد بالایی داشته باشد که مدیریت آن برای طبقه بندی کننده های دیگر دشوار است.

از دیگر مزایای الگوریتم بیز ساده می توان به موارد زیر اشاره کرد:

- به داده های آموزشی زیادی نیاز ندارد.
- داده های پیوسته و گسسته را مدیریت می کند.
- با تعداد پیش بینی ها و نقاط داده زیاد، باز هم مقیاس پذیر است.
- سریع است و می توان از آن به منظور پیش بینی در زمان های واقعی استفاده کرد.
- به ویژگی های نامربوط حساس نیست.

معایب الگوریتم بیز ساده

از جمله معایب الگوریتم بیز ساده، عبارت اند از:

- **فرکانس صفر:** فرکانس صفر زمانی اتفاق می افتد که یک متغیر طبقه بندی در مجموعه آموزشی وجود نداشته باشد. به عنوان مثال، فرض کنید که در حال تلاش به منظور یافتن برآوردگر حداکثر احتمال برای کلمه "آقا" با توجه به کلاس "هرزنامه" هستید؛ اما کلمه "آقا" در داده های آموزشی وجود ندارد. احتمال در این حالت، صفر خواهد بود. از آنجایی که این طبقه بندی کننده همه احتمالات شرطی را در هم ضرب می کند، به این معنا خواهد بود که احتمال پسین نیز صفر است. به منظور جلوگیری از بروز این چنین مشکلاتی، می توان از هموارسازی لاپلاس (Laplace) استفاده کرد.
- **فرض اصلی غیرواقعی:** در حالی که فرض استقلال مشروط به صورت کلی دارای عملکرد خوبی است، این فرض همیشه برقرار نیست و در مواردی سبب طبقه بندی های نادرستی می شود.

بهبود کارکرد الگوریتم بیز ساده

در اینجا چند نکته برای بهبود کارکرد الگوریتم بیز ساده آورده شده است:

- اگر ویژگی های پیوسته، توزیع نرمالی ندارند، باید از تبدیل یا روش های مختلف به منظور تبدیل آن به توزیع نرمال استفاده کرد.
- در صورتی که مجموعه داده های آزمون دارای مشکل فرکانس صفر باشند، می توان از تکنیک های هموارسازی تصحیح لاپلاس برای پیش بینی کلاس مجموعه داده های آزمایشی استفاده کرد.
- ویژگی های همبسته را حذف کنید؛ زیرا ویژگی های بسیار همبسته در مدل، به گونه ای قرار گرفته اند که دو بار به آن ها رأی داده می شود و می تواند منجر به اهمیت بیش از حد آن ها شود.
- ممکن است در فکر اعمال برخی تکنیک های طبقه بندی کننده مانند هم آمیزی، بسته بندی و تقویت باشید؛ اما این روش ها کمک نخواهند کرد. در واقع هم آمیزی، تقویت و بسته بندی کمکی نمی کنند؛ زیرا هدف آن ها کاهش واریانس است و الگوریتم بیز ساده هیچ واریسی برای به حداقل رساندن ندارد.
- طبقه بندی ساده بیز، گزینه های محدودی برای تنظیم پارامترهایی مانند alpha = 1 برای صاف کردن، fit – prior= [True | False] برای یادگیری احتمالات پسین کلاس و برخی از گزینه های دیگر دارد. بهتر است روی پیش پردازش داده ها و انتخاب ویژگی ها متمرکز شوید.