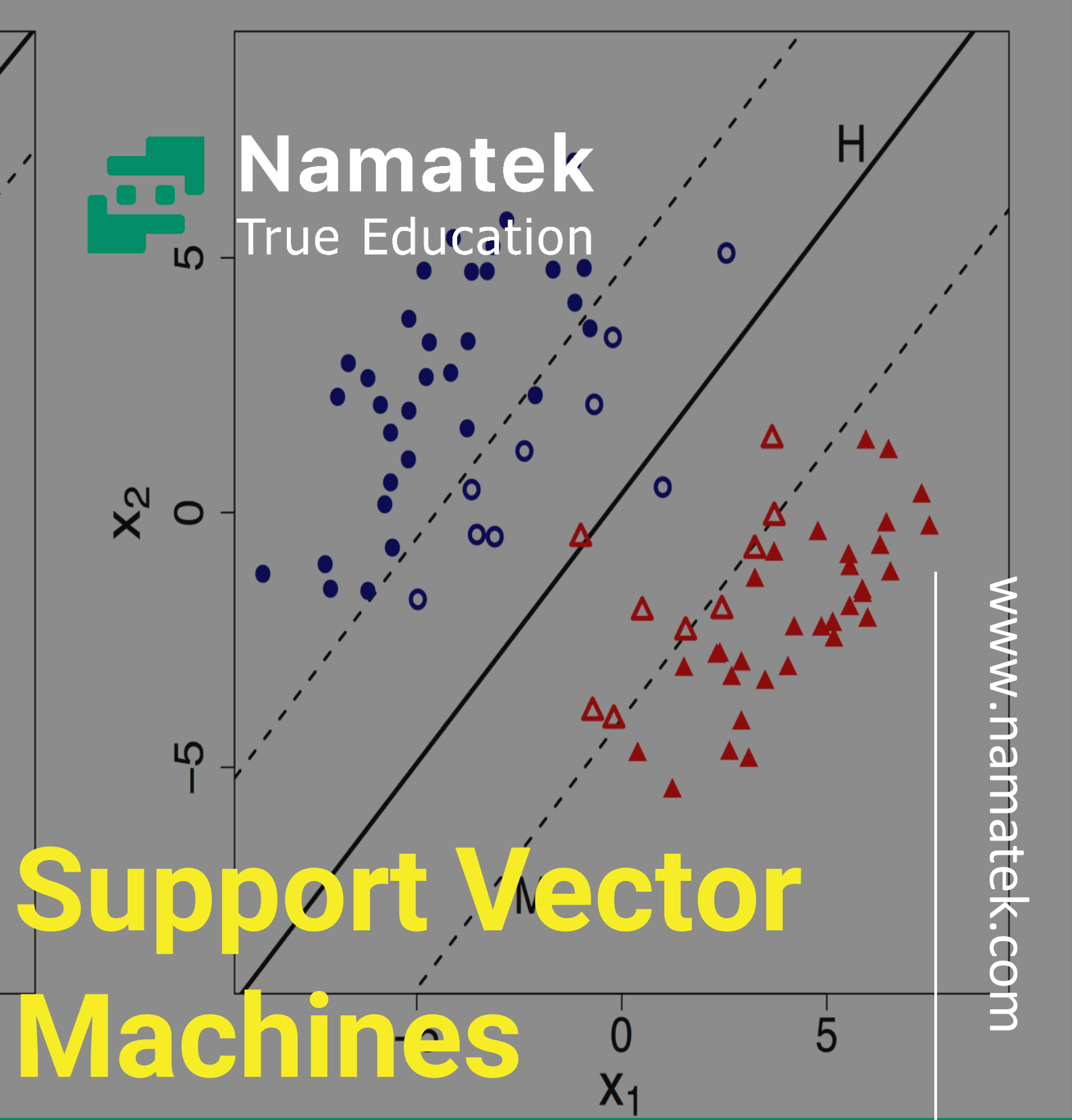


Support Vector Machines

ماشین بردار پشتیبان

www.namatek.com

فهرست مطالب

ماشین بردار پشتیبان، یکی از محبوب ترین الگوریتم های یادگیری نظارتی است که برای مسائل طبقه بندی استفاده می شود. برای مثال فرض کنید گربه عجیبی می بینیم که برخی از ویژگی های سگ ها را نیز دارد.
بنابراین به دنبال مدلی خواهیم بود که گربه یا سگ بودن این موجود عجیب را تشخیص دهد. می توان با استفاده از الگوریتم ماشین بردار پشتیبان، چنین مدلی ایجاد کرد. در ابتدا، مدل را با تصاویر زیادی از گربه و سگ آشنا می کنیم تا بتواند با ویژگی های گربه ها و سگ ها آشنا شود و سپس آن را با این موجود عجیب تست می کنیم. پس الگوریتم، حالات این جانور را دسته بندی می کند و مواردی که بیشتر مشاهده می کند را انتخاب خواهد کرد؛ فرض می کنیم، مدل این گونه در نظر گرفته که موجود مذبور بیشتر ویژگی های یک گربه را دارد. در این مقاله به بررسی ماشین بردار پشتیبان، انواع آن، ویژگی ها، نحوه عملکرد، کاربردها و مزایای و معایب این الگوریتم می پردازیم.

ماشین بردار پشتیبان چیست؟

ماشین بردار پشتیبان یا SVMs (Support Vector Machine) یک الگوریتم یادگیری ماشین تحت نظارت است که می تواند برای چالش های

طبقه بندی و رگرسیون استفاده شود. با این حال بیشتر در مسائل طبقه بندی مانند دسته بندی متن استفاده می شود. پس از ارائه یک مدل ماشین بردار پشتیبان، مجموعه ای از داده های آموزشی برچسب گذاری شده برای هر دسته به وجود خواهد آمد و آنها می توانند متن جدید را دسته بندی کنند.

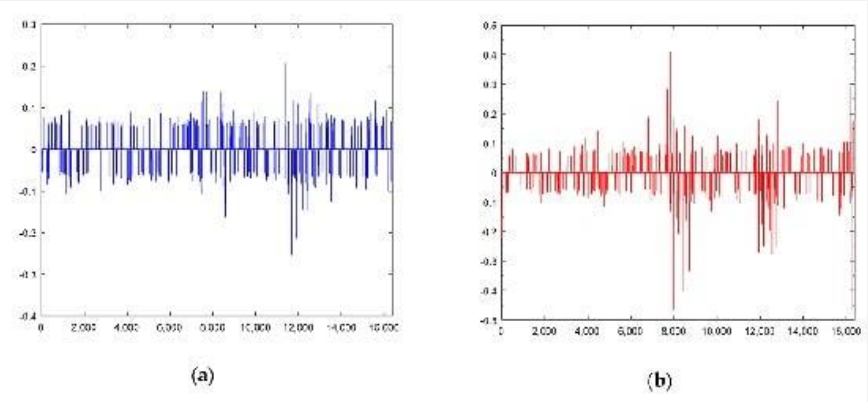
در الگوریتم ماشین بردار پشتیبان، هر آیتم داده را به عنوان یک نقطه در فضای n بعدی رسم می کنیم (که در آنn ، تعداد ویژگی های موجود است) و مقدار هر ویژگی، مقدار یک مختصات خاص است. سپس با پیدا کردن میزان بهینه هایپر پلان (Hyperplane) که این دو کلاس را به خوبی از یکدیگر متمایز می کند، طبقه بندی انجام خواهد شد.

در مقایسه با الگوریتم های جدیدتری مانند شبکه های عصبی، ماشین بردار پشتیبان دارای دو مزیت است، یکی سرعت بالاتر و دیگری عملکرد بهتر با تعداد محدودی از نمونه. این موضوع سبب می شود که الگوریتم SVMs، برای مسائل طبقه بندی متن بسیار مناسب باشد، جایی که دسترسی به مجموعه داده های چند هزار نمونه ای برچسب گذاری شده معمول و متداول است.

ویژگی های ماشین بردار پشتیبان

آنچه ماشین بردار پشتیبان را به یک چارچوب یادگیری ماشینی جذاب تبدیل کرده، ویژگی های آن است که در ادامه با آن آشنا می شویم.

تکنیک پراکنده

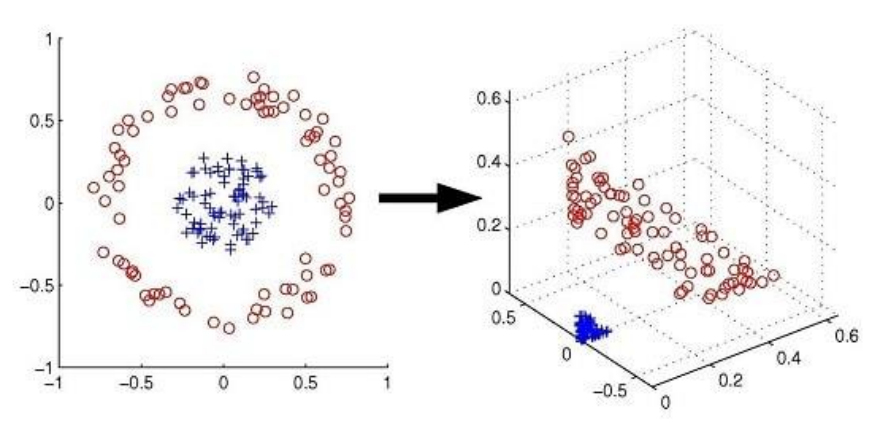


همانند روش های غیرپارامتریک، SVMs نیاز دارد که تمامی داده های آموزشی را در دسترس داشته باشد؛ یعنی در طول مرحله آموزش، زمانی که پارامترهای مدل آموخته می شوند، در حافظه ذخیره خواهند شد. با این حال، زمانی که پارامترهای مدل شناسایی شوند، ماشین بردار پشتیبان برای پیش بینی آینده تنها به زیر مجموعه ای از این نمونه های آموزشی با نام بردارهای پشتیبانی، بستگی خواهد داشت. بردارهای پشتیبان حاشیه هایپر پلان ها (Hyperplanes) را تعریف می کند و بردارهای پشتیبانی پس از یک مرحله بهینه سازی (که شامل یک تابع هدف است و با یک عبارت خطا و یک محدودیت منظم می شود) با استفاده از ریلکسیشن لانگراژی (Lagrangian Relaxation) یافت خواهد شد.

پیچیدگی کار طبقه بندی با ماشین بردار پشتیبان به تعداد بردارهای پشتیبانی بیشتر بستگی دارد نسبت به ابعاد فضای ورودی. تعداد بردارهای پشتیبان که در نهایت از مجموعه داده های اصلی حفظ می شوند، وابسته به داده است و بر اساس پیچیدگی های داده ها که با ابعاد داده و قابلیت تفکیک پذیری کلاس در نظر گرفته می شود، متفاوت خواهد بود. حد بالایی

برای تعداد بردارهای پشتیبانی، نصف اندازه مجموعه داده های آموزشی است؛ اما در عمل به ندرت چنین چیزی رخ می دهد.

تکنیک هسته ای



ماشین بردار پشتیبان از ترفند هسته به منظور نگاشت داده ها در فضایی با ابعاد بالاتر، قبل از حل کار یادگیری ماشین به عنوان یک مسئله بهینه سازی محدب استفاده می کند که در آن بهینه ها به صورت تحلیلی و نه اکتشافی، همانند سایر تکنیک های یادگیری ماشین یافت می شوند. اغلب، داده های واقعی به صورت خطی در فضای ورودی اصلی، قابل تفکیک نیستند. به عبارت دیگر، نمونه هایی که دارای برچسب های متفاوتی هستند، فضای ورودی را به گونه ای به اشتراک می گذارند که مانع از جداسازی صحیح کلاس های مختلف درگیر در این طبقه بندی توسط یک هایپر پلان خطی شوند. تلاش برای یادگیری یک مرز جدا کننده غیرخطی در فضای ورودی، نیازهای محاسباتی را در مرحله بهینه سازی، افزایش می دهد؛ زیرا سطح جدا کننده حداقل از مرتبه دوم خواهد بود.

در عوض SVMs، داده ها را با استفاده از توابع هسته ای که از پیش تعریف شده اند، در فضایی جدید، اما با ابعاد بالاتر ترسیم می کند، جایی که یک

جدا کننده خطی می تواند بین کلاس های مختلف تمایز قائل شود. بنابراین، مرحله بهینه سازی ماشین بردار پشتیبان، مستلزم یادگیری تنها یک سطح متمایز خطی در فضای نقشه برداری است. البته توجه به این نکته لازم است که انتخاب و تنظیمات تابع هسته برای بهینه سازی SVMs نیز بسیار مهم خواهد بود.

جدا کننده حداکثر حاشیه

فراتر از به حداقل رساندن خطا یا تابع هزینه، بر اساس مجموعه داده های آموزشی (مانند سایر تکنیک های یادگیری ماشین متمایز) ماشین بردار پشتیبان یک محدودیت اضافی را بر مسئله بهینه سازی تحمیل می کند: هایپر پلان باید به گونه ای باشد که در حداکثر فاصله از کلاس های مختلف قرار گیرد.

چنین عبارتی، مرحله بهینه سازی را وادار به پیدا کردن هایپر پلانی می کند که در نهایت بهتر می تواند تعمیم یابد؛ زیرا در فاصله مساوی و حداکثر از کلاس ها قرار دارد. این امر ضروری است؛ زیرا آموزش بر روی نمونه ای از جامعه انجام شده، در حالی که پیش بینی باید بر روی نمونه هایی صورت گیرد که هنوز دیده نشده اند و ممکن است توزیعی داشته باشد که کمی با توزیع زیر مجموعه آموزش دیده متفاوت باشد.

انواع ماشین بردار پشتیبان

ماشین بردار پشتیبان می تواند دو نوع باشد:

- خطی
- غیرخطی

که در ادامه با آن ها آشنا خواهیم شد.

ماشین بردار پشتیبان خطی

ماشین بردار پشتیبان خطی برای داده های قابل جداسازی خطی استفاده می شود. به این معنا که اگر یک مجموعه داده را بتوان با استفاده از یک خط مستقیم به دو کلاس طبقه بندی کرد، آنگاه این داده ها را داده های قابل جداسازی خطی می نامند و SVMs طبقه بندی کننده را با عنوان طبقه بندی کننده خطی می شناسند.

ماشین بردار پشتیبان غیرخطی

ماشین بردار پشتیبان غیرخطی برای داده های جدا شده غیرخطی استفاده می شود. به این معنا که اگر یک مجموعه داده را نتوان با استفاده از یک خط مستقیم طبقه بندی کرد. این داده ها را داده های ماشین بردار پشتیبان غیرخطی گفته و طبقه بندی استفاده شده را طبقه بندی غیرخطی می نامند.

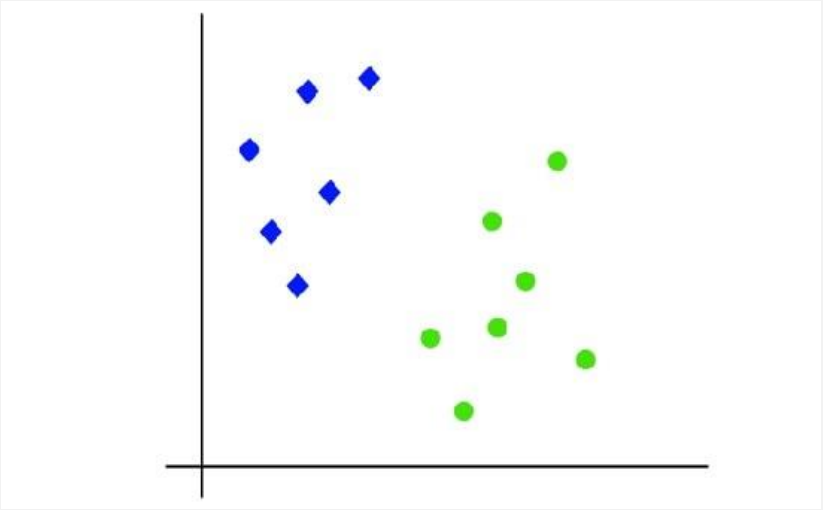
چگونگی کارکرد ماشین بردار پشتیبان

در ادامه چگونگی عملکرد SVMs را به صورت خطی و غیرخطی بررسی می کنیم.

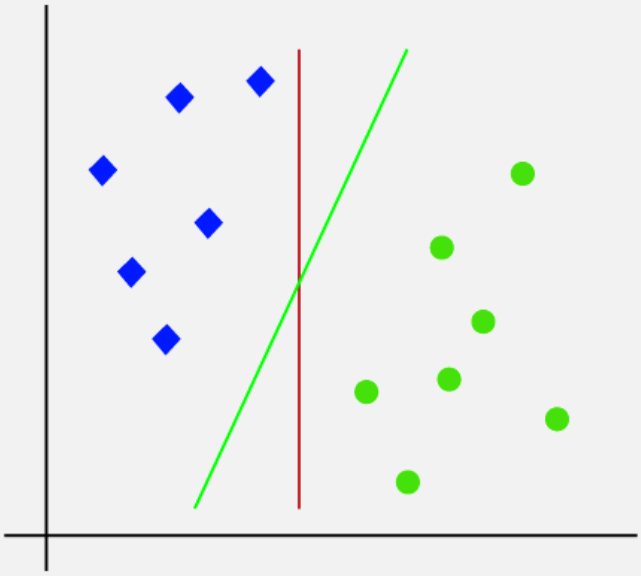
نحوه عملکرد ماشین بردار پشتیبان خطی

عملکرد الگوریتم SVMs خطی را می توان به آسانی و با استفاده از یک مثال بیان کرد. فرض کنید مجموعه ای از داده با دو برچسب آبی و سبز داریم و مجموعه داده دارای دو مختصات x_1 و x_2 است. ما طبقه بندی ای

می خواهیم که بتواند هر دوی مختصات x_1 و x_2 را به رنگ های سبز یا آبی طبقه بندی کند. همانگونه که در تصویر زیر آمده است:



بنابراین، از آنجایی که فضا دو بعدی است، تنها با استفاده از یک خط مستقیم می توان به راحتی این دو کلاس را از یکدیگر جدا کرد. اما ممکن است چندین خط به صورت همزمان وجود داشته باشد، که قادر باشند، این کلاس ها را از یکدیگر جدا کند؛ همانند تصویر زیر:



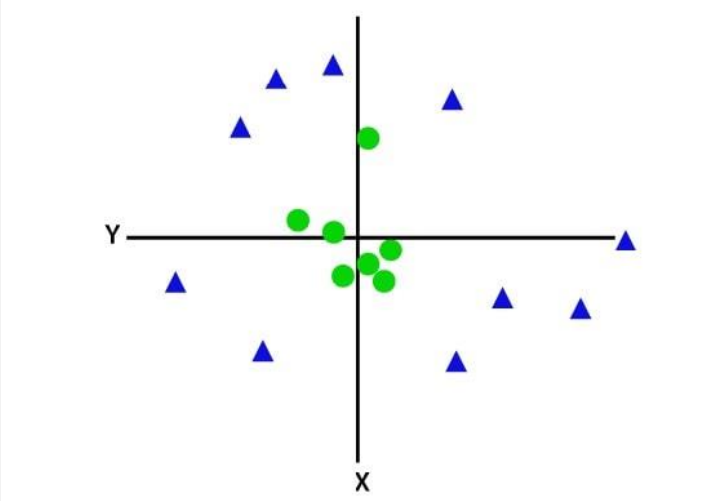
در این حالت، الگوریتم ماشین بردار پشتیبان به یافتن بهترین خط یا مرز تصمیم کمک خواهد کرد که بهترین مرز یا منطقه و با نام هایپر پلان خواهد بود. الگوریتم ماشین بردار پشتیبان، نزدیکترین نقطه خطوط از هر دو کلاس

را پیدا می کند. همان گونه که در بخش های قبل گفته شد، به این نقاط بردارهای پشتیبان می گویند. فاصله بین بردارها و هایپر پلان را حاشیه می نامند و هدف الگوریتم ماشین بردار پشتیبان، به حداکثر رساندن این حاشیه است. هایپر پلان با حداکثر حاشیه را هایپر پلان بهینه می گویند.

نحوه عملکرد ماشین بردار پشتیبان غیرخطی

اگر داده ها به صورت خطی چیده شوند، می توان با استفاده از یک خط مستقیم آن ها را جدا کرد؛ اما برای داده های غیرخطی نمی توان از یک خط مستقیم استفاده کرد.

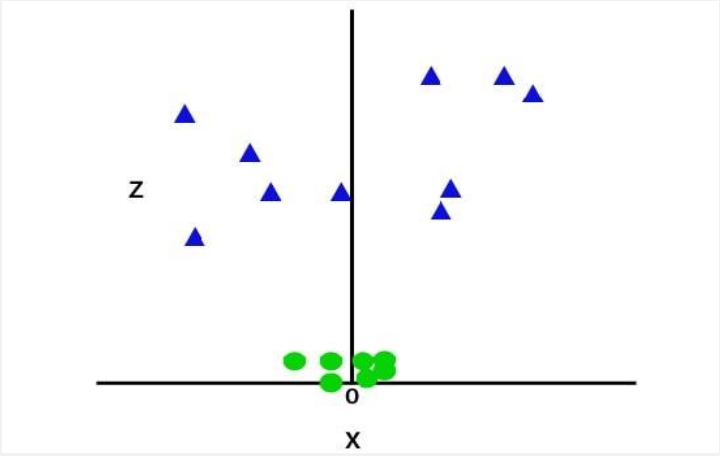
تصویر زیر را در نظر بگیرید:



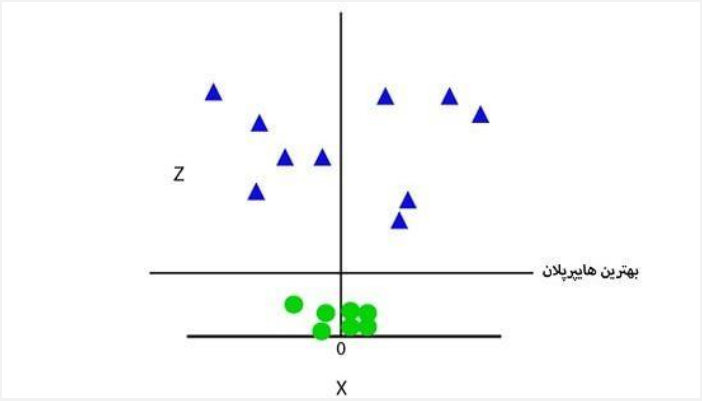
برای داده های خطی از دو بعد x و y استفاده کرده ایم. بنابراین به منظور جداسازی این نقاط داده، باید یک بعد دیگر نیز اضافه کرد و بعد سوم یعنی z به نمودار اضافه خواهد شد که می توان آن را به صورت زیر محاسبه کرد:

$$Z = x^{۲} + Y^{۲}$$

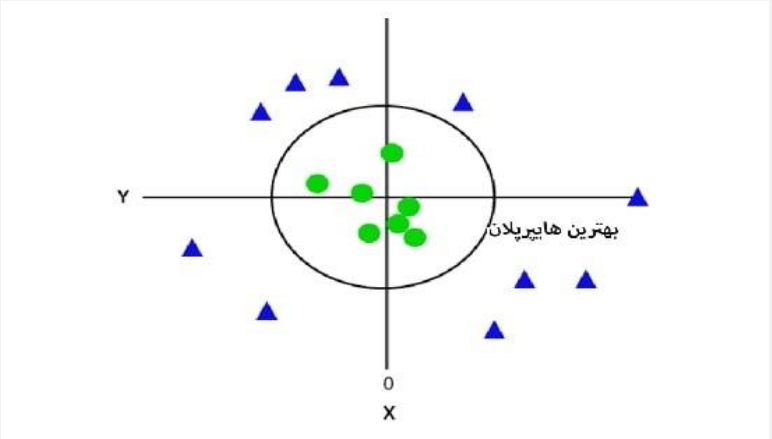
با اضافه شدن بعد سوم، فضای نمونه مانند تصویر زیر خواهد شد:



بنابراین و در این مرحله، ماشین بردار پشتیبان، مجموعه داده ها را به صورت زیر به کلاس هایی تقسیم می کند؛ همانند تصویر زیر:



از آنجایی که ما در فضای ۳ بعدی هستیم، این تصویر شبیه صفحه ای موازی با محور x است. اگر آن را در فضای دو بعدی با z = 1 تبدیل کنیم، به صورت زیر خواهد شد:



بنابراین، در مورد داده های غیرخطی، محیطی با شعاع یک به دست می آوریم.

کاربردهای ماشین بردار پشتیبان

SVMs دارای کاربردهای متفاوتی در صنعت هستند که از جمله آن ها می توان به موارد زیر اشاره کرد.

طبقه بندی متن

ماشین بردار پشتیبان، معمولاً در پردازش زبان طبیعی برای کارهایی مانند موارد زیر استفاده می شود:

- تجزیه و تحلیل احساسات
- تشخیص هرزنامه
- مدل سازی موضوع

این الگوریتم با داده هایی با ابعاد بالا، عملکرد بهتری ارائه می دهد.

طبقه بندی تصویر

ماشین بردار پشتیبان در کارهای طبقه بندی تصویر مانند تشخیص اشیا و بازیابی تصویر استفاده می شود. همچنین این الگوریتم در حوزه های

امنیتی بسیار مفید عمل می کند و می تواند یک تصویر را در دسته عکس های دستکاری شده قرار دهد.

بیوانفورماتیک

ماشین بردار پشتیبان، همچنین به منظور طبقه بندی موارد زیر استفاده می شوند:

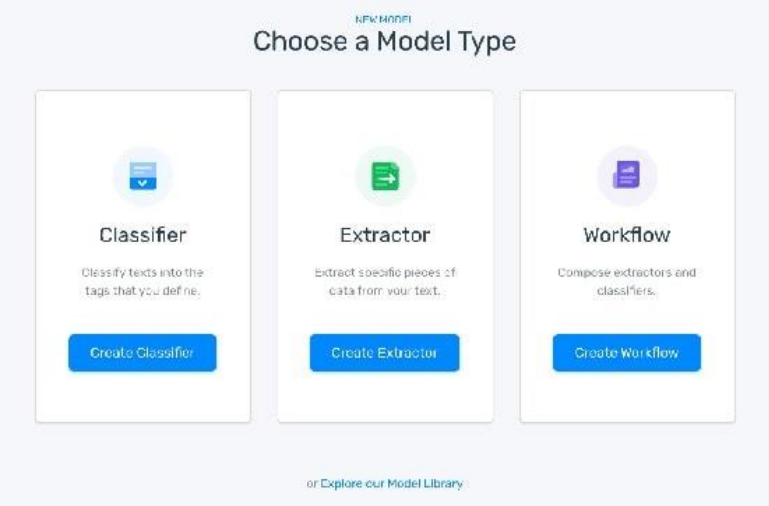
- پروتئین
- تجزیه و تحلیل و بیان ژن
- تشخیص بیماری

از این الگوریتم اغلب در تحقیقاتی که در مورد سرطان ها است، استفاده می شود؛ زیرا می توانند روندهای ظریف در مجموعه داده های پیچیده را تشخیص دهند.

سیستم اطلاعات جغرافیایی

SVMs می تواند، ساختارهای ژئوفیزیکی لایه های زیرِ زمین را نیز تجزیه و تحلیل و نویز داده های الکترومغناطیسی را فیلتر کنند. این الگوریتم، همچنین به پیش بینی پتانسیل روان گرایی لرزه ای خاک که مربوط به رشته مهندسی عمران است، کمک می کند.

نحوه طبقه بندی ماشین بردار پشتیبان



برای ایجاد طبقه بندی در SVM، بدون استفاده از بردارها و هسته ها، می توان یکی از مدل های طبقه بندی از پیش ساخته شده MonkeyLearn را برای شروع استفاده کرد. طبقه بندی ماشین بردار پشتیبان در ۸ مرحله صورت می پذیرد؛ ولی قبل از آن باید در MokeyLearn به صورت رایگان ثبت نام کنید.

این مراحل به صورت زیر هستند:

۱. یک طبقه بندی جدید ایجاد کنید. به داشبورد بروید، روی Create a Model کلیک کنید و Classifier را انتخاب کنید.

۲. نحوه طبقه بندی داده های خود را انتخاب کنید. برای مثال مدل Topic Classification را انتخاب کنید.

۳. داده های آموزشی خود را وارد کنید. داده هایی که برای آموزش نیاز دارید را انتخاب و آپلود کنید. به خاطر داشته باشید که طبقه بندی کننده ها یاد می گیرند و با ارائه داده های آموزشی، هوشمندتر می شوند. می توانید داده ها را از فایل CSV یا اکسل وارد کنید.

٤. برچسب ها را برای طبقه بندی کننده SVM خود تعریف کنید. برای شروع حداقل دو تگ اضافه کنید؛ در ادامه می توانید برچسب های بیشتری اضافه کنید.
٥. داده ها را برای آموزش طبقه بندی کننده خود، برچسب گذاری کنید. پس از برچسب گذاری دستی برخی از نمونه ها، طبقه بندی کننده به تنهایی شروع به پیش بینی می کند. اگر می خواهید مدل دقیق تری داشته باشید باید نمونه های بیشتری را برای ادامه آموزش مدل خود تگ کنید.
٦. الگوریتم خود را روی SVM تنظیم کنید. به بخش تنظیمات رفته و مطمئن شوید که الگوریتم SVM را در بخش پیشرفته انتخاب کرده باشید.
٧. طبقه بندی کننده خود را آزمایش کنید. روی RUN و سپس Demo کلیک کنید و متن خود را بنویسید و ببنید مدل چگونه داده های جدید را طبقه بندی می کند.
٨. طبقه بندی موضوع را یکپارچه کنید.

مزایا و معایب ماشین بردار پشتیبان

مزایا و معایب ماشین بردار پشتیبان را در ادامه بررسی خواهیم کرد.

مزایا

مزایای ماشین بردار پشتیبان عبارت اند از:

- در فضایی که ابعاد بالایی دارد، مؤثر است.
- در مواردی که تعداد ابعاد از تعداد نمونه بیشتر باشد، به صورت مؤثری عمل خواهد کرد.
- از زیر مجموعه ای از مجموعه آموزشی در تابع تصمیم (به نام بردارهای پشتیبانی) استفاده می کند، بنابراین دارای حافظه کارآمدی است.
- با حاشیه های جدا کننده ای که واضح هستند، به خوبی کار می کند.
- چند منظوره است، به این معنا که:
  - توابع کرنلی مختلفی را می توان برای تابع تصمیم مشخص کرد.
  - هسته های مشترکی در آن ارائه شده است.
  - امکان تعیین هسته های سفارشی نیز وجود دارد.

معایب

از جمله معایب ماشین بردار پشتیبان می توان به موارد زیر اشاره کرد:

- زمانی که مجموعه داده های بزرگ داریم، عملکرد خوبی ندارند، زیرا زمان لازم به منظور آموزش مورد نیاز بیشتر خواهد شد.
- هنگامی که مجموعه داده ها، نویز بیشتری دارد، یعنی کلاس های هدف با یکدیگر هم پوشانی داشته باشند، عملکرد چندان خوبی ندارند.
- ماشین بردار پشتیبان، به صورت مستقیم برآوردهای احتمال را ارائه نمی کند و با استفاده از اعتبارسنجی متقابل که ۵ برابر قیمت بیشتری دارند، محاسبه می شود که در روش SVC مربوط به کتابخانه اسکیت لرن (Scikit Learn) پایتون است.