1-مطالعه مبانی نظری و استخراج متغیرهای کمی موثر بر مدیریت سود

2-جمع آوری اطلاعات مربوط به شرکت های دارای مدیریت سود به صورت صفر و یک

3-متن کاوی گزارشات مالی

4-آموزش مدل و ارزیابی و مقایسه نتایج

5- انتخاب بهترین الگوی ارائه شده جهت پیش بینی

مرحله اول

برای انجام مرحله اول تحقیق ابتدا مطالعه و جستجوی اکتشافی در متون مربوطه انجام خواهد پذیرفت تا متغیرهای کمی مدیریت سود استخراج شود.

مرحله دوم

داده های پژوهش از طریق گزارش ها و اطلاعیه های منتشر شده در شبکه کدال، بانک های اطلاعاتی سازمان بورس و اوراق بهادار و بانک اطلاعاتی شرکتهای پردازش اطلاعات مالی جمع آوری خواهد گردید.

مرحله سوم

در این مرحله گزارشات مالی مورد متن کاوی قرار خواهد گرفت ولی نخست لازم است که الگوی مورد استفاده جهت متن کاوی مشخص گردد:

الف)کلاس‌بندی یا دسته‌بندی

کلاس بندی یک کار علمی برای پیش بینی مقدار متغیر طبقه بندی شده (هدف یا کلاس) با ساختن یک مدل بر اساس یک یا چند متغیر عددی و / یا دسته ای (پیش بینی کننده یا ویژگی) است. این روش یکی از تکنیک های پرکاربرد ماشین یادگیری است که هدف آن طبقه بندی مجموعه مشخصه های مشابه در یک کلاس است.. داشتن داده های از پیش برچسب گذاری شده شده و طبقه بندی شده از الزامات این روش است. به این صورت که در فاز اول، یادگیری بر اساس مجموعه ای از داده های برچسب گذاری شده انجام میپذیرد سپس در فاز دوم به منظور ارزیابی دقت مدل، خروجی حاصل از داده های آزمون با واقعیت بررسی میگردد.

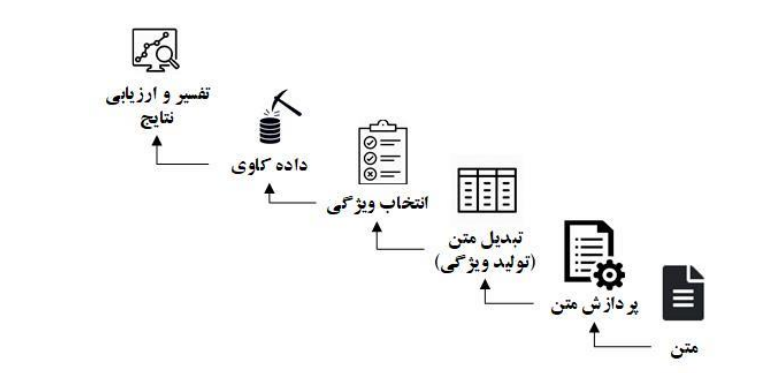
ب) خوشه‌بندی[[1]](#footnote-1)

خوشه‌بندی به فرایندی گفته میشود که درآن مجموعه ای از داده ها به گونه ای به گروه های مجزا افراز می شوند که اعضای هر خوشه در عین حال که بیشترین شباهت را با یکدیگر دارند دارای بیشترین تفاوت با اعضای خوشه های دیگر هستند . مهمترین تفاوت این روش با کلاس بندی در این است که در این روش هیچ گروه از پیش تعیین شده ای وجود ندارد.

ج) کاوش قوانین وابستگی

این روش از طریق کشف و یافتن قوانین در رویدادها و رخداد های دیگر وقوع موردی را درآینده پیش بینی می­کند. کاوش قوانین وابستگی ، بدلیل عدم استفاده از هیچ برچسب قبلی برای آموزش الگوریتم ها در زمره الگوریتم های یادگیری بدون نظارت جای می گیرد. (کومار و راوی[[2]](#footnote-2)،2016).

پس از انتخاب الگوی متن کاوی،گزارشات هیئت مدیره مورد متن کاوی قرار خواهد گرفت. متن کاوی در برگیرنده مراحل مختلفی است که میتوان پیش پردازش، استخراج ویژگی و انتخاب ویژگی را به عنوان مراحل اصلی آن عنوان کرد:



مراحل مختلف پیش پردازش متن

**پیش پرازش متن**

پیش پردازش متن در این تحقیق بر روی متون گزارشات فعالیت هیئت مدیره اعمال خواهد شد. پیش پردازش شامل مراحلی است که در آن سندهای متن خام به عنوان ورودی داده شده و خروجی آن مجموعه ای از کلمات است که می تواند در مدل بردار استفاده شود. در این قسمت به بررسی روش های پیش پردازش متن شامل تکه تکه کردن، تکه تکه کردن، تکه تکه کردن، ریشه یابی، حذف کلمات توقف و هرس کردن می پردازیم. (ایمانی، ۱۳۹۱)

**نرسازی و اصلاح نویسه ها(Normalize) :**

یکی از مشکلات زبان فارسی وجود چند نمونه ی مختلف از یک نویسه است که کار جستجو در متون فارسی را مشکل می کند. در این مرحله کاراکترهای غیراستاندارد با کاراکترهای استاندارد جایگزین می شوند و کاراکترهای اضافی نیز بسته به نوع پردازش از بین می روند تا واژه های یکسان در تمامی متن به یک صورت نوشته شده باشند .

برای این منظور میتوان از ابزار و روش های زیراستفاده کرد:

یکسان ساز : زبان فارسی از زبان هایی است است که دارای پیچیدگی زیادی است و کارهای پیش پردازش آن چالش برانگیز می باشد( اسرابی و دیگران، ۲۰۱۳)، پردازش زبان فارسی از جهاتی با پردازش زبان انگلیسی تفاوت دارد. در زبان انگلیسی تمامی حروف و تمامی کلمات جدا از هم و با قانونی مشخص نوشته می شوند و این در حالی است که در زبان فارسی بعضی از حروف به هم چسبیده هستند، برخی از حروف جدا از هم نوشته می شوند، بعضی از کلمات یکپارچه اند، بعضی از کلمات با فاصله با نیم فاصله به دو یا چند بخش تقسیم می شوند. تمامی حوزه های مرتبط با پردازش زبان طبیعی به نحوی با متون واقعی سروکار دارند. یکی دیگر از مهمترین مشکلات پردازش متن فارسی، وجود کد گذاری های مختلف در کارکترهای متنی آن است( محتاج[[3]](#footnote-3) و دیگران، ۲۰۱۸). صورت های غیر استاندارد نویسه ها و کلمات به وفور در این نوع متون نوشته دیده می شوند. قبل از این که بتوان از این متون به منظور استفاده در سیستم های تبدیل متن به گفتار، ترجمه ماشینی، بازشناسی حروف فارسی، خلاصه ساز فارسی، جستجو در متون فارسی و غیره استفاده کرد و یا در پایگاه داده ذخیره نمود، باید ابتدا پیش پردازشی روی آن ها انجام گیرد تا صورت های غیر استاندارد به شکل استاندارد تبدیل گردند. اگر حروف، نشانه های نگارشی و کلمات فارسی به شکل یکسانی نوشته نشوند، متون مورد استفاده قابل تحلیل توسط سامانه های رایانه ای نخواهند بود. طی فرایند ترمال سازی، علايم نگارشی، حروف، فاصله های بین کلمات، اختصارات و غیره بدون ایجاد تغییرات معنایی در متن به شکل استاندارد تبدیل می گردند. بنابراین، بایستی از یک استاندارد مشترک برای پیش پردازش و پردازش متون استفاده کرد.

در زبان فارسی رایج ترین ابزارها برای نرمال سازی متون که به صورت عمومی و رایگان در درسترس می باشند کتابخانه های هضم (hazm) توسعه داده شده توسط شرکت دانش بنیان راهکار پردازش ژرف (سبحه) و کتابخانه ی پارسیور (parsivar) توسعه داده شده توسط مرکز تحقیقات مخابرات ایران می باشند.( محتاج و دیگران، ۲۰۱۸). از مهم ترین ویژگی های ابزار کتابخانه های هضم میتوان به نکات زیر اشاره کرد:

یکسان سازی حروف عربی از جمله «ی » و « ک » عربی به «ی» و «ک» فارسی و حذف حروف همزه، تشدید و تبدیل «آ» به« ا»

يکسان سازی فاصله های اضافی و نیم فاصله ها در متن

تبديل ارقام عربی و انگلیسی به فارسی . . .

یکسان سازی حروف کشیده شده به صورت نرمال به عنوان مثال «دوســـــــت» به « دوست»

حذف کارکتر های اضافی از متن .

حذف خطوط خالی اضافی از متن

چسباندن پسوندهای «تر»، «ترین» و ... به آخر واژه ها .

تبديل «ه» به «ــــــهی» .

اصلاح فاصله گذاری پسوندها

اصلاح اعراب و حذف فتحه، کسره و ضمه و همچنین تنوین ها

از مهم ترین ویژگی های ابزار کتابخانه های پارسیور میتوان به نکات زیر اشاره کرد .

یکسان سازی حروف عربی از جمله «ی » و « ک » عربی به «ی» و «ک» فارسی حذف حروف همزه، تشديد و تبدیل « آ» به « ا».

یکسان سازی فاصله های اضافی و نیم فاصله ها در متن

تبديل ارقام عربی و انگلیسی به فارسی .

یکسان سازی حروف کشیده شده به صورت نرمال به عنوان مثال «دوســـــــت» به « دوست».

حذف کارکتر های اضافی از متن

حذف خطوط خالی اضافی از متن

چسباندن پسوندهای «تر»، «ترین» و ... به آخر واژه ها

تبديل «ه» به «ــــــهی» .

اصلاح فاصله گذاری پسوندها

اصلاح اعراب و حذف فتحه، کسره و ضمه و همچنین تنوین ها

شناسایی کلمات مرکب و اطلاح شیوه نگارشی آنها به شیوه مختلف با استفاده از پیکره کلمات مرکب

نرمال سازی تاریخ

تبدیل حروف فینگلیش به حروف فارسی

**تکه تکه کردن یا تبدیل متن به جملات و سپس به کلمات(Tokenization)**

در این مرحله جملات تکه تکه شده و به صورت مجموعه ای از کلمات در می آید. روش های پیچیده تکه تکه کردن متن از پردازش زبان طبیعی برای این کار بهره می گیرد، به این صورت که از تجزیه ساختار گرامری متن برای به دست آوردن کلمات پرمعناتر از قبیل اسم ها استفاده می کنند. در فرایند تحلیل متون اغلب بعد از یکسان سازی کردن متن عمل جداسازی کلمات با ایزار واحدساز انجام میشود.قابلیت های ابزار واحدساز هضم عبارتند از :

جداسازی و جایگزینی ایمیل ها از متن

جداسازی و جایگزینی هشتگ ها از متن

جداسازی و جایگزینی لینک ها از متن

جداسازیایموجی ها

جداسازی اعداد

جداسازی شناسه های عددی

ادغام سازی فعل های مرکب

**حذف کلمات توقف(stop words):**

کلمه ایست به کلمه ای گفته می شود که به تنهایی معنای خاصی را نمی رساند و در واقع به عنوان یک عنصر از مدل فضای بردار اطلاعات مفیدی در بر ندارد. در واقع منظور از کلمات توقف، کلماتی است که از لحاظ معنایی دارای اهمیت کمی بوده ولی در جملات به کررات استفاده میشود و معمولا در جملات، کاربرد ربطی دارند مثل «اگر»، «و»، «با». این نوع کلمات بدلیل تعدد بسیار و بار معنایی کم ، در اغلب فعالیت های حوزه پردازش زبان طبیعی در فاز پیش پردازش حذف میشوند.

**ریشه یابی لغات(stemming word)**

ریشه یابی لغات به گونه ای است که کلماتی که دارای ریشه مشابهی هستند به عنوان یک لغت درنظر گرفته خواهد شد. با توجه به ذات زایا و اشتقاق پذیر زبان فارسی و همچنین نیاز برنامه های کاربردی مختلف مرتبط با پردازش زبان طبیعی و بازیابی اطلاعات، ریشه یابی و لمیابی از مسائل مهم پیش پردازشی در پردازش زبان طبیعی فارسی به شمار می رود. ریشه یابی در لغت به معنای حذف پسوندها، پیشوندها و میانوندهای کلمه و به دست آوردن ریشه کلمه است. اگرچه در هر زبان، واژه ها با توجه به نقش معنایی و نحوی خود در جمله به شکل های متفاوتی ظاهر میشوند، اما با توجه به این که آنها از یک ریشه مشتق شده اند، از نظر معنایی به هم نزدیک هستند. پس در واقع ریشه یابی واژه ها نه به معنای زبان شناسی آن بلکه به معنای دسته بندی کلمات در گروه های معنایی یکسان تعریف می شود. بنابراین ریشه یابی مسئله ای است که در بسیاری از زمینه های پردازش زبان طبیعی مورد نیاز می باشد و در بسیاری از کاربردها، نیاز داریم تا همه مشتقات یک واژه را به ریشه ی آن، که همان شکل ساده واژه میباشد، تبدیل نماییم و البته باید توجه داشت که نیاز کاربردهای مختلف برای تحوه ریشه یابی متفاوت است. در این زمینه اغلب با دو واژه ریشه یابی و لمیابی روبرو هستیم. مطابق تعریف (سرابی و دیگران، ۲۰۱۳) در ریشه یابی عموما وندها جدا می شوند ولی در لم یابی با تحلیل مورفولوژی و بررسی کلمات زمینه، بن کلمه بازگردانده می شود. انتخاب از میان این دو وابسته به شرایط مورد نیاز در کاربردهای مختلف مثل عمق ریشه یابی مورد نیاز در برابر پیچیدگی زمانی الگوریتم است.

از مهم ترین قابلیت های لم یابی ابزار هضم میتوان به نکات زیر اشاره نمود

حذف کلینیک ها، پیشوندها و پسوندها از فعلها و بدست اوردن ریشه فعل بر اساس پیکره .

حذف کلیتیک ها، و پسوندها از اسم ها

ریشه یابی فعل های مرکب

مهم ترین قابلیت های لم بابی ابزار پارسیور عبارتند از : حذف کلینیک ها، پیشوندها و پسوندها از فعلها و بدست اوردن ریشه فعل بر اساس پیکره بسیار کامل تر از هضم

•حذف کلیتیک ها و پسوندها از اسم ها

• تبديل كلمات مکسر به ریشه مفرد

• ریشه یابی از فعلها با بیش از ۲۰ نوع حالت صرفی مختلف و بسیار کامل تر از هضم

**حذف کلمات و کاراکترهای زائد**

در این مرحله تمامی کاراکترهای اضافی همچون(\*،{}،@،؟،+و...)حذف میگردند.

**استخراج ویژگی و مدل سازی:**

پس از انجام مراحل پیش پردازش داده ها، فاز استخراج ویژگی و مدل سازی جهت آموزش در سیستم طبقه بندی و الگوریتم پیشنهادی یادگیری ماشین صورت خواهد گرفت

**استخراج کلمات کلیدی:**

مهم ترین و پایه ای ترین بخش سیستم طبقه بندی کننده متون، تهیه کلمات کلیدی است. روش های گوناگونی برای استخراج کلمات کلیدی وجود دارد که معروف ترین آنها حاصل ضرب TF-IDF است که در آن TF متناسب با بسامد یک کلمه در مستند و IDF یک فاکتور وزنی که بیانگر معکوس میزان پراکندگی یک کلمه در سندهای مختلف است. در این روش کلماتی به عنوان کلمه کلیدی انتخاب می شوند که ضمن داشتن بسامد وقوع بالا در تعداد محدودی سند واقع شده باشند و به عبارتی در اسناد مربوط به طبقه خاصی به تعداد زیاد دیده شوند. به صورت اختصار نحوه عمل ماتریس TF-IDF توضیح داده شده است:

ماتریس TF-IDF(تکرار کلمه \* تکرار سند معکوس) براساس فرمول زیر امتیاز هر کلمه را برای مرحله انتخاب ویژگی مشخص میکند.

که درآن تعداد تکرار کلمه در متن، تعداد متونی که در برگیرنده کلمه بوده،N تعداد متون مورد بررسی و نیزز وزن واژه در متن j است.

در این روش تکرار کلمات در سند نشان دهنده اهمیت اصطلاح در اسناد می باشد و تکرار سند در این اصطلاح (df، درصد اسناد حاوی این اصطلاح) نشان دهنده میزان اهمیت درکل متن می باشد. مقدار df پایین نشان دهنده منحصر به فرد بودن این اصطلاح در اسناد می باشد چرا که در بسیاری از اسناد ظاهر نمیشود. از این رو در این پژوهش به منظور وزن دهی از idf به جای df استفاده خواهد شد. لذا وزن بالا در روش tf\*idf تکرار زیاد یک کلمه در یک سند و اسناد کم حاوی این متن را نشان خواهد داد. در این مرحله کلماتی که بیشترین امتیاز را دارند ویژگی ها را مشخص خواهند کرد.

**مرحله پنجم: مدلسازی با الگوریتم های ماشین یادگیری**

پس از این مرحله، از ترکیب ماتریس کلمات و داده های مالی منتخب و مدیریت سود یا عدم مدیریت سود شرکت،ها ماتریس ورودی مدل تشکیل خواهد شد. در قدم بعدی ، اقدام به اجرای الگوریتم ماشین بردار پشتیبان برای ارائه مدل پیش بینی خواهد شد.

در پژوهش حاضر برای تشخیص مدیریت سود از رابطه زیر استفاده خواهد شد:

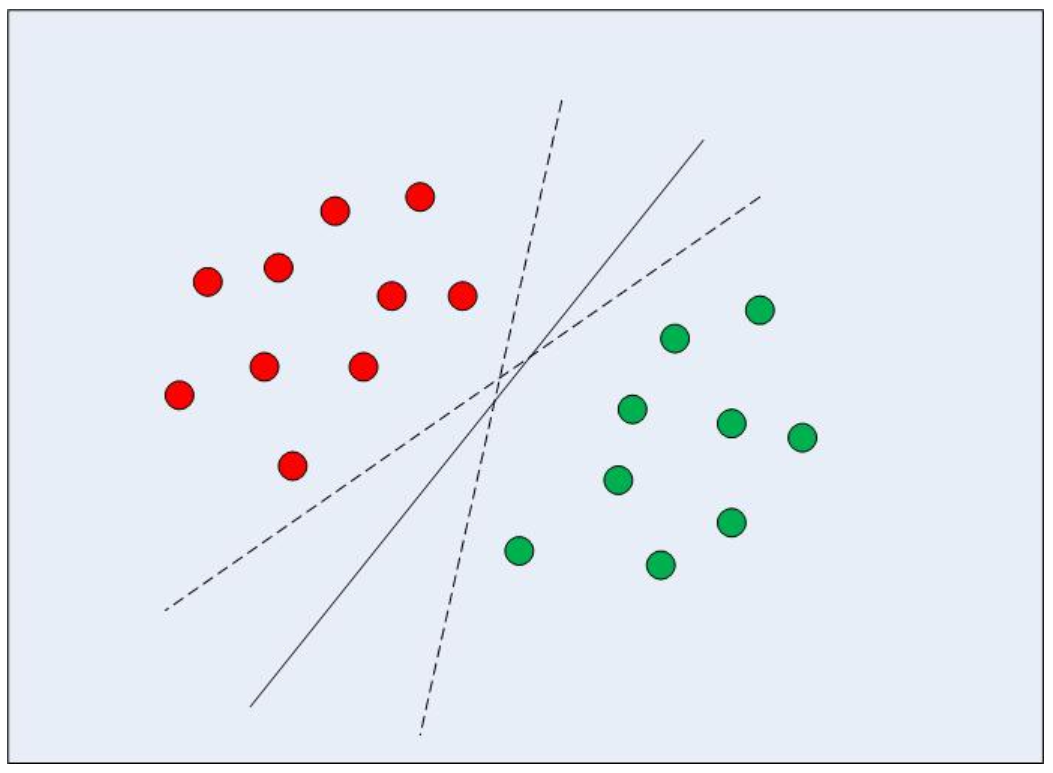
متغیر وابسته (Y)به صورت یک ماتریس m ×1است که اعضای آن، برای شرکتهایی که شاخص مدیریت سود در « تقسیم بندی اولیه» برای آنها بالا تشخیص داده شده است معادل یک و برای سایر شرکتها معادل صفر است. همچنان ماتریس متغیرهای توضیحی ( X) یک ماتریس m × n است که در آن mمعادل تعداد شرکت سالها و n معادل تعداد واژه ها و شاخص های مالی است که به عنوان متغیرهای توضیح دهنده شاخص مدیریت سود ، استفاده شده و باید از بین آنها واژه ها و شاخص هایی که بیشترین قدرت را در تببین مدیریت سود دارند، مشخص شوند. هر عضو این ماتریس که محل تلاقی یک سطر با یک ستون است، فراوانی نسبی(TF-IDF) یک واژه را نشان می دهد. پس از تشکیل ماتریس مذکور، وارد فاز آماری ومدلسازی با الگوریتم های ماشین یادگیری خواهیم شد.

**الگوریتم‌های مورد استفاده برای ساخت مدل**

در این بخش به بررسی الگوریتم ماشین بردار پشتیبان با کرنل‌های خطی و غیرخطی که در ساخت مدل استفاده شده است، می‌پردازیم.

**SVM با کرنل خطی**

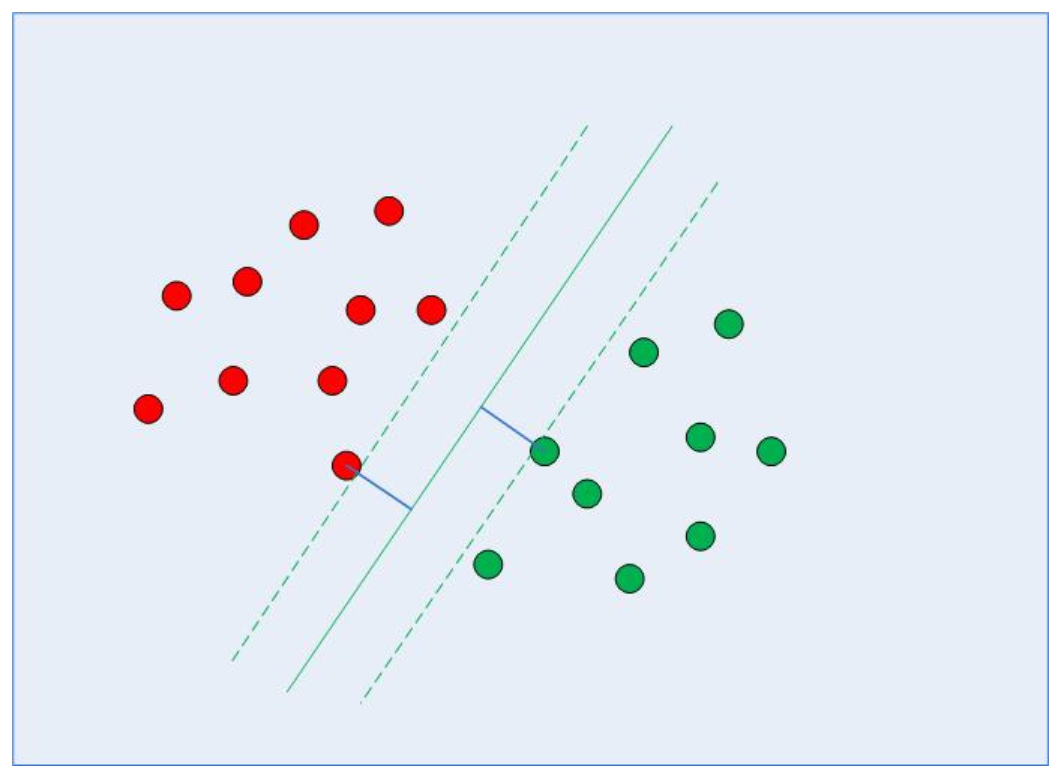
در رابطه با داده‌های خطی جدا از هم در دو بعد، همان‌طور که در شکل ‏3‑7 نشان داده‌شده است، در یک الگوریتم یادگیری ماشین با کرنل خطی سعی می‌شود داده‌ها مرزها طوری تعیین شوند که خطا به حداقل برسد. اگر به شکل ‏3‑7 دقت کنید، چندین مرز وجود دارد که نقاط داده را به‌درستی تقسیم می‌کنند. دو خط نقطه‌چین، داده‌ها را همانند خط ساده به‌درستی دسته‌بندی می‌کنند.



شکل ‏3‑7 -مرزهای تصمیم‌گیری چندگانه

SVM با الگوریتم‌های طبقه‌بندی دیگر که به نحوی مرز تصمیم را فقط تعیین می‌کنند متفاوت است. SVM تنها یک مرز تصمیم‌گیری نیست، بلکه بهینه‌ترین مرز یا ابر صفحه را مشخص می‌کند.

به کمترین فاصله بین خط جداکننده و نقاط مرزی حاشیه گفته می‌شود. خط یا ابر صفحه‌ای که از بیشترین حاشیه را دارد بهینه‌ترین مرز تصمیم‌گیری است. ، همان‌طور که در شکل ‏3‑8 مشاهده می‌شود، بردارهای پشتیبان همان نزدیک‌ترین نقاط در مرز تصمیم‌گیری‌اند که فاصله بین مرز تصمیم و نقاط را به حداقل می‌رساند. با فرض اینکه نمونه‌های یادگیری ما به‌صورت i=1,2,…I {xi,yi} باشند، که xi بردار ویژگی‌ها و yi در مجموعه {+1,-1} و کلاس‌های ما باشند، هدف جدا کردن این دو کلاس است. معادله این ابر صفحه به‌صورت x.w+b=0 است که x بردار ورودی ویژگی‌ها، w بردار وزن و b بایاس این ابر صفحه است. x.w همان نزدیک‌ترین نقطه داده از حاشیه جداسازی است. (برگس[[4]](#footnote-4) ،1998).

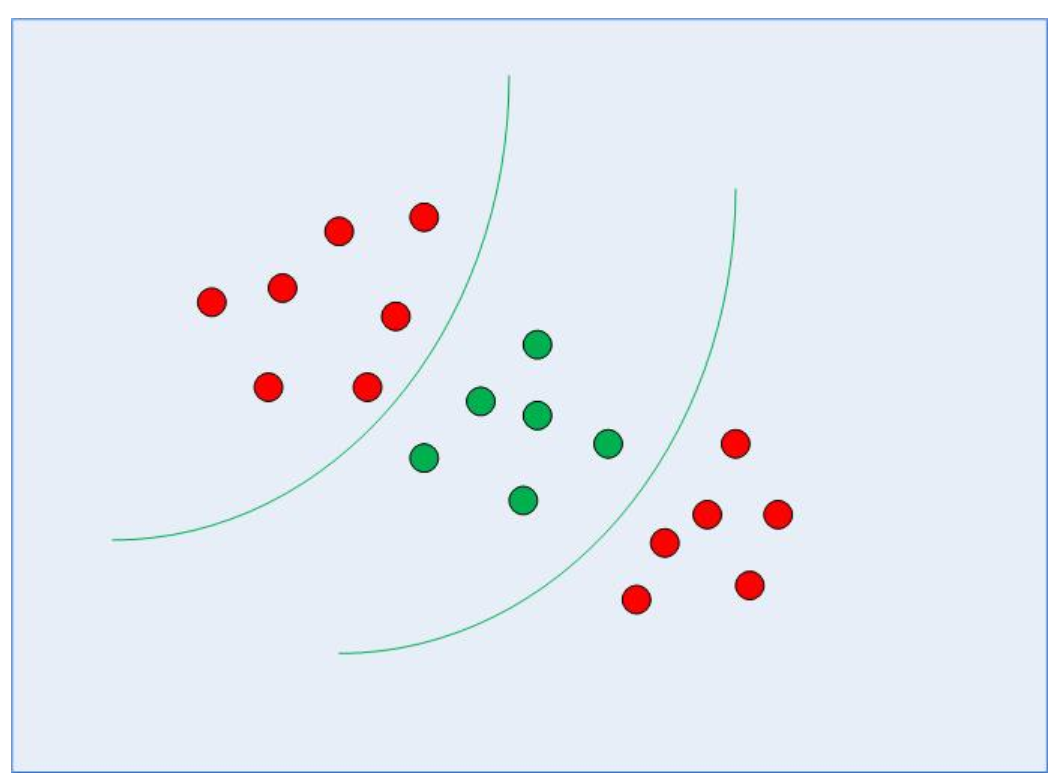


شکل ‏3‑8-مرز تصمیم‌گیری با بردارهای پشتیبان

در پشت پرده یافتن بردارهای پشتیبانی ریاضیات پیچیده‌ای پنهان‌شده، محاسبه فاصله بین مرز تصمیم‌گیری و بردارهای پشتیبانی و به حداقل رساندن این فاصله از آن جمله است. در این پژوهش به جزئیات ریاضیات نمی‌پردازیم، بلکه بیشتر خواهیم دید که چگونه SVM از طریق کتابخانه Python Scikit-Learn پیاده‌سازی می‌شود.

**-SVM با کرنل غیرخطی**

در بخش قبل مشاهده کردیم که چگونه می‌توان از الگوریتم SVM خطی برای یافتن مرز تصمیم‌گیری کرد. بااین‌حال، در مورد داده‌های غیرخطی جدا از هم، همانند شکل ‏3‑9، یک خط مستقیم نمی‌تواند به‌عنوان یک مرز تصمیم‌گیری استفاده شود.



شکل ‏3‑9-داده‌های غیرخطی جداگانه

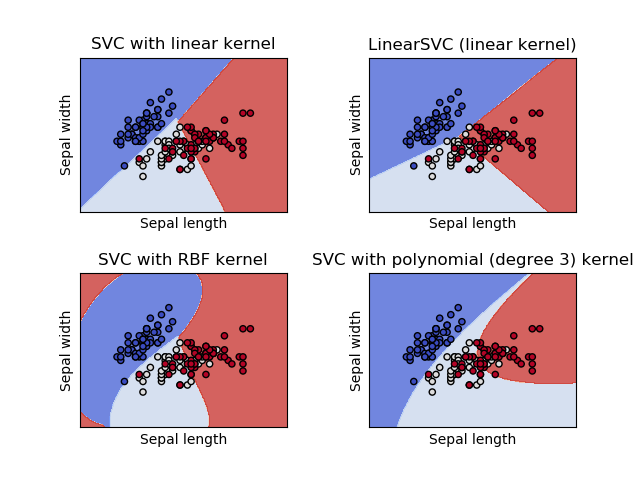
درصورتی‌که داده‌های غیرخطی از یکدیگر جدا باشند، نمی‌توان از الگوریتم ساده SVM استفاده کرد. در عوض، از یک نسخه اصلاح‌شده SVM بانام SVM غیرخطی استفاده می‌شود. این مدل داده‌های غیرخطی با ابعاد پایین‌تر تبدیل به داده‌های خطی با ابعاد بالاتر می‌شوند. بدین منظور از توابع کرنل استفاده می‌شود. سه تابع کرنل که در کتابخانه scikit-learn استفاده می‌شود عبارت‌اند از:

تابع چندجمله‌ای[[5]](#footnote-5) درجه d:

تابع شعاعی[[6]](#footnote-6):

تابع سیگموئید[[7]](#footnote-7):

در که از سایت کتابخانه scikit-learn استخراج شده است، تفاوت کرنل‌های مختلف و نحوه دسته‌بندی در الگوریتم SVM را مشاهده می‌کند.



-نحوه دسته‌بندی داده‌ها با کرنل‌های مختلف در الگوریتم SVM (scikit-learn, 2018)

باز هم، ریاضیات پیچیده‌ای در این مورد دخیل است، اما برای استفاده از SVM نیازی به نگرانی در مورد ریاضیات پیچیده نیست. بلکه می‌توانیم به‌سادگی از کتابخانه Scikit-Learn Python که برای پیاده‌سازی و استفاده از SVM غیرخطی مورد استفاده قرار می‌گیرد، استفاده کنیم. پارامترهای کرنل‌های مختلف از جلمِ درجه تابع چندجمله‌ای با سعی و خطا تعیین می‌شوند تا بهترین نتیجه به دست آید. ازآنجایی‌که تابع سیگوئید دو مقدار 0 و 1 را بر می‌گرداند برای طبقه‌بندی‌های دوتایی مناسب‌تر است.

**ارزیابی قدرت پیش بینی مدل ارائه شده**

پس از مدلسازی ، به منظور ارزیابی قدرت پیش بینی مدل ارائه شده داده ها به دو دسته آموزش و ارزیابی تقسیم خواهند شد . برای این منظور نخست یادگیری مدل پژوهش از طریق داده های آموزش انجام و سپس داده های ارزیابی به منظور محاسبه دقت و صحت الگوریتم ها روی داده هایی که تاکنون مشاهده نکرده است، استفاده خواهد شد. قابل ذکر است به منظور ارزیابی مناسب نیاز است که اجرای الگوریتم ها چندین بار تکرار شود. لذا زمانی که یک مجموعه داده در اختیار است به منظور ارزیابی نهایی نیاز است که بخشی از آن کنار گذاشته شود و از مابقی داده ها جهت یادگیری استفاده شود و مجددا دو مجموعه را تغییر داد و دوباره مدل مورد ارزیابی قرار گیرد. اعتبار سنجی 10 قسمتی یکی از روش هایی است که برای این منظور مورد استفاده قرار میگیرد در این روش مجموعه داده ها به صورت تصادفی به k قسمت مساوی، تقسیم می گردد. در اجرای اول قسمت اول از K قسمت به منظور ارزیابی، K −1 قسمت باقیمانده برای یادگیری استفاده میشود. در اجرای دوم قسمت دوم از K قسمت به منظور ارزیابی، K −1قسمت باقیمانده برای یادگیری استفاده میشود الگوریتم مورد اشاره K مرتبه به همین روال اجرا میگردد و در نهایت میانگین نتایج پس از اجرای روش اعتبارسنجی 10 قسمتی به عنوان شاخص ارزیابی مدل ارائه شده در سال t در نظر گرفته خواهد شد.

متدهای مختلفی جهت ارزیابی عملکرد مدل یادگیری و نتایج دسته بندی وجود دارد. ماتریس درهم ریختگی یک روش اندازه گیری عملکرد برای مساله طبقه بندی یادگیری ماشین است .ماتریس درهم ریختگی نتایج حاصل از طبقه‌بندی را بر اساس اطلاعات واقعی موجود، نمایش می‌دهد و تعداد پیش‌بینی‌های صحیح و نادرست هر کلاس را مشخص میکند. . جدول زیر ماتریس درهم ریختگی برای یک مدل دو تایی را نشان میدهد:

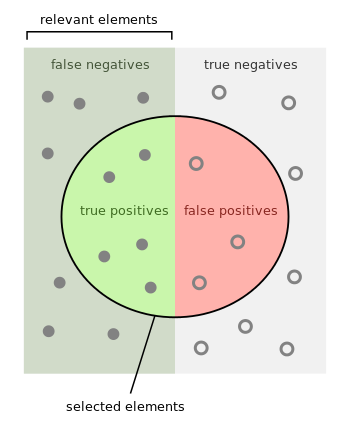
|  |  |  |
| --- | --- | --- |
| کلاس/تشخیص | تشخیص مثبت | تشخیص منفی |
| مثبت | TP | FN |
| منفی | FP | TN |

TP: کلاس مثبت که به درستی مثبت تشخیص داده شده است.

FN: کلاس مثبتی که به اشتباه منفی شده است.

FP: کلاس منفی که به اشتباه مثبت تشخیص داده شده است.

TN: کلاس منفی که به درستی مثبت تشخیص داده شده است.



نمایش ماتریس درهم ریختگی به صورت گرافیکی

در این مرحله دقت پیش بینی مدل ها با استفاده از روش های معتبر اعتبارسنجی بررسی خواهد شد.

در این مرحله دقت الگوی پیش بینی شده براساس داده های مالی، داده های متنی و تلفیق داده های مالی و متنی براساس الگوریتم های ماشین یادگیری با یکدیگر مقایسه و بهترین مدل جهت پیش بینی مدیریت سود مالی ارائه خواهد شد.

1. Clustring [↑](#footnote-ref-1)
2. Kumar& Ravi [↑](#footnote-ref-2)
3. Mohtaj [↑](#footnote-ref-3)
4. Burges [↑](#footnote-ref-4)
5. Polynomial [↑](#footnote-ref-5)
6. Radial basis function kernel - Wikipedia [↑](#footnote-ref-6)
7. Sigmoid [↑](#footnote-ref-7)