گزارش ۲:

یادگیری ماشین

Support Vector Machine

(SVM)

سید علی مجتبوی

گزارش ۲ - بررسی و تحویل انواع روش های SVM

دراین گزارش با انواع روش های SVM روی انواع مختلفی از دیتاست آشنا خواهیم شد

SVM خطی – دیتاست iris

دیتاست:

یک مجموعه داده شناخته شده در زمینه یادگیری ماشینی و آمار است. مجموعه زنبق شامل ۱۵۰ نمونه گل زنبق است که هر کدام به یکی از سه گونه ستوزا، ورسیکالر یا ویرجینیکا تعلق دارد. هدف اغلب استفاده از این ویژگیها برای پیشبینی گونههای گل زنبق میباشد. این مجموعه داده در بسیاری از کتابخانههای یادگیری ماشین، از جمله scikit-learn نیز موجود است. دراین گزارش نیز برای بارگزاری این دیتاست از کتابخانه scikit-learn استفاده شده است

۱. لینک دیتاست: https://archive.ics.uci.edu/dataset/53/iris

۲. مشخصات دیتاست:

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Biology	Classification
Feature Type	**# Instances**	**# Features**
Real	150	4

۳. فیچرهای دیتاست:

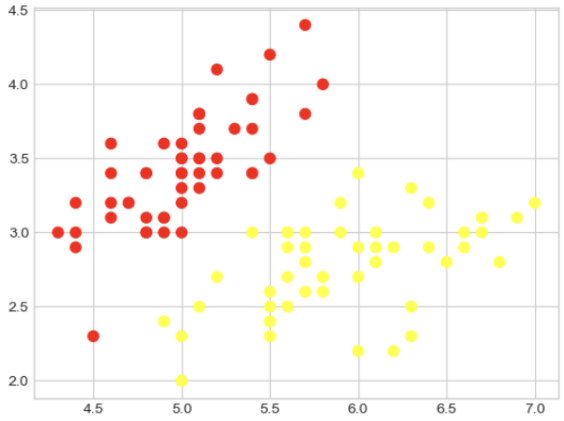
Variables Table

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
sepal length	Feature	Continuous			cm	no
sepal width	Feature	Continuous			cm	no
petal length	Feature	Continuous			cm	no
petal width	Feature	Continuous			cm	no
class	Target	Categorical		class of iris plant: Iris Setosa, Iris Versicolour, or Iris Virginica		no

ارائه مدل SVM خطی:

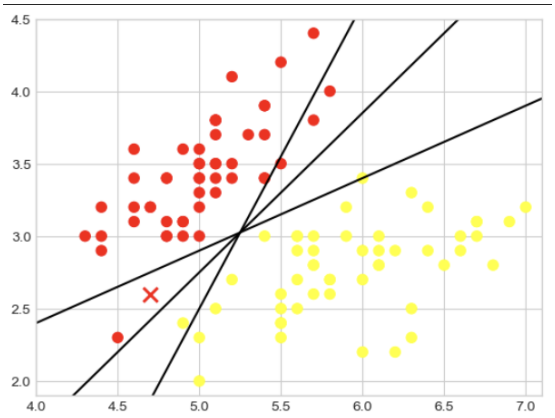
محور افقی نمودارها نمایانگر طول کاسبرگ و محور عمودی مشخص کننده عرض کاسبرگ میباشد

دایرههای قرمز نمونه Setosa دایرههای زرد رنگ نمونه های Versicolor را نمایش میدهد.

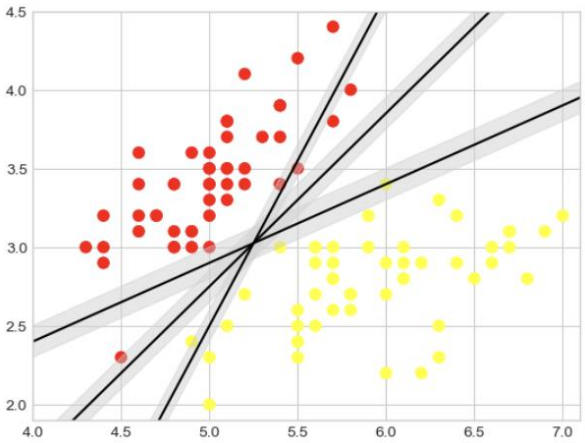


یک طبقه‌بندی‌کننده متمایز خطی سعی می‌کند خط مستقیمی را ترسیم کند که دو مجموعه داده را از هم جدا می‌کند. و در نتیجه یک مدل مناسب برای طبقه بندی ایجاد کنید. برای داده های دو بعدی مانند آنچه در اینجا نشان داده شده است، این عمل را میتوانیم بصورت شهودی و یا استفاده از رگرسیون انجام دهیم. اما بلافاصله با یک مشکل مواجه هستیم: بیش از یک خط تقسیم احتمالی وجود دارد که می تواند بین این دو کلاس کاملاً تمایز قائل شود!

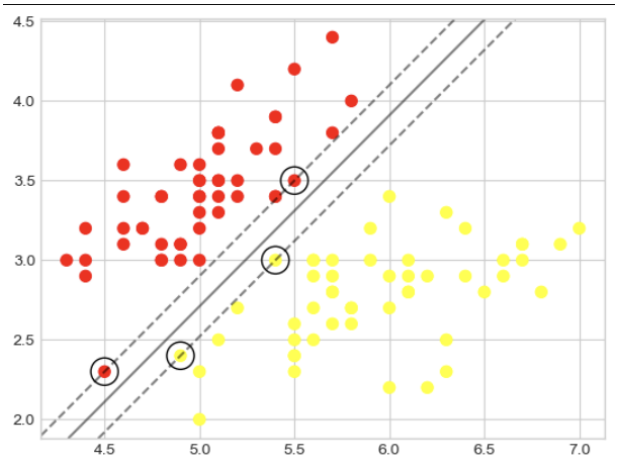
همانطور که در شکل زیر میتوانیم مشاهده کنیم سه خط جداکننده برای این مدل پیشنهاد شده است که ۲ دسته را بخوبی از یکدیگر جدا می کنند.اما بسته به اینکه کدام را انتخاب کنید، یک نقطه داده جدید (به عنوان مثال، نقطه ای که با "X" در این نمودار مشخص شده است) یک برچسب متفاوت به آن اختصاص داده می شود! بدیهی است که روش شهودی ما برای "خط کشی بین دسته ها" کافی نیست و باید کمی عمیق تر فکر کنیم.



SVM یک راه برای بهبود این موضوع ارائه می‌دهند. راه حل این است: به جای اینکه صرفاً یک خط با عرض صفر بین کلاس ها بکشیم، می توانیم در اطراف هر خط حاشیه ای با عرض کم تا نزدیکترین نقطه بکشیم.



این خط تقسیم است که حاشیه بین دو مجموعه نقطه را به حداکثر می رساند. توجه داشته باشید که تعدادی از نقاط مجموعه داده ها فقط margin را لمس می کنند: آنها با دایره های سیاه در شکل زیر نشان داده شده اند. این نقاط عناصر محور و اصل این روش هستند و به عنوان Support Vector شناخته می شوند این روش نیز به این دلیل به نامگذاری شده است

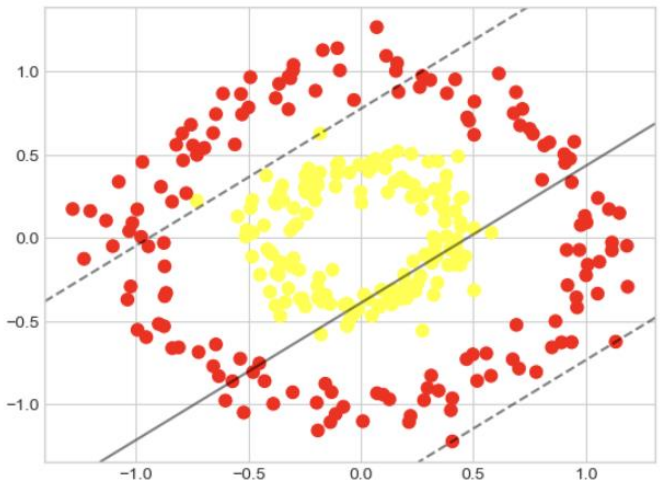


SVM غیرخطی – ساخت دیتاست با استفاده از تابع make_circles کتابخانه scikit-learn

دیتاست:

جهت تست بهتر روش SVM غیرخطی و استفاده از کرنل جهت بردن فیچر ها به فضای جدید میتوانیم با استفاده از تابع make_circles یک دیتاست تست با ۲ فیچر بسازیم

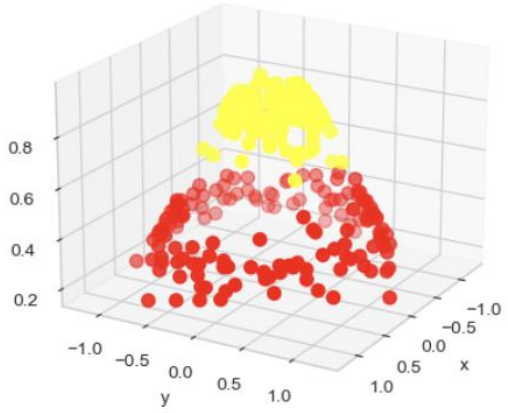
X, y = make_circles(300, factor=.4, noise=.1)



واضح است که هیچ خطی هرگز قادر به جداسازی این داده ها نخواهد بود. اما میتوانیم با استفاده از یک کرنل داده ها را به ابعادی بالاتر بفرستیم به طوری که یک جداکننده خطی کافی باشد. به عنوان مثال، یک تابع ساده که می‌توانیم از آن استفاده کنیم، radial basis function یا به اختصار RBF است:

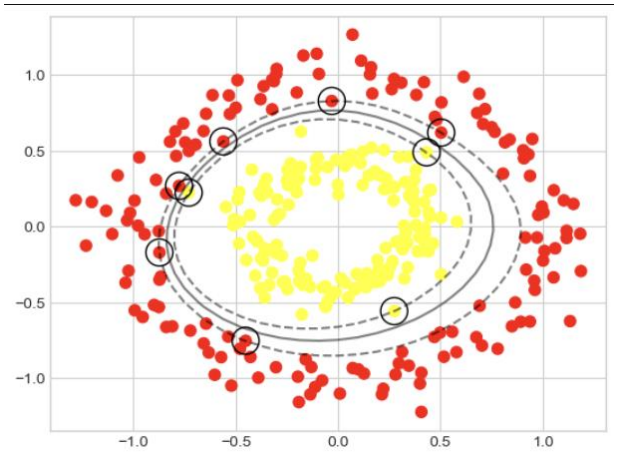
```
r = np.exp(-(X ** 2).sum(1))
```

ما می توانیم این بعد داده اضافی را با استفاده از یک نمودار سه بعدی تجسم کنیم. می‌توانیم ببینیم که با این بعد اضافی، داده‌ها به صورت خطی قابل تفکیک می‌شوند



در Scikit-Learn، می‌توانیم SVM با کرنل را به سادگی با تغییر کرنل خطی خود به یک کرنل RBF، با استفاده از هایپرپارامتر مدل اعمال کنیم:

```
clf = SVC(kernel='rbf', C=1E6)
clf.fit(X, y)
```



SVM غیرخطی Softening Margins

تا به اینجا گزارش با داده هایی سروکار داشتیم که بسیار تمیز بودند و نقاط دو مجموعه با یکدیگر همپوشانی نداشتند، اما اگر داده های شما مقداری همپوشانی داشته باشد. برای رسیدگی به این مورد، در پیاده‌سازی SVM مقداری بنام fudge-factor وجود دارد.به این معنا که اجازه می‌دهد درصورتی که دقت مدل بیشتر شود برخی از نقاط به margin داخل شوند، حساسیت مارجین به نقاط، توسط یک پارامتر تنظیمی کنترل می شود که اغلب به عنوان C شناخته می شود. برای C بسیار بزرگ، مارجین سخت گیر است و نقاط نمی تواند در آن داخل شوند. و برای C کوچکتر، مارجین ملایم تر است و می تواند برخی نقاط را در بر بگیرد.