دانشگاه شهید بهشتی

دانشکده علوم ریاضی

گروه علوم کامپیوتر

تمرین های سری سوم

درس داده کاوی

شکیلا جابری

99422046

اساتید محترم:اقای دکتر فراهانی-اقای دکتر خرد پیشه

در این تمرین از ما خواسته شده در مورد کرنل های پر کاربرد موجود در روش (اس وی ام) تحقیق کنیم. و ببینیم که چرا از اس وی ام استفاده میکنیم .و میخواهیم به این نکته بپردازیم که ایا در خصوص کرنل ها و استفاده از انها میتوان حکم کلی داد یا خیر ؟

حال برای پاسخ به این سوال و بررسی کردن سوال مطروحه نیاز داریم کمی در مورد اس وی ام صحبت کنیم و سپس وارد بحث کرنل ها شویم.

ساپورت وکتور ماشین یا ماشین های بردار پشتیبان

در ابتدا از این روش برای حل مسایل کلاسیفیکیشن دو کلاسه استفاده میکنیم و بعد به چند کلاسه تعمیم میدهیم.

در ابتدا فرض میکنیم با یک مسیله کلسیفیکیشن دو کلاسه مواجه هستیم. فرض میکنیم یک تعدادی نقطه داریم و دو تا فیچر داریم که x1 و x2 هستند.

مثلا x1=2 و x2=0 که در کلاس بنفش قرار گرفته و x1=0وx2=1.5 که در کلاس ابی قرار دارند .

حال چه کاری انجام میدهیم؟ در svm هدف این است که یک خط پیدا کنیم که بتواند data را به وسیله یک خط جدا کند . و به وسیله یک لاین این کار انجام دهد .

که این خط یک حالت خاصی از مفهومی به نام hyper plaine است که توضیح خواهم داد...

Feature space چیست ؟ بالا تر گفتیم دو کلاس به رنگ ابی و بنفش داریم حالا میخواهیم کاری کنیم که به وسیله این خط دو منطقه ی ابی و بنفش کاملا از هم جدا شوند .

اما گاهی نمیتوان به واسطه یک خط data هارا به طور دقیق از هم تفکیک کرد پس باید راهکار دیگری عرضه دهیم.

Data هایی که نمیتوان از یکدیگر جدا کرد خطی جدایی پذیر نیستند....

برای جدا کردن data توسط هایپر پلین یک فضایی داریم که به وسیله hyper plain میخواهیم سعی کنیم data هارا کلاس بندی کنیم..

در بعضی داده ها data کاملا خطی جدایی پذیر است .

در برخی دیگر میتوانند جدا شوند اما خطاهایی هم داریم

اما در برخی دیگر اصلا خطی جدایی پذیر نیستند

در نوع اول یک خط رسم میکنیم و فضای فیچر ها را به کلاس هایی جدا میکنیم و به شکل

مناسبی تقسیم بندی میکنیم .

در نوع دوم عملیات TAFENING انجام میدهیم یعنی یک مقدار شرایط را نرم میکنیم یعنی

اجازه میدهیم مقداری از داده ها که میدانیم خطا دارند وارد شوند .

در واقع در SVM خطی را در نظر میگیریم که بیشترین مارجین یا حاشیه را تولید کند .

حال معادله خط را مینویسیم :

Y=AX+B

برای ما A و B اهمیت دارند.

B+AX-Y=0

B0+B1X+B2Y=0

B0+B1X1+B2X2=0

حالا فرض میکنیم که این مارجین به اندازه M باشد

ما میخواهیم MAXIMIZE کنیم M را به ازای B های مختلف تا بفهمیم مارجین چه وقت MAX میشود.

پس این اولین مرحله در اس وی ام است .

در این مرحله داده ها کاملا از هم جدا میشوند و وقتی این اتفاق میافتد خطی را فیکس میکنیم که بیشترین مارجین را داشته باشد .

حال به بررسی مرحله ی دو میپردازیم :

DATA هایی که به طور کامل توسط یک خط جدا نمیشوند ...

دراین مرحله اجازه میزهیم تا یک سری داده MISS CIASSIFY شوند یعنی یک هزینه ای متقبل میشویم .

اگر مارجین خیلی ماکس بلشد عملا بدرد بخور نخواهد بود .

در این جا یک احتمالی نسبت میدهیم.

فرض کنید احتمال ما در CLASSIFICATION 51 درصد باشد.. و داده در کلاس یک است و اگر 48 درصد باشد در کلاس دو است.

حال فرض کنید که یک مقدار حساس باشد مثلا یک پزشک میخواهد احتمالات بیماری مریض را تشخیص دهد

مثلا سکته قلبی و مسمومیت دارویی و نحوه درمان هم کاملا متفاوت است.

سکته قلبی با احتمال 51 درصد و مسمومیت 48 درصد ...

وقتی درصدها به یکدیگر نزدیک هستند پزشک احتیاط بیشتری به خرج میدهد ..

حال وقتی تابعی داشته باشیم که مقادیر احتمال که میگیرد بالای 75 درصد یا کمتر از 25

درصد باشد احتمال و تابع خیلی بهتری خواهد بود..

اگر مارجین را بیش از حد بزرگ در نظر بگیریم داده ای ما عملا پرت میفتد ..

در مرحله سوم لاین را نمیتوان هیچ کاری کرد..

هیچ خطی وجود ندارد که بتواند داده ها را با هزینه کم برای خطا کنترل کند.

در این جا میایم چکاری انجام میدیم؟ اینجا دیگر مرز خطی بدرد ما نمیخورد .

کاری که میکنیم این است صفحه ی x1و x2 داریم..

میریم در فضای سه بعدی صفحه xy را در نظر میگیریم.. نقاط ابی و بنفش در صفحه پخش هستند

یک برگه a4 در نظر میگیریم و خمش میکنیم...

حالا این برگه میتواند داده های مارا جدا کند .

Kernels and support vector machins

فرض کنید دو بردار داریم .

فرض کنید data frame به این صورت است :

X1=(x11,x12,…,x1p)

Xi=(xi1,xi2,…xip

Xn=(xn1,xn2,…xnp)

بجای این ایکس ها فریم هارا قرار میدهیم.

کرنل چیست؟

ضابته های متفاوتی دارند که در ان خط به رویه تبدیل میشود .

چرا از کرنل استفاده میکنیم ؟

چون این داده هایی که به صورت خطی جدایی پذیرنیستند را میتوانیم با رویه های تراز و داده هارا جدا کنیم .

حال اگر svm بیشتر از دو کلاس داشتیم چه کاری انجام میدهیم

دو روش داریم

1-ova

2-ovo

یکی در مقابل همه

یکی در مقابل یکی

که بیشتر از روش دوم استفاده میکنیم.

در سوال یک در خصوص کرنل های پرکاربرد اس وی ام سوال شده است .

الگریتم های اس وی ام از مجموعه ای از توابع ریاضی که به عنوان کرنل تعریف میشوند استفاده میکنند .

وظیفه ی کرنل این است که داده هارا به عنوان ورودی گرفته و انها به شکل مورد نیاز تبدیل کند .

الگریتم های مختلف اس وی ام از انواع مختلف توابع کرنل استفاده میکند .

این توابع میتوانند انواع متفاوتی داشته باشند:به عنوان مثال خطی-غیر خطی –چند جمله ای

-تابع پایه شعاعی یا rbf وسیگمویید .

توابع کرنل برای داده های ترتیبی – نمودار ها –متن ها – تصاویر و همچنین بردار ها

معرفی میشوند .

پرکاربرد ترین نوع تابع کرنل ار بی اف.زیرا دارای پاسخ محلی و متانهی در کل بازه محور ایکس است .

توابع کرنل ضرب داخلی بین دو نقطه ذر یک فضای ویژگی مناسب را برمیگردانند

بنابراین با هزینه ی محاسبتابی کم حتی در فضا های با ابعاد بالا مفهومی از شباهت را تعریف میکنند

نمونه هایی از کرنل های رایج اس وی ام :

کرنل های چند جمله ای که در پردازش تصویر پرکاربرد است .

کرنل گاوسی برای اهداف عمومی است و هنگامی که هیچ دانش پیشینی در مورد داده ها وجود ندارد استفاده میشود.

تابع پایه ی شعاع گاوسی:این کرنل برای اهداف عمومی کاربرد دارد و هنگامی که هیچ

دانش پیشینی در مورد داده ها وجود نداشته باشد به کار میرود

در ادامه ی سوال اول از ما خواسته شده تا بررسی کنیم که چرا به طور کلی از ایده ی اس وی ام استفاده میکنیم .

الگریتم اس وی ام از مدل های قدرتمند و همه کاره ماشین لرنینگ است

این الگریتم میتواند برای طبقه بندی های خطی یا غیر خطی رگرسیون و شناسایی داده های پرت استفاده شود

چند مفهوم و کاربرد :

طبقه بندی با linear svm

قبلا گفتیم که دو کلاس به راحتی میتوانند توسط یک خط جدا شوند گاهی ممکن است مدلی که مرض تصمیم گیری است به قدری بد عمل کند که اصلا نتواند کلاس هارا به درستی

جدا کند.

سوال دو: در این سوال از ما خواسته شده که با استفاده ار دیتا ست ارسالی روش svm را اجرا کنیم.

با استفاده از پکیج skleran

در ابتدا نیاز داریم که مثل همیشه پکیج هارا فراخوانی کنیم که با استفاده از پکیج pandas و numpy این کار را انجام میدهیم و سپس دیتا ست ارسالی را به کمک فایل csv میخوانیم که با توجه به جدول موجود در فایل لرسالی داده ها قابل مشاهده هستند..

با خواند دیتا ست میتوانیم دید کلی نسبت به دیتا پیدا کنیم

گفتیم به دو پکیج نیاز داریم . 1-numpy که برای کار با دیتا به ان نیاز داریم

2-pandas که برای کارهای پردازشی به ان نیازمندیم

در قدم اول باید با یک فایل csv دیتا را بخوانیم .

در خط سوم با دستور shap میتوانیم بفهمیم چه تعداد نمونه داریم که این دستور تعداد ستون هارا مشخص میکند و به ما اطلاع میدهد که باید یک سری کار اماری انجام دهیم

پس میرویم به سراغ پاسخ دادن به سولات خواسته شده

اول از همه نیاز داریم که دیتا ست را به دو قسمت test و train تقسیم میکنیم

بعد مدل svm را با کرنل خواسته شده در صورت سوال بر روی داده ی train انجام میدهیم ..و بعد خروجی را برای داده های test بدست میاوریم .

در قسمت اول سوال یک توضیح دادیک که باید از یک تابع احتمال درsvm استفده کنیم تا یک احتمال و درصد دقیقی به ما بدهند برای تصمیم گیری در مورد داده ها که بیماری قلبی و مسمویت را مثال زدیم.

در این سوال میبینیم که دقت مدل ما تقریبا 96 درصد است.

که در خط ششم کدها این کا را انجام داده ایم.

و در خط هفتم با استفده از دستور accuracy دقت مدل را بررسی کردیم.

بعد از انجام این کار مدل svm را با کرنل خواسته شده در صورت سوال بر روی داده های train اجرا میکنیم و دوباره دقت مدل را با استفاده از دستور قبل بدست میاوریم 92که درصد است.

سوال سوم: در این سوال از ما خواسته شده چند مورد از حالت های مختلف مثل کرنل ها را بررسی کنیم و نتایج را گزارش کنیم

در خط دوازدهم از تابع sigmoid به عنوان نمونه استفاده کردیم برروی داده های اموزش اجرا میکنیم و بعد خروجی که بدست میاوریم را بر روی داده های تست ازمایش میکنیم.

و سپس دقت این مدل را هم بدست میاوریم که 21 درصد است.

این درصد اصلا درصد خوبی نیست و بسیار پایین است..

در خط پانزدهم مدل دیگری را ارزیابی میکنیم

مدل poly : این مدل و پارامتر درجه 3بر روی داده های اموزش اجراا میکنیم و بعد خروجی آن را برای داده های تست ازمایش میکنیم.

و بعد دقت این مدل را هم محاسبه میکنیم.

که 94 درصد است و عدد بسیار خوبی است و این کار را در خط هفدهم کدها انجام دادیم..

بعد از این دوباره باز هم از همین کرنل استفاده میکنیم اما با یک تغییر

این کار برای پارامتر درجه 5 انجام میدهیم. بر روی داده های اموزش اول اجرا میکنیم و بعد خروجی که بدست میاوریم را بر روی داده های تست دوباره مثل قبل ازمایش میکنیم.

دوباره طبق روال قبل ماتریس آن را بدست میاوریم و با استفاده از آن دقت مدل را بدست میاوریم.

این دقت 94 درصد است که نسبتا درصد خوبی است و در خط بیستم فایل ارسالی این کار را انجام داده ایم.

سوال 4: در این سوال از ما خواسته شده تا سعی کنیم مبحث soft margin و hard margin را برای یوال دوم مورد بررسی قرار دهیم و نتایج را گزارش کنیم..

ابتدا نیاز داریم که توضیح مختصری در مورد این دو موضوع بدهیم..

تئوری و پیاده سازی hard margin svm:

داده خطی دو کلاسه جدایی پذیر داده ای که در آن بتوان داده های دو کلاس را بایک خط جدا کرد.

و ساده ترین مسئله کلاس بندی استکه در آن هدف پیدا کردن مرز خطی بهینه است. روشهای مختلف با رویکرد های متفاوتی مرز بهینه را پیدا میکند.

تئوری و پیاده سازی soft margin svm

این الگوریتم در بین الگوریتم های موجود بهینه ترین خط ممکن را پیدا میکند با وجود اینکه این الگوریتم رئیکرد بسیار خوبی دارد در دو حالت با مشکل مواجه میشود

1-داده های آموزش داده کلاس کاملا تفکیک پذیر نباشند

2-نمونه های نویزی در داده های آموزشی وجود داشته باشد

برمیگردیم به حل سوال

در دیتا ست هایی که به صورت خطی جدایی پذیر هستند soft margin بهتر عمل میکند

به همین دلیل اگر از hard margin استفاده کنیم. اگر یک داده ی پرت داشته باشیم میتواند در عملکرد ما تاثیر گذاشته و کار مارا خراب کند.

برای بررسی این دو نوع مارجین باید پارامتر c را تغییر دهیم.

هر چفدر این پارامتر را افزایش دهیم از هارد به سمت سافت میرویم.

حال مثل روال سابق ماتریس را با استفاده از کد بدست می اوریم و مجدد دقت را اندازه گیری میکنیم.

اول c را 0.1 در نظر میگیریم

دقت در این قسمت برابر با 90 درصد خواهد بود که این را در خط بیستو سوم کدهای فایل ارسالی مشاهده میکنیم..

پس از این c را تغییر میدهیم و مقدار آن را به 0.5 میرسانیم و ماتریس را به دست آورده دقت را محاسبه میکنیم

دقت این مدل برابر با 92 درصد خواهاد بود.( خط بیست و ششم )

حال دوباره c را تغییر میدهیم و مقدار آن را برابر 5 قرار میدهیم.

ماتریس را بدست میاوریم و سپس دقت مدل را میسنجیم

دقت مدل برابر با 94 درصد است. (خط بیست و نهم)

حال مجددا مقدار c را به 10 تغییر میدهیم ماتریس میکشیم و بعد دوباره دقت را بدست میاوریم .

دقت مدل 94درصد خواهد بود.خط سی و دوم.

نتایج و درصدهایی که تا به الان بدست اورده ایم نشان میدهد که روش soft margin روش مناسبی است.

سوال پنج-قسمت الف:

از ما خواسته شده در خصوص مهندسی ویژگی یک سری کار روی دیتا ست انجام دهیم.

اول از همه از روش bining استفاده میکنیم و طبق خواسته سوال سه دسته برای قدرت باتری موبایل در نظر میگیریم.

و طبق جدول موجود در فایل ارسالی دیتا ست را برای این سه دسته میتوانیم مشاهده کنیم.

حال نیاز داریم فیچر هایی که تا به الان با آن ها کار میکردیم حذف میکنیم و یک مدلسازی جدیدی روی این دیتا ست انجام میدهیم.

ماتریس آن را بدست میاوریم.

دقت مدل 79 درصد و تقریبا نزدیک به 80 درصد در نظر میگیریم (خط چهل)

این نتیجه ای که ما میکیریم به ما نشان میدهد که استفاده کردن از روش bining باعث میشود که عملکرد مدل کاهش یابد.

سوال پنج قسمت دوم: در این سوال خواسته شده برای فیچرهای کتگوریکال موجود در دیتا ست one hot encoding را اعمال کنیم .

ابتدا توضیح مختصری در مورد این موضوع میدهیم:

روش کدبندی one hot یکی اآز پرکاربرد ترین رویکردها است و عملکرد آن بجز در مواردی که متغیر دسته ای مقادیر خیلی زیادی بگیرد بسیار خوب است

معمولا از این روش برای متغیرهایی که بیشتر از پانزده مقدار متفاوت بگیرند مناسب نیست.

در برخی از مواردی که تعداد متغیر ها کمتر است نیز امکان دارد گزینه مناسبی نباشد)

کدبندی one hot ستون های دودویی باینری جدیدی میسازد که هر یک مربوط به یکی از مقادیری هستند که متغیر به خود میگیرد.

در دیتا ست گاهی با ستون هایی رو به رو هستیم که در آن اعداد اشاره به گونه خاصی دارند.. به این ستون ها کتگوریکال میگوییم.

برای بالا بردن کارایی مدل از one hot encoding استفاده میکنیم.

دیتا ست پس از استفاده از این روش به شکل جدول موجود در فایل ارسالی در خواهد بود

(خط چهل و هشتم).

حال دوباره فیچرهای قبلی را حذف میکنیم و مدلسازی را دوباره انجام میدهیم.

سوال پنج قسمت ج:

در این قسمت از ما خواسته تا بررسی کنیم ایا استفاده از تبدیل هایی مثلlog transform را اعمال کنیم. و بگوییم چرا از این ها استفاده میکنیم.

میخواهیم در این سوال از تبدیل لگاریتمی و نمایی استفاده کنیم و آن ها را نرمال کنیم.

در خط پنجاه و هشتم این کار را انجام داده ایم با استفده از clock –speed .

میشود فهمید که توزیع داده ها ما نرمال نیستند .

در خط پنجاه و نه از تابع battry-power استفاده کرده ایم و درمیابیم که باز هم توزیع داده ه نرمال نیست 0

پس تابع لگاریتمی روی آن پیاده سازی میکنیم.

در خط شصتم تابع m-dep را تعریف کرده ایم0

باز هم متوجه میشویم که توزیع داده ها نرمال نیستند و باز هم تابع لگاریتمی روی آن پیاده میکنیم.

سپس تابع taik-time را تعریف میکنیم.

این تابع نشان میدهد توزیع داده ها نرمال نیست و کار قبل را تکرار میکنیم.

دیتا ست ما به شکل موجود در کد خط شصت و دو تغییر خواهد کرد.

بعد از تبدیل ستون داده ها را به دو قسمت تقسیم میکنیم.

تست و train. و دوباره مدل سازی میکنیم.

ماتریس بدست می اوریم و دقت مدل را اندازه گیری میکنیم که ماتریس ما در خط 66 و دقت ما در خط 67 قرار دارد و بربر است با 78 درصد .

استفاده کردن از این روش باعث باعث کم شدن دقت مدل ما میشود .

سوال پنجم قسمت د: از ما خواسته شده یک فیچر جدید بسازیم به نام مساحت یا حجم گوشی

فیچر مساحت را به دیتا ست اضافه میکنیم .

نتیجه نشان میدهد که اضافه شدن فیچر مساحت دقت مدل را به شدت کاهش میدهد .

حالا تمام روش های قسمت پنجم را با هم ادغام میکنیم و نتایج را بررسی میکنیم .

دیتا ست جدید در خط 78 کدهای ما قابل مشاهده است.

بعد مدل سازی را انجام میدهیم و نتایج را بررسی میکنیم.

نتایج بدست امده نشان میدهد که اعمال روش ها روی مدل باعث کم شدن بازدهی مدل ما میشود.