گزارش تمرین شماره 3

درس داده کاوی

جناب آقای دکتر فراهانی

جناب آقای دکتر خرد پیشه

استادیار جناب آقای شریفی

اشکان صفوی سهی

۹۸۴۲۲۰۹۶

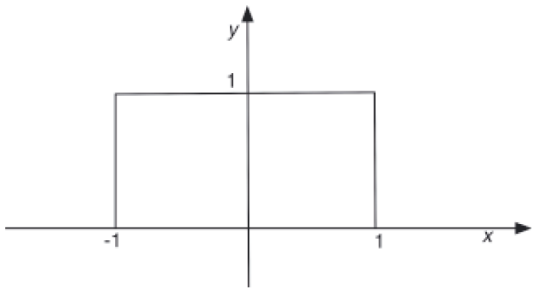
سوال ۱) الگوریتم های SVM از مجموعه ای از توابع ریاضی که به عنوان کرنل تعریف می شوند، استفاده می کنند. وظیفه کرنل این است که داده ها را به عنوان ورودی گرفته و آن ها را به شکل مورد نیاز تبدیل کند. الگوریتم های مختلف SVM ، از انواع مختلف توابع کرنل استفاده می کنند. این توابع می توانند انواع متفاوتی داشته باشند. به عنوان مثال خطی ، غیرخطی ، چند جمله ای ، تابع پایه شعاعی (RBF) و سیگموئید.توابع کرنل ، برای داده های ترتیبی ، نمودار ها ، متن ها ، تصاویر و همچنین بردار ها معرفی می شوند .پرکاربردترین نوع تابع کرنل، RBF است. زیرا دارای پاسخ محلی و متناهی در کل بازه محور x است.توابع کرنل ، ضرب داخلی بین دو نقطه در یک فضای ویژگی مناسب را برمی گردانند. بنابراین ، با هزینه محاسباتی کم، حتی در فضاهای با ابعاد بالا، مفهومی از شباهت را تعریف می کنند.

قواعد کرنل

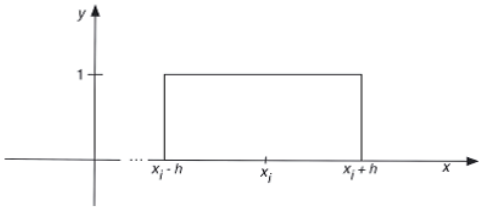
تعریف کرنل یا یک تابع پنجره به شرح زیر است:

$$K\left(\overline{x}\right)=\begin{matrix}1 & \text{if}\left\|\overline{x}\right\|\leq 1\\ 0 & \text{otherwise}\end{matrix}$$

مقدار این تابع، در داخل یک شکل بسته به دامنه ۱ و مرکز مبدا مختصات برابر ۱ و در غیر این صورت ۰ است. همانطور که در شکل زیر نشان داده شده است:



برای xi ثابت، در داخل شکل بسته با دامنه h و مرکز xi ، تابع برابر است با K (z – xi / h) = 1 و در غیر این صورت ۰ می باشد. همانطور که در شکل زیر نشان داده شده است:



بنابراین ، با انتخاب آرگومان K(.) ، پنجره را حرکت داده اید تا با دامنه h در مرکز xi قرار گیرد.

نمونه هایی از کرنل های SVM

بیاید برخی از کرنل های رایج مورد استفاده در SVM ها و کاربرد های آن ها را مشاهده کنیم:

۱ -کرنل چند جمله ای

این کرنل در پردازش تصویر پرکاربرد است. معادله آن به صورت زیر است:

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d$$

که در آن d درجه چند جمله ای است.

۲ -کرنل گاوسی

این یک کرنل برای اهداف عمومی است. و هنگامی که هیچ دانش پیشینی در مورد داده ها وجود ندارد استفاده می شود .معادله آن به صورت زیر است:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

۳ -تابع پایه شعاعی گاوسی (RBF)

این کرنلی برای اهداف عمومی کلربرد دارد. و هنگامی که هیچ دانش پیشینی در مورد داده ها وجود نداشته باشد، مورد استفاده قرار می گیرد. معادله آن به صورت زیر است:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2)$$

و برای $\gamma > 0$

گاهی اوقات با استفاده از پارامتر زیر استفاده می شود:

$$\gamma = 1/2\sigma^2$$

۴ -کرنل RBFلاپلاس

این هم یک کرنل برای اهداف عمومی است. و هنگامی که هیچ دانش پیشینی در مورد داده ها وجود ندارد استفاده می شود .معادله آن به صورت زیر است:

$$k(x,y) = \exp\left(-\frac{||x-y||}{\sigma}\right)$$

۵ -کرنل تانژانت هیپربولیک(tanh)

می توانیم از آن در شبکه های عصبی استفاده کنیم .معادله مربوط به آن عبارت است از:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\kappa \mathbf{x_i} \cdot \mathbf{x_j} + c)$$

در برخی موارد (نه همیشه $k > 0$ و $c < 0$)

٦ -کرنل سیگموئید

می توان این کرنل را در شبکه های عصبی مورد استفاده قرار داد. معادله مربوط به آن عبارت است از:

$$k(x,y) = \tanh(\alpha x^T y + c)$$

۷ -کرنل تابع بِسِل (Bessel)از نوع اول

ما می توانیم از آن برای حذف مقطع عرضی در توابع ریاضی استفاده کنیم. معادله آن عبارت است از:

$$k(x,y) = \frac{J_{v+1}(\sigma||x-y||)}{||x-y||^{-n(v+1)}}$$

که J تابع بسل از نوع اول است.

۸ -کرنل پایه شعاعی ANOVA

ما می توانیم از آن در مسائل رگرسیون استفاده کنیم. معادله مربوط به آن عبارت است از:

$$k(x,y) = \sum_{k=1}^{n} \exp(-\sigma(x^k - y^k)^2)^d$$

۹ -کرنل splineخطی بصورت یک بعدی

این کرنل، هنگام کار با بردارهای بزرگ داده پراکنده ، کاربرد زیادی دارد. این کرنل اغلب در دسته بندی متن مورد استفاده قرار می گیرد. کرنل spline همچنین در مسائل رگرسیون عملکرد خوبی دارد. معادله آن عبارت است از:

$$k(x,y) = 1 + xy + xy\ min(x,y) - \frac{x+y}{2}\ min(x,y)^2 + \frac{1}{3}\min(x,y)^3$$

سوال ۲) جواب سوال در فایل ضمیمه موجود می باشد.

سوال ۳) جواب سوال در فایل ضمیمه موجود می باشد.

سوال ۴) ما انتظار داریم که SVM با حاشیه نرم حتی زمانی که مجموعه داده های آموزش به طور خطی قابل تفکیک باشد ، بهتر عمل نماید .دلیل این امر این است که در hard margin ، یک حاشیه جداگانه می تواند مرز را تعیین کند ، که باعث می شود طبقه بندی بیش از حد به نویز موجود در داده حساس شود .در دیتاستهایی که بصورت خطی جدایی پذیر هستند، soft-marginم ناسبتر است، به این خاطر که اگر از hard margin استفاده کنیم، وجود یک داده پرت میتواند در عملکرد الگوریتم طبقه بندی ما تاثیر منفی داشته باشد. برای بررسی soft-margin و hard-margin باید پارامتر C را تغییر میدهیم. هرچه پارامتر C را افزایش دهیم از هارد مارجین به سمت سافت مارجین حرکت خواهیم کرد .نتایج تغییرات پارامتر cرا در جدولی به شرح زیر جمع آوری شده است:

C parameter	Accuracy
1	**0.846250**
5	**0.8675**
10	**0.873750**
60	**0.876875**

طبق نتایج ذکر شده در جدول بالا soft margin برای این دیتاست مناسب میباشد.

سوال ۵)الف)جواب سوال در فایل ضمیمه موجود می باشد.

سوال ۵)ب)در دیتاست با ستونهایی مواجه خواهیم شد که در آنها اعداد اشاره به گونه یا نوع خاصی دارد، این ستونها کتگوریکال هستند. اگر در دیتاست ستونهای عددی نیز وجود داشته باشد، در این صورت مدل یادگیری ماشین گیج شده و در تشخیص اعداد دچار مشکل میشود. برای حل این مشکل و بالا بردن کارایی مدل، از روش one hot encoding استفاده میشود، به این صورت که به ستونهای کتگوریکال یک ماتریس اختصاص میدهد. تعداد ستون های هر ماتریس برابر با تعداد نوع های آن ستون است. در هر سلول مقدار 1 یعنی وجود داشتن آن نوع و مقدار صفر یعنی عدم وجود آن نوع خاص .
روند استفاده از این روش در فایل ضمیمه موجود میباشد.

سوال ۵)ج) جواب سوال در فایل ضمیمه موجود می باشد.

سوال ۵)د) جواب سوال در فایل ضمیمه موجود می باشد.

سوال ٦) جواب سوال در فایل ضمیمه موجود می باشد.