Support Vector Machines

باز شناسی الگو

جلسه نهم

ماشین های بردار پشتیبان

فهرست مطالب

- مقدمه
- ایده Margin و تاثیر آن در قابلیت تعمیم
- تشریح تئوری support vector machines (svm)
- تعمیم ایده SVM به مسائلی که دارای جدایی پذیری خطی نیستند
- معرفی کرنل های متداول
- تعمیم ایده SVM برای مسائل چند کلاسه

جایگاه بحث

- مسائل مطرح در این بحث:
  - بازنمائی نمونه ها
  - طراحی کلاسیفایر با یک دسته از نمونه ها و یک نحوه انتخاب شده از بازنمائی
  - نحوه تخمین و ارزیابی قابلیت کلاسیفایر طرح شده

- مسائل عملی:
  - تخمین (تخمین توزیع و یا تخمین پارامتر ها)
  - بهینه سازی

مقدمه:

- دیدیم که یکی از مهمترین بخش های کار، طراحی کلاسیفایر است.
- صورت مسئله طراحی کلاسیفایر:
  - با در دست بودن تعدادی از مثال ها (یا نمونه ها)، هدف آن است که سیستم(ساختار ریاضی) خاصی طرح شود که بتوان به کمک آن، نمونه ها را از یکدیگر جدا نمود
  - ممکن است نمونه های یاد شده به صورت با مربی یا بدون مربی در دست قرار گیرند
  - مسئله عموما به یک مسئله بهینه سازی تبدیل می شود

مقدمه ...

- در فاز عملی، تفاوت کلاسیفایر ها، در معیار آن برای بهینه *است*.
- کم کردن خطای طبقه بندی نمونه های آموزشی، معمولترین معیار است اما این مشکل وجود دارد که این معیار معیاری گسسته است، در حالیکه اکثر روش های جستجو(بخصوص فرمولی) نیاز مند معیار یا تابع هزینه ای پیوسته دارند.
- این مسئله، بخصوص در زمانی که تعداد نمونه ها کم باشد مشکلات عملی زیادی ایجاد می کند:
  - مشکل زمان جستجو(عدم امکان استفاده از مشتق گیری)
  - مشکل وقتی که تعداد خطا های رخ داده به مقدار می نیمم(صفر)رسیده است (عدم وجود راهنما برای ایجاد تغییر در پارامتر های قابل تنظیم)

مقدمه...

- اما واقعیت آن است که حتی در بین نمونه های آموزش هم، ارزش آنها در یادگیری و از آن مهم تر در تعمیم با هم مساوی نیستند.
- جدا سازی برخی نمونه های آموزشی بسیار آسان و برای برخی دیگر این امر دشوار است.
- عدم توجه به قابلیت تعمیم(کمینه کردن خطا روی نمونه های تست)، با توجه به عدم در اختیار بودن واقعی آنها قابل توجیه است.

تبیین مشکل

- در فرآیند بهینه سازی، معمولترین معیار مجموع(یا متوسط) مربعات خطا روی نمونه هاست.
- در بسیاری از موارد، با تغییر یک پارامتر، تعدادی از نمونه هائی که قبلا سیستم روی آنها دچار خطا می شد، قابل جدا سازی خواهد شد و در عوض(احتمالا) برخی از نمونه ها که قبلا درست طبقه بندی می شد، در وضعیت خطا قرار می گیرند.
- راه حل معمول:
  - در صورتی که تعداد نمونه هائی که در وضعیت موفق قرار می گیرند، بیش از نمونه هائی باشد که دچار مشکل می شود، ما تغییر پارامتر را اعمال کنیم

سوالات اساسی:

- آیا در طرح کلاسیفایر، ارزش همه نمونه های یادگیری، در ایجاد قابلیت تعمیم یکسان است؟ به تعبیر دیگر آیا نمونه هایی وجود دارد که در صورت یادگیری عملکرد صحیح در آنها، دسته بزرگتری از داده های دیده نشده را بتوان به درستی کلاسه بندی کرد؟
- آیا همه طرح ها (کلاسیفایرها)یی که روی نمونه های آموزشی، درصد عملکرد(و یا حتی عملکرد) مشابهی دارند، قابلیت تعمیم یکسانی هم دارند؟
- آیا با کمینه (یا حتی صفر شدن)خطا روی نمونه های آموزشی، نمی توان به نحوی، پارامتر های طراحی کلاسیفایر را تغییر داد که عملکرد آن روی نمونه های دیگر بهبود یابد؟

بعد VC (Vapnik-Chervonenkis)

- محقق روسی بنام Vladimir Vapnik در سال 1965 ، با پرداختن به سوالات یاد شده، گامی بسیار مهم در طراحی کلاسیفایرها برداشت و نظریه آماری یادگیری را بصورت مستحکم تری بنا نهاد.
- مبنای نظریه مزبور، بر قائل شدن تفاوت بین نمونه های مختلف در حین یادگیری است

- قبلا در بحث طراحی کلاسیفایر خطی در حالت چند بعدی، با ابر صفحات برخورد داشتیم.
- دیدیم که ممکن است (در صورت وجود جدایی پذیری خطی که معادل است با امکان رسیدن به خطای صفر در یادگیری، با صفحات جدا کننده یا کلاسیفایر خطی) بتوانیم تعداد زیادی از صفحات چند بعدی را برای جدا سازی پیدا کنیم.
- در چنین وضعیتی، می توان این سوال را طرح کرد که از بین این صفحات، کدام یک را باید انتخاب کنیم؟

فاصله یا حد

- بطور عقلایی، به نظر می رسد بهترین انتخاب برای صفحه جدا ساز، صفحه ای باشد که فاصله آن با نزدیکترین نقطه از هر کلاس، ماکزیمم باشد.
- به عبارت دیگر اگر بتوان دو نقطه (در حالت دو کلاسه) را پیدا کرد که هر یک به یکی از کلاس ها متعلق باشند و نزدیکترین فاصله را در بین دو کلاس داشته باشند، این دو نقطه بهتر است مبنای کار قرار گیرد.
- اگر چه این امر برای ممکن است برای جدا سازی داده های یادگیری، بی تاثیرباشد، ولی برای داده های دیگر(تست) می تواند بسیار موثر باشد.

ایده اساسی SVM

- *Vapnik* ثابت کرد که بعد *VC* برای طبقه‌بندی کننده‌هایی از نوع ابر صفحات کانونی، دارای یک کران بالاست که این کران بالا با توان دوم نرم بردار وزن یعنی نسبت مستقیم دارد.
- در واقع اگر ما این *نرم* را محدود کرده و مینیمم کنیم، معیاری به نام بعد *VC* طبقه‌بندی کننده را مینیمم کرده‌ایم
- در این حالت تخمین ما از مقدار ریسک واقعی بصورت احتمالی دقیق‌تر بوده و خاصیت تعمیم دسته‌بندی کننده بیشتر خواهد شد.
- *این ایده، مبنای ساخت کلاسیفایری قرار گرفت که به SVM موسوم است.*

Support vector machines

- اگر فعلا توجه خود را بر جدائی پذیری خطی و جدا ساز های خطی متمرکز کنیم؛ تشریح ایده به کلاسیفایری موسوم به SVM منجر خواهد شد.
- svm برای حل مشکلات کلاسیفایر معمول، در گستره وسیعی از کاربرد ها استفاده می شود که در آن فضای مشخصه را از طریق ا بر صفحه (Hyperplane) بهینه، به 2 کلاس متمایز تقسیم می کند .
  - در حالت کلی، (hyperplane) یک سطح هندسی با ابعاد متغیر می باشد و تعداد ابعاد یک (hyperplane) ماهیت آن را مشخص می کند .
  - به لحاظ عملی، ابعاد چنین صفحه ای، به توانائی برای سنجش و اندازه گیری محدود می شود، هر چند به لحاظ ریاضی ممکن است چنین محدودیتی وجود نداشته باشد
  - اما افزایش بعد (با تعداد ثابت نمونه ها) چنانکه می داانیم منجر به پدیده نحوست ابعاد می شود

ادامه

- یکی از هدفهای SVM برای شناسایی الگوها این است که در عین دوری از پدیده هایی چون نحوست ابعاد، یک شبه صفحه (hyper plane) بهینه و مطلوب برای مینیمم سازی تابع هزینه پیدا کند
  - یک hyperplane مطلوب، به اندازه کافی از 2 کلاس دور است
  - چنین امری، با داشتن margin مطلوب بیان می شود
  - یعنی اگر حداکثر marginرا داشته باشیم فاصله بین کلاس ها و صفحه جدا کننده طوری است که ماکزیمم جدا سازی بین کلاس ها را در صورت جابه جایی نقاط (یعنی نقاط جدید یا تست) تامین می کند.

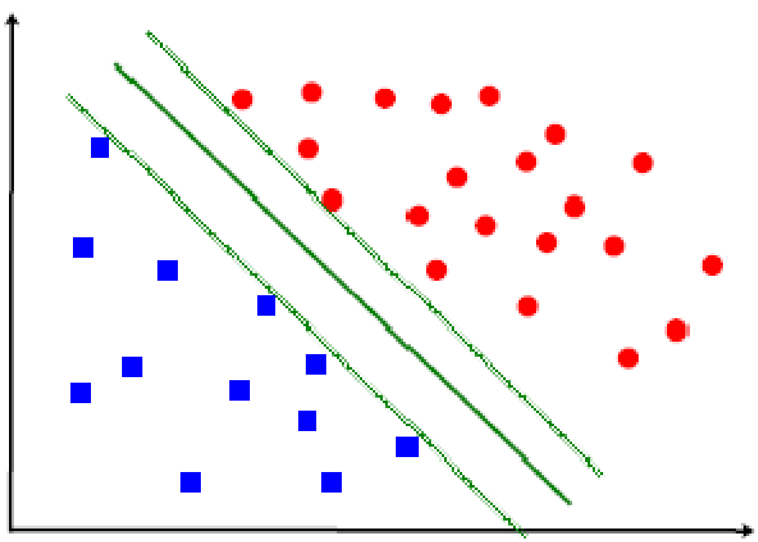
Margin و بردار های پشتیبان

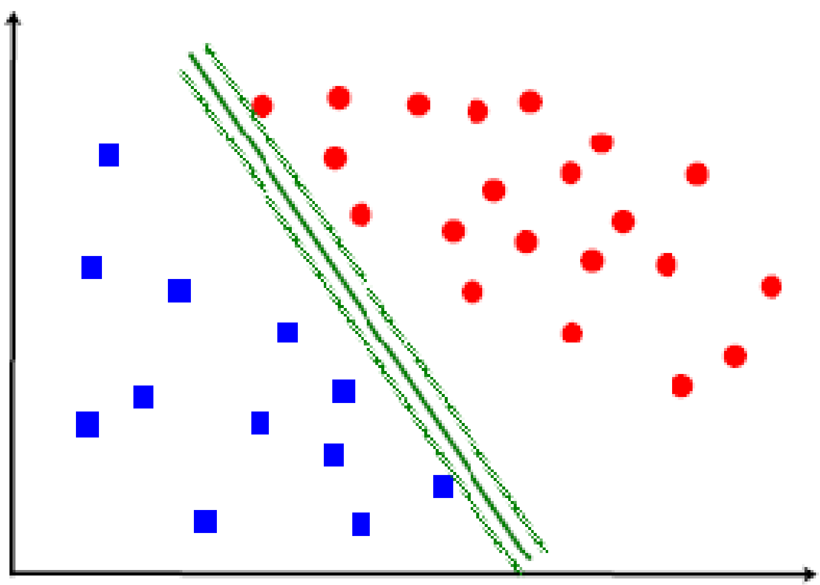
- چنانکه گفته شد، ایده اصلی برای افزایش قدرت تعمیم یک کلاسیفایر(حتی با وجود اینکه معیار مبتنی بر شمارش خطا روی نمونه ها کاهش نیابد) آن است که از بین کلاسیفایر های ممکن، کلاسیفایری انتخاب شود که margin بزرگتری داشته باشد.
- ابتدائی ترین ایده Margin عبارتست از حداقل فاصله بین ابرصفحه جدا ساز و اعضای کلاس ها (که با بردار یا نقطه نشان داده می شوند)
- بردارهائی(نقطه هایی) که margin از روی آنها محاسبه می شود، Support vectors یا بردار های پایه(پشتیبان) نام دارند.

نقاط (بردار ها) پایه

- SV ها نقش اساسی در کلاسیفایر SVM ایفا می کند و مزیتی که SVM نسبت به دیگر تکنیک ها دارد،این است که تخمین احتمال خطا در آن(و مبنای بهینه سازی)، به جای اینکه به تمام نمونه ها و درنتیجه تمام ابعاد فضای اولیه وابسته باشد، تنها وابسته به Support Vectors می باشد
- بنابراین SVMS را می توان برای ابعاد بالا مورد استفاده قرار داد بدون اینکه به نحوست ابعاد دچار شد.

نمایش مفهوم ابر صفحات با حداکثر margin





نمایش مفهوم margin وبردار های پشتیبان

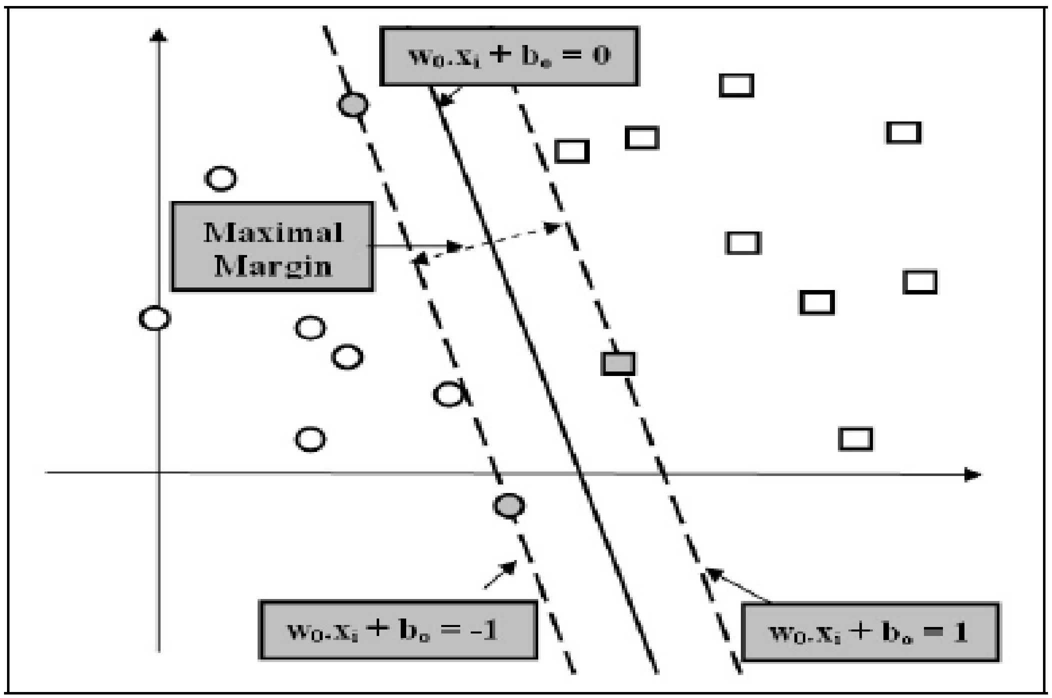


Figure 3: Hyperplane with Support Vectors and Maximal Margins shown.

تعبیر هندسی برای ابر صفحه های بهینه

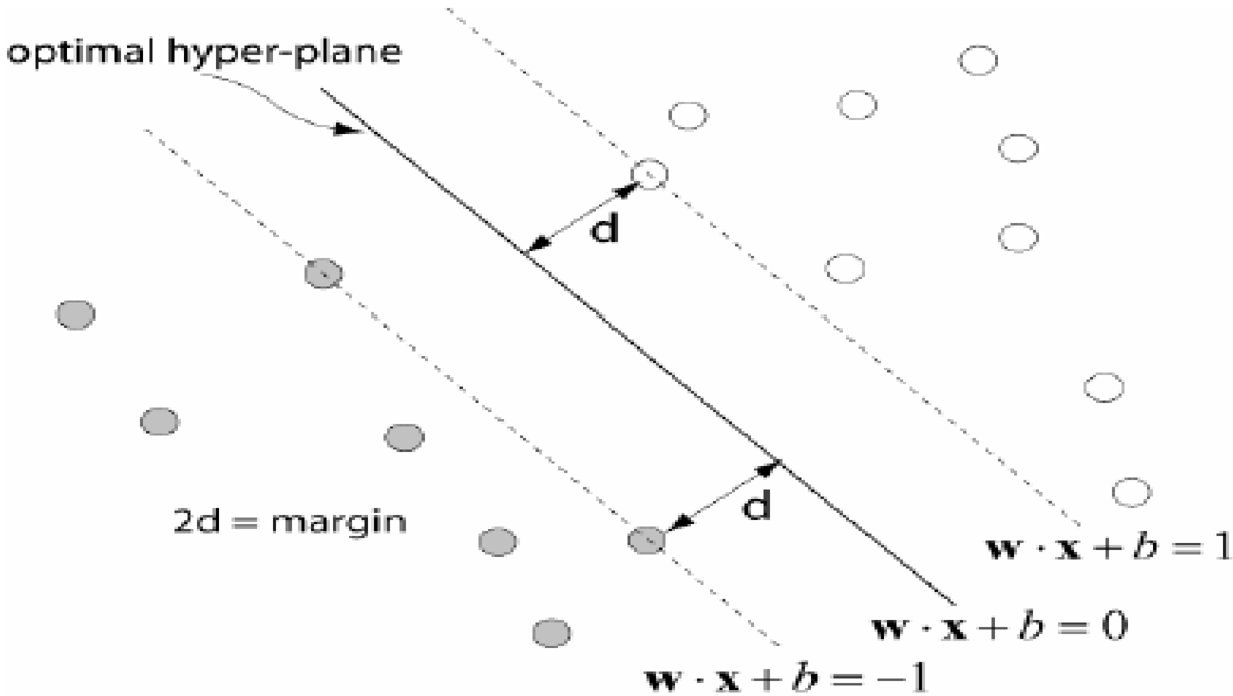


Fig. 1 Geometric interpretation of the optimal hyper-plane

فرآیند بهینه سازی

- برای اینکه فرآیند طراحی کلاسیفایر SVM بتواند شکل اتوماتیک به خود بگیرد آن را به یک فرآیند بهینه سازی تبدیل میکنند.
- در این فرآیند، ابتدا مفهوم یا معیاری اخذ می شود که باید بیشینه یا کمینه شود و سپس ارتباط این معیار با پرامتر های در دست تبیین می شود.
- در صورت امکان بر اساس مفهوم تغییرات (مشتق) به یافتن مقادیر بهینه مبادرت می شود

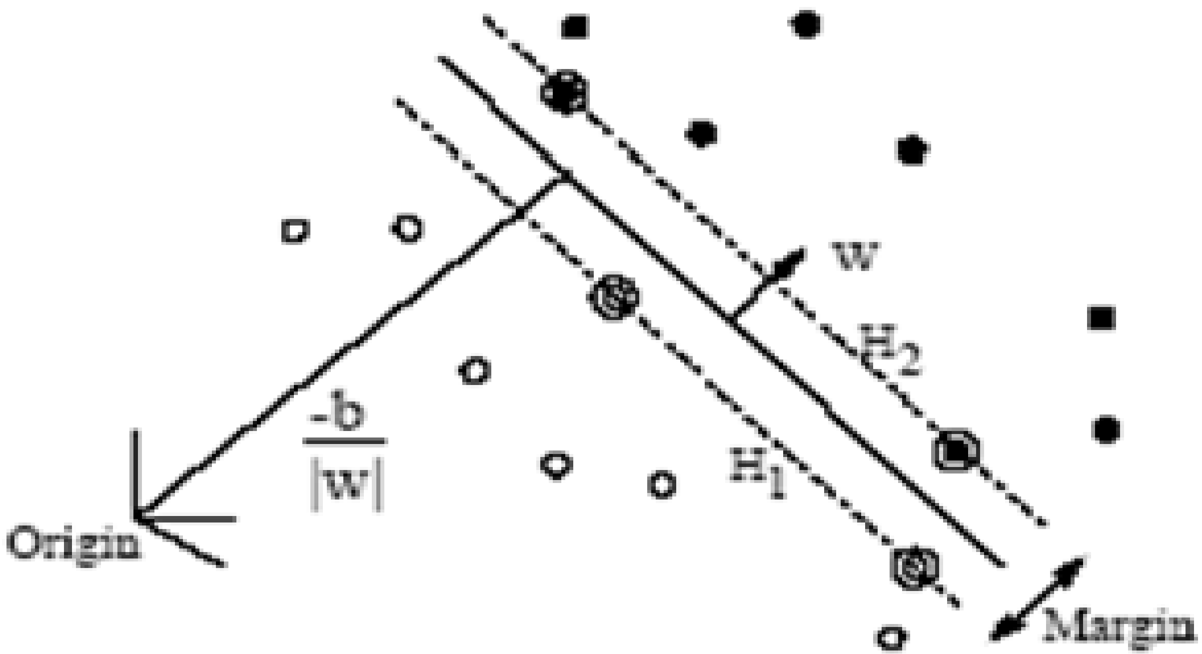
اتخاذ معیار بهینه سازی

- به بردارهایی که در یک طرف hyperplane با حداقل margin ویا فراتر قرار می گیرند می تواند عدد 1 یا بزرگتر را نسبت داد و به بردارهایی که در سمت دیگر hyperplane دیگر، یعنی زیر حداقل margin و یا فروتر قرار می گیرند می توان عدد 1- یا کوچکتر را نسبت داد .
- فرض بر این است که بین این دو صفحه هیچ نمونه ای قرار نمی گیرد
- می توان بین این دو صفحه، صفحه ای میانی جست.
- برای Support vector برچسب 1 و 1- را داریم .
  - (۱) $y_i=+1$ برای $x_i.w+b\geq+1$
  - (۲) $y_i=-1$ برای $x_i.w+b\leq+1$

که میتوان این دو را باهم تلفیق کرد و به نامساوی زیر رسید:

$$y_i(x_i.w+b)-1\geq0 \qquad (۳)$$

نمايش ابر صفحه ميانی يا مبنا



معیار بهینه سازی، Maximum margin:

- با توجه به شکل نزدیکترین نقاط مثبت x_+ و x_- به صورت $x_+.w+b=+1$ و $x_-.w+b=-1$ مشخص می شوند.
- برای محاسبه margin داریم:

$$\frac{1}{2}\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\mathbf{x}_+ + \frac{b}{\|\mathbf{w}\|} - \frac{\mathbf{w}}{\|\mathbf{w}\|}\mathbf{x}_- - \frac{b}{\|\mathbf{w}\|}\right) = \frac{1}{2\|\mathbf{w}\|}(\mathbf{w}\mathbf{x}_+ + b - \mathbf{w}\mathbf{x}_- - b) = \frac{1}{\|\mathbf{w}\|}$$

- به ازای تمام i همواره داریم: $y_i(x_i.w+b)-1\geq 0$
- با توجه به معادلات بالا تمام نمونه های هر کلاس در فاصله ای بزرگتر یا مساوی با $1/\|w\|$ نسبت به hyperplan قرار میگیرند.

پیاده سازی بهینه سازی
: (oh) optimal hyperplane

- Margin به سادگی برابر $2/\|w\|$ می باشد. بنابراین می توانیم ماکزیمم margin را با مینیمم کردن $\|w\|^2$ بدستت می اوریم.

  - Minimize $\tau(w) = \frac{1}{2}\|w\|^2$
  - Subject to $y_i\left((w \cdot x_i) + b\right) \geq 1, i = 1, \ldots, l$

- حال برای مینیمم کردن $\|w\|^2$ مشکل آن است که روابط نامساویند و ضمنا در گام بعدی (کلاسیفایر غیر خطی) مفهوم W به این شکل وجود ندارد.
- برای رفع این مشکلات، در مینیمم کردن، (تعریف تابع هزینه) سراغ فرمول لاگرانژ می رویم

دلایلی برای استفاده از فرمول لاگرانژ

- اول : محدودیت های $y_i(x_i.w+b)-1\geq 0$ با محدودیت های ضرایب لاگرانژ ، که کار کردن با آنها آسانتر خواهد بود، جایگزین می شود.
- دوم:دراین فرمول بندی، داده های training درالگوریتم های تست و trainingواقعی به شکل ضرب نقطه ای بردارها ظاهر می شود.
- این یک خاصیت مهم و حیاتی است که به ما اجازه می دهد، یک پروسه را به موارد غیرخطی تعمیم دهیم.
- بنابراین ما ضرایب لاگرانژمثبت $a_i,\ a=1\ldots,\ell$ را معرفی می کنیم که هر یک از آنها نقش یکی از مشخصه ها را تبیین می کند.

فرمول لاگرانژ

$$L(w,b,\alpha)=\tfrac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i(y_i\cdot((x_i\cdot w)+b)-1)$$

- اکنون ما باید L را نسبت به b,w مینیمم کنیم :

$$\frac{\partial}{\partial b}L(w,b,\alpha)=0 \quad \frac{\partial}{\partial w}L(w,b,\alpha)=0$$

محدودیت ها:

- اینکه گرادیان Lp نسبت به b,w صفر شود منجر به محدودیت های زیر می شود:

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i$$

با اعمال شرط زیر داریم:

$$\sum_{i=1}^{l} a_i y_i = 0$$

$$\alpha_i \cdot [y_i((x_i \cdot w) + b) - 1] = 0, \quad i = 1, \ldots, l$$

- از شرط فوق می توان نتیجه گرفت که:
  - a_i برای بردارهای پشتیبان (SV) مخالف صفر می باشدو برای دیگر نقاط training صفر می باشد.

فرمول لاگرانژ.....

- با اعمال محدودیت های قبل به فرمول لاگرانژ به معادله زیر می رسیم

$$L_D = \sum_i \alpha_i - \tfrac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$$

$$\alpha_i \cdot [y_i((x_i \cdot w) + b) - 1] = 0, \quad i = 1, \ldots, l$$

فرمول لاگرانژ..

- توجه شود که برچسب های مختلف لاگرانژ تاکید بر این نکته دارد که ($L(w,b,a)$ برای اصلی و D برای دوگان) از تابع هدف یکسان ولی با شرایط متفاوت؛ دو راه-حل،از دو مسیر زیر بدست می آید:
  - مینیمم کردن $L(w,b,a)$
  - ماکزیمم کردن L_D
- محاسبه a_i ، به حل معادله بالا، موسوم به quadratic problem منجر شود.

حل لاگرانژ

- برای بدست آوردن ضرایب لاگرانژ لازم است که معادله L_D ماکزیمم شود(البته با در نظر گرفتن شرایط زیر)

• Maximize $$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\tfrac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)$$

- با در نظر گرفتن شرایط زیر:

$$\alpha_i\geq 0,\ i=1,\dots,l,\ \ and\ \sum_{i=1}^{l}\alpha_i y_i=0$$

فرمول لاگرانژ.

- ضرایب لاگرانژ a_i در محاسبه تابع تصمیم دخیل هستند.
- تابع تصمیم Decision function:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i \cdot (x \cdot x_i) + b\right)$$

- معادله مهم بالا زمانی برقرار می باشد که اطلاعات training به خطی جداپذیر هستند. حال اگر اطلاعات بطور خطی جدا پذیر نباشند معادله بالامفید واقع نمی شود

جدائی پذیری خطی و غیر خطی، مفهوم کرنل

- در حالتی که جدا پذیری بصورت خطی نباشد (جدا پذیری غیر خطی) ایده اصلی این است که نمونه ها را به یک فضای با بعد بالا (feature space) فضای مشخصه نگاشت دهیم که در فضای جدید مشخصه ها، نمونه ها می توانند به صورت خطی از هم جدا شوند .
- این امر نیاز به اعمال یک تابع هسته (کرنل) را به همراه خواهد آورد.

نمایش دستیابی به جدایی پذیری خطی، برای مسئله ای که دارای این خاصیت نیست، به کمک نگاشت

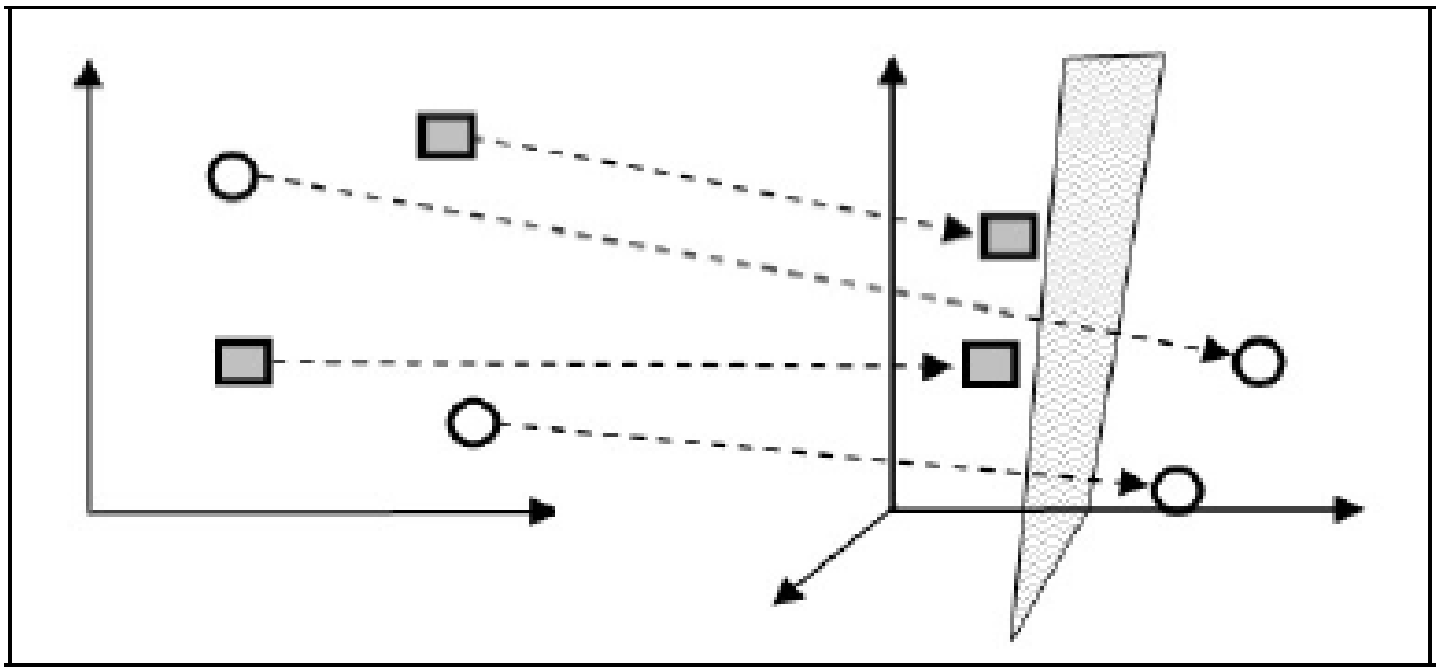


Figure 4: Kernel Mapping from input to feature space.

توابع کرنل متداول

- در این روش از تابع غیر خطی کرنل (kernel) k(φ1.,φ2.) استفاده می کنیم . یعنی به جای معادلات x و x_i از k(x,y)=(φ1 (x). φ2 (y)) استفاده می کنیم . تابع کرنل هم بصورت مختلف بر حسب کاربرد وجود دارد معمولا" توابع کرنل که در عمل استفاده می کنیم عبارتند از :
- Linear kernel

$$K(\mathbf{x}_j,\mathbf{x}) = \mathbf{x}\cdot\mathbf{x}_j$$

- Polynomial kernel

$$K(\mathbf{x}_j,\mathbf{x}) = (\gamma\mathbf{x}\cdot\mathbf{x}_j + r)^d,\ \gamma > 0$$

- Gaussian kernel

$$K(\mathbf{x}_j,\mathbf{x}) = \exp(-\gamma\|\mathbf{x}-\mathbf{x}_j\|^2),\ \gamma > 0.$$

- Sigmodial kernel
- انتخاب γ بسته به این است که بهترین کلاسیفایر ممکن را برای اطلاعات training بدست آوریم.

....

- چنان که قبلا نیز ذکر شد، با استفاده ازاین توابع، ابتدا به جدایی پذیری خطی دست پیدا می شود و سپس با استفاده از مفهوم SVM بهترین تفکیک حاصل می گردد.
- با کلاسیفایر غیر خطی(استفاده از کرنل) تابع تصمیم به صورت زیر می باشد

$$\boxed{f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right)}$$

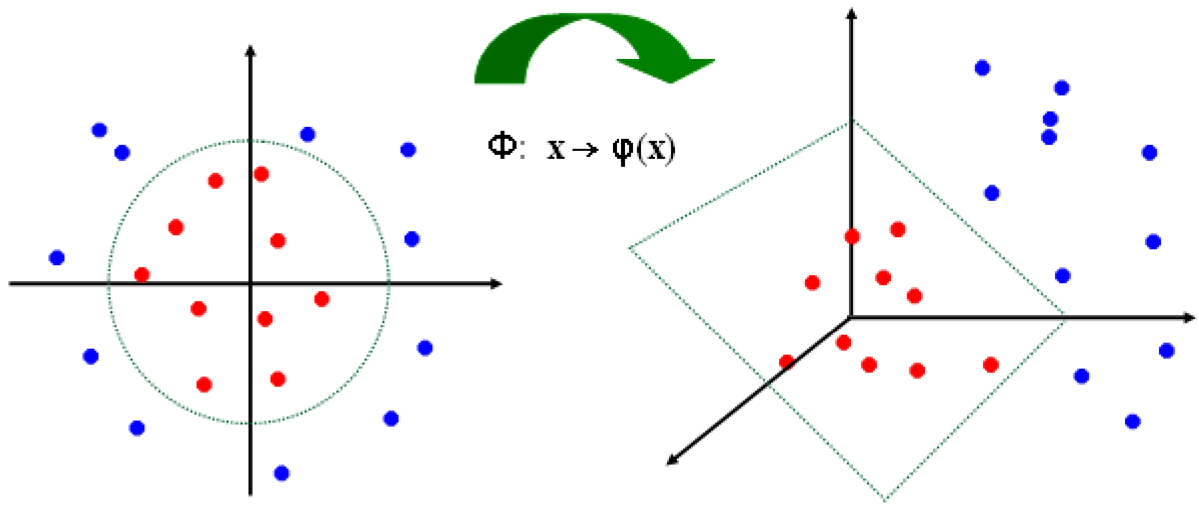
خلاصه مراحل اصلی کار

- **نحوه بکارگیری (Classification) SVM :**
  - آماده سازی ماتریس اطلاعات داده ها
  - انتخاب تابع کرنل مناسب
  - به اجرا در آوردن training با استفاده از حل quadratic برای محاسبه ..
  - در نهایت داده های testing با استفاده از Support vectors کلاس بندی می شود .

ماشین بردار پشتیبان غیرخطی

- ابتدا باید با نگاشت داده به یک فضای ویژگی آنها را بصورت خطی جداپذیر نمود:

$$x \rightarrow \phi(x)$$



انواع کرنل درماشین بردار پشتیبان غیرخطی

- در ماشین بردار پشتیبان انتقال داده از فضای ورودی به فضای داده توسط توابع کرنل صورت می گیرد.

- کرنل چند جمله ای(polynomial)

- کرنل RBF

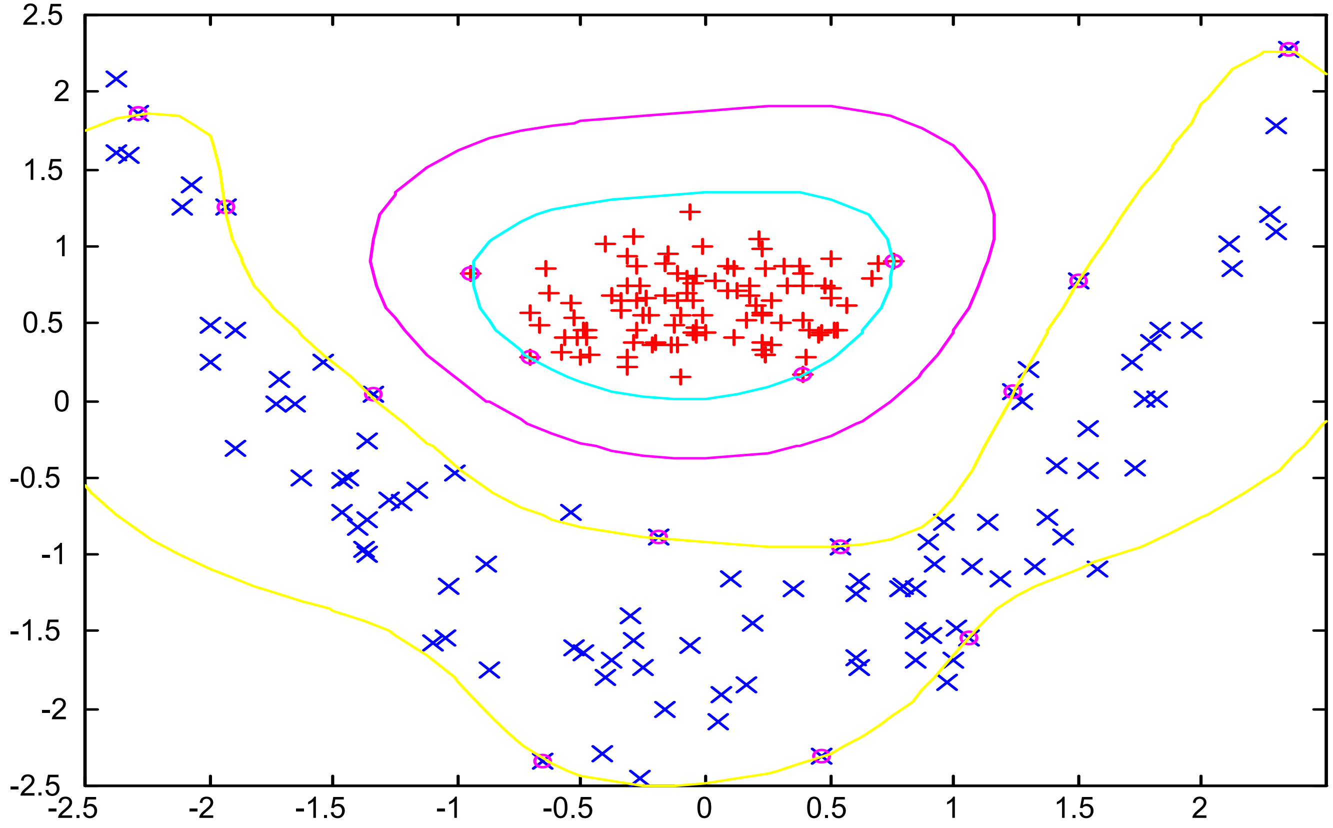
- کرنل تانژانت هیپربولیک

$$k(x,y)=(x.y+1)^p \quad p=2,3,...$$

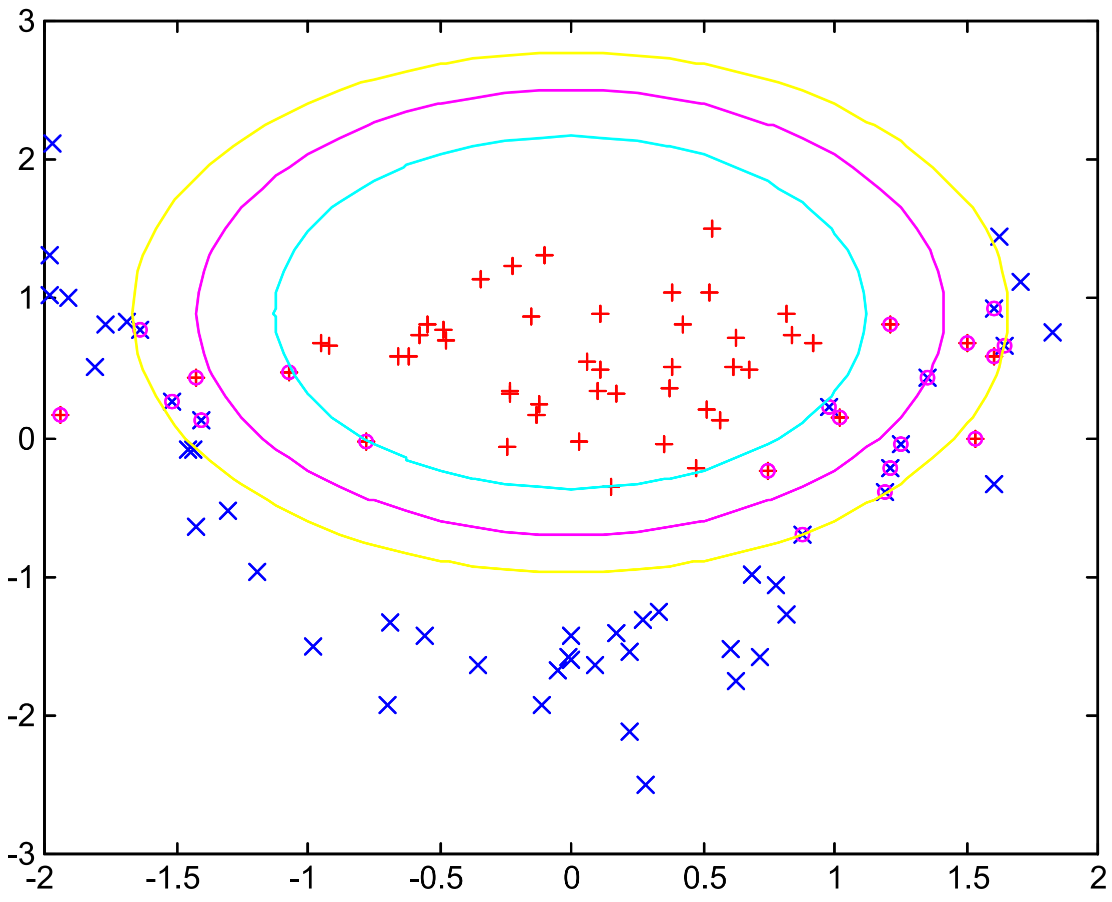
$$k(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

$$k(x,y)=\tanh(x.y+\theta)$$

کرنل RBF



کرنل quadratic



خلاصه بحث

- در پیاده سازی فرآیند های یادگیری در طراحی کلاسیفایر، مشکل اساسی آن است که معیار ها معمولا به خطا (loss) بستگی دارد که مقداری گسسته است و لذا
  - اولا با فرآیند های جستجوی فرموله، همخوانی ندارد
  - ثانیا در گام تغییر آن، هیچ راهنمائی برای ایجاد تغییرات وجود ندارد
- به همین دلایل مفهومی به نام حد (Margin) ارائه می شود
- این مفهوم در کنار لزوم رفع مشکلاتی مانند نحوست ابعاد که به لزوم کاهش نمونه ها منجر خواهد شد، ارائه مفاهیمی جدید ضروری است

خلاصه

- دو مفهوم اساسی :
  - مفهوم اول: بعد VC
  - مفهوم دوم: نقاط(بردار های) پایه SV
- با تکیه بر جدائی پذیری خطی در گام اول، به این نتیجه می رسیم که اندازه بردار وزن می تواند به عنوان معیار قرار گیرد
- با توجه به مشکلات عملی، برای فرآیند بهینه سازی استفاده از ضرایب و فرمول لاگرانژ در دستور کار قرار گرفت
- برای تعمیم روش به مسائل غیر خطی، اسفتفاده از توابع کرنل توصیه گردید

پایان

سوال؟