

ML-13

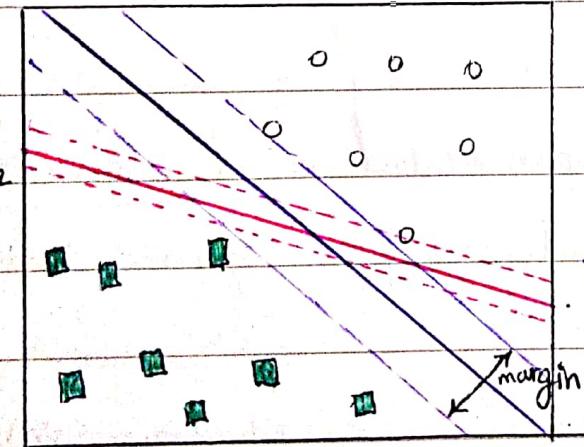
ML-1-Dr.Kavousi

Introduction to statistical learning theory, kernel methods,

and Support Vector Machines (SVM)

اگر خواهیم ساده شدی حدسی SVM را بینم، باید کل نزدیک باشد.

B<sub>1</sub>



دسته باند خلف در مدلی دو بعدی معرفی شده است.

به دنبال طراحی linear classifier هستیم.

محیط توانیم از خود رام از خطوط  $B_1$  و  $B_2$  استفاده کنیم.

به دنبال خطی صفتی که با فاصله margin را داشته باشد.

خسین دو طبقه را از همی خفوطا داشتیم.

کند.

- جای اینکه بین حیثیت دوستم، مسافت بین خطوط فرایند آنچه دارد
- در واقع observation صفت خارا را داشت که نزدیک در گردیدی هم داشتم
- توانم بخوبی مسافت بین خطوط را بتوانم -
- optimization کرد و در واقع بتوانم margin را بزرگ کنم
- این است که پس از تبادل اینکه توانم margin را بزرگ کنم و تبدیل به دنبال راه حل بگیرم

## History of SVM

- The study on statistical learning theory was started in 1960s by Vapnik.
- statistical learning theory is the theory about Machine learning principle from a small sample size.
- support vector Machine is a practical learning method based on statistical learning theory.

a simple SVM could beat a sophisticated neural networks with elaborate features in a handwriting recognition task.

از جمله این گذشتگانی که از جواد در حقیقت گفته، SVM و Deep learning را در زیر می‌داند.

## Empirical Risk / True error

فرض کنید  $f$ ، تابعی است که حس تواند مسندی کند. که خطای دو بعدی دیگر صفر دوستی به تکمیل حس تواند مسندی باشد. (ایده دوستی دو بعدی، کرد دوستی آنکه)

classification Engine

$$f: \mathbb{R}^N \rightarrow \{\pm 1\}$$

این دستین که نظری را از مقادیر  $N$  features (label  $\{\pm 1\}$ ) که داریم  $N$  dim را در کاری را بعنوان خروجی معرفی شرطی

Training Dataset

$$(x_1, y_1), \dots, (x_l, y_l) \in \mathbb{R}^N \times \{\pm 1\}$$

↓  
خوب داده

Testing Dataset

$$(\bar{x}_1, \bar{y}_1), \dots, (\bar{x}_l, \bar{y}_l) \in \mathbb{R}^N \times \{\pm 1\}, \text{ satisfying } \{\bar{x}_1, \dots, \bar{x}_l\} \cap \{x_1, \dots, x_l\} = \emptyset$$

محض نظری اثباتی هم را در Train و Test داده و  $\bar{y}_1, \dots, \bar{y}_l$  را هم عن داشت

با این فرضیت حس توامی function empirical risk عی طبقه بندی کنند مثل مسندی مثل خطوط فرط دو... در اطراف پریم کردنی به عنوان محض نظری

برای سبزیابیتر اینم دادن  $C$  ایمclassification tuning parameter

را جذب انتخاب کنیم. چیزی دوستی سه بعدی orientation صفت را تاکه می دهد



دروافعه رک بردار سرفال عوردي صفحه دیگر نھما را که قبل باقاطع صفحه پايس از جو رک ۳ کاره متصدي  
3 بودي حسنه هى تو اند به عنوان پاراصر رک هى آنها صفحه در تظرير رخته سود.  
(چون هى خواهيم پاراشر در تظرير رک اشتباه علی هى كنم دكتره ۳ نصفه هم در مصري ۳ بلکه هى )  
اين همه هى orientation صفحه هست

\* Estimate  $f(\vec{x}_i, \alpha)$  from a finite set of observations by minimizing some kind of an error function, for example:

## empirical risk

$$R_{\text{emp}}(\alpha) = \frac{1}{2l} \sum_{i=1}^l |y_i - f(\vec{x}_i, \alpha)|$$

$\alpha$ : parameters of learning engine.  $\rightarrow w, w_s$

class labels:  $y_i = \begin{cases} +1 & \text{if } \vec{x}_i \in \omega_1 \\ -1 & \text{if } \vec{x}_i \in \omega_2 \end{cases}$

$$y_i = \begin{cases} +1 & \text{if } \vec{x}_i \in \omega_1 \\ -1 & \text{if } \vec{x}_i \in \omega_2 \end{cases}$$

- اعلن اوقات بحسب  $\lambda$  کردن خط احتمال. حال آنکه  $\lambda$  روی داده‌گیری - empirical risk

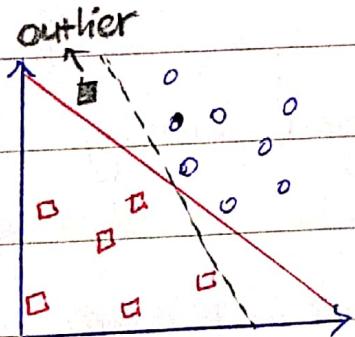
- هست را کنم کنیم، آیا ازدواج منجز به درست generalization خوب برای داده‌گیری train نماید یا نه؟

اما اس خانہ ایک تو نہ سنت و دو قسم اسی خانہ را روی دادھکر train حصی کشم کشم ، دل روی -  
دادھکر over hitting حصی fit حصی train رخ خی دھدر.

اپنے ازدواجی صورتِ حاصل برسوں اس سے:

۱) از محان خطا با صفت روشی داریم که  $\text{empirical risk}$ ,  $\text{min}_{\theta} \text{train loss}$

- خطاها عنوان ER classifer یا classifier یا انداخت  
- این برای استفاده کنید، ممکن است در جایی نباشد که ممکن باشد.



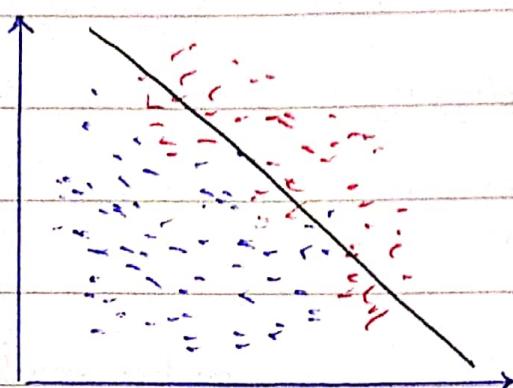
خوبی در اینجا داریم.  
خوبی در اینجا داریم.  
خوبی در اینجا داریم.

اگر برای داده‌ی  $x$   $ER = 0$  شود این  
خط را یک کلاسifier می‌نامیم.  
هر کدام را بعنوان classifier خوب درنظر نمیریم.  
این نشان حی دهد که  $min\{ER\}$  نزدیک‌ترین  
خطی است که می‌تواند.

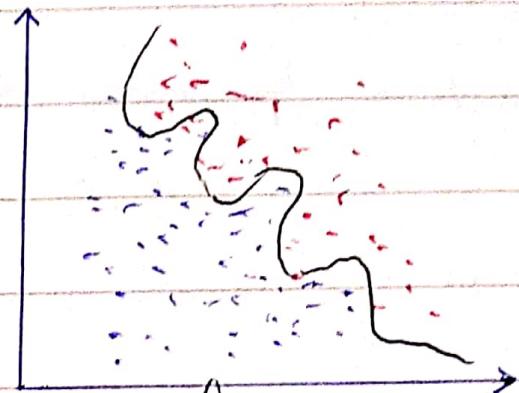
2nd QLR Test error,  $ER=0$  so global classifier wins  $f^*$

## capacity & VC dimension

از قدرت درستی را من می‌توانم بازی کنم Capacity



low capacity



high capacity

همه چیز را می‌دانم خوب است اما نیز خوب نیست

- classification error و high capacity تو،  
ایجاد می‌کند overfitting

- Zuhar Vapnik-Chervonenkis statistical learning (ریاضی)
- True error را می‌دانم و این را با نسبت به VC dimension
- با این روش می‌توانم generalization error test error را پیش بینی کنم

\* VC dimension can predict a probabilistic upper bound on the test error, of a classification model.

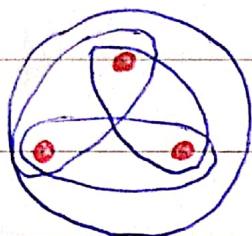
:  $\sqrt{\frac{1}{n} \sum_{i=1}^n \text{Pr}[Y_i \neq f(X_i)]}$   
to minimize, ER کو کوئی (1)

این low VC dimension چه چیزی (2)



## VC-dimension - Def(1)

Set Shattering → a subset  $S$  of instances of a set  $X$  is shattered by a collection of functions  $F$  if  $\forall S \subseteq S$  there is a function  $f \in F$  such that:



$$f(x) = \begin{cases} 1 & x \in S' \\ 0 & x \in S - S' \end{cases} \rightarrow$$

یعنی از اینها پیدا شود که -

اگر عضوی هستند در مکان ۱  
قرار چون نداشته باشند

شترین شکسته شدن از تبعیق موقوفه های بسیار زیگزگی ای از جمیع را shattering

پر shatter کردن یعنی در وکتور متریک دادن (اینجا) و عبارت بودا خواهش دید  
نحویت binary classification، VC dimension، حریط اجرا شوند.

## VC-dimension - Def (2)

The VC-dimension of a function set  $F$  ( $\text{VC-dim}(F)$ ) is the cardinality of the largest dataset that can be shattered by  $F$ .

ملحق debt، معدن اسک زریخویه کی زیادی را سیداً کننم تا بتوان با آن خواهد از تابعیت شد. Shatter ازین عالم آن زریخویه که آن زریخویه ای که بترین مقادیر غصه را در سُناسی کننم، آن جوچه اس زریخویه حی شود آن خواهد از تابعیت شد. VC-dimension

## VC dimension examples - Rectangles

فناوری اطلاعات  $\rightarrow$  عالم مستطیل‌گشایی ممکن درستی دو بعدی است اما اطلاعات دو بعدی نیست

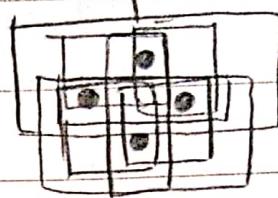
اعمق ععودی است، بعضی مواردی  $x, y$  صفاتی.

The VC dimension is 4, why?  $\rightarrow$  باید مطلب چهارم شکسته شود.

ععودی اعمق حی تو این آنرا شکسته کنم.

No set of 5 points

can be shattered.



بعضی کاری سخت که تمام حالات.

کامپیوئر داده‌گذاری نمایند.

هر چهارمی دلیلی باشی 4 تا اصنف کنم، مثلاً کسی که اصناف اعمق ععودی را نمی‌تواند خود را شکسته کردن نقاط در داده‌گذاری نمایند.



\* مثل دوم: نشان دهد نه بگای جویی که طول و عرض افقی (ورودی، اسیز) 3، VC dim.

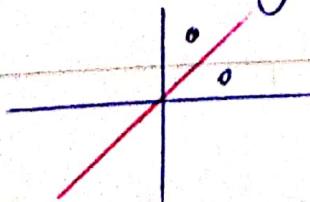
\* مثل سوم: مثلاً چهاری با طول و عرض بین 0 و 1 (اون، عودی و ...)

\* مثل چهارم: چهاری خواهد شد که می‌توانیم این معنی بیندیشیم؟

\* Machine  $f$  can shatter a set of points  $x_1, x_2, \dots, x_l$  if and only if  $\rightarrow$  for every possible training set of the form:

$(x_1, y_1), \dots, (x_l, y_l) \rightarrow$  there exists some value of  $\alpha$  that for which the  $f(\vec{x}_i, \alpha)$  gets zero training error.

\* Can the following  $f$  shatter the following points:



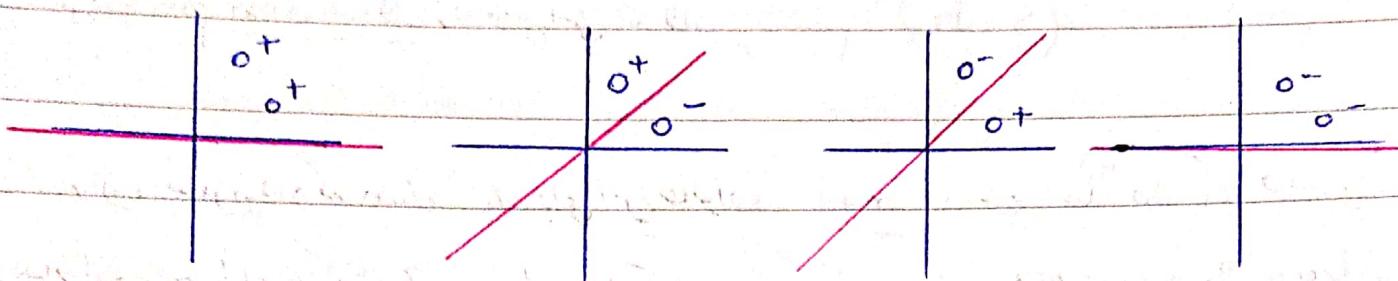
$$f(x, w) = \text{sign}(x \cdot w)$$

$$+1 \leftarrow +\text{لکسی} \leftarrow \text{sign } g^+$$

$$-1 \leftarrow - \text{لکسی} \leftarrow \text{sign } g^-$$



There are 4 training sets to consider.



$$w = (0, 1)$$

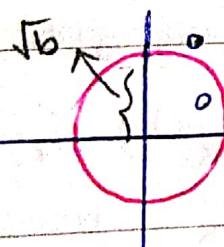
$$w = (-2, 3)$$

$$w = (2, -3)$$

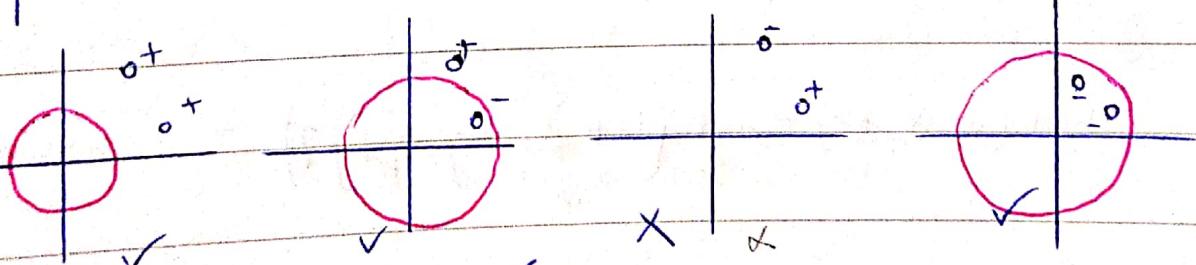
$$w = (0, -1)$$

برای هر یکی از مجموعه‌های آموزشی  $\{x_i, y_i\}_{i=1}^n$  می‌تواند یکی از خطوط را شکست کند. این مجموعه‌ها را shatter نمی‌شوند. مجموعه‌هایی که همچنان شکست نمی‌شوند دو نوع دارند: ۱) مجموعه‌هایی که همچنان شکست نمی‌شوند و همچنان شکست کنند. ۲) مجموعه‌ای که همچنان شکست کنند و همچنان شکست نمی‌شوند.

\* Can the following  $f$  shatter the following points:



$$f(x, b) = \text{sign}(x_1 x_2 - b) = \text{sign}(x_1^2 + x_2^2 - b)$$



آنرا درست نمی‌شود و آنرا shatter نمی‌شود. این مجموعه را sign می‌گویند.

VC dim for LC is linear classifier. (SVM, ...)

فرضیه  $\text{sign}(w \cdot x + b)$  یک دسته دو تایی باشد و  $m$ -dim خواهد بود. این خواهد بود که  $h$  چگونه باشد؟

حیوان شان داد که  $\dim_{VC}$  برابر این خانواده  $m+1$  است مگر در حصای دلخیری -  
خانواده‌ی خفهای حدکسر ۳ نقطه را هم توایند shatter نمی‌شود در حصای ۳ دلخیری ، حدکسر ۴ نقطه و  
الی آخر.

proof:

براهمنی کو اس سے کوئی نظر نہیں رکھا جاوے۔

$$h \geq m$$

$$x_1 = (1, 0, 0 \dots, 0)$$

$$x_2 = (0, 1, 0, \dots, 0)$$

$$x_m = (0, 0, 0, \dots, 0)$$

مavarid حی تو دنخ  $2^m$  (متعدد ایکی) combinations of class labels;

به ازای هر کدام از این labeling چن توائم معتبر بردار انتقال ایجاد شود و در این مورد راهنمایی محسن

→ label k میں  $\text{sign}(w_i x_i + b)$  نے کیا

$$\sum_{k=1}^n w_k = y_k \rightarrow b = 0$$

$$\Downarrow \text{sign}(w \cdot x_k + b) = \text{sign}\left(b + \sum_{j=1}^m w_j \cdot x_k[j]\right)$$

$$\begin{aligned} & \text{- proof } h \geq m+1 \\ & \text{- proof } h \leq m+2 \end{aligned} \left. \right\} \rightarrow h \geq m$$

## VC dimension & margin of separation

فرصه کيسي بحد ادني تقطع داريم و حجم خواهش ما شين باس اينجا خود داشم و هرچه  
نمایه شدن را باشيم (باشند) shattering، اما حجم خواهش بجزئی کنتم اين کار  
خوب است یا نه؟

min  $\frac{1}{2} \|\alpha\|^2$  در محدوده margin را maximize کرد Vapnik

- در محدوده hyper-plane (خط) باشند کار کردن در محدوده

حجم خواهش اينکه

\* The optimal hyperplane is the one giving the largest margin of separation between the classes.

\* A bound on the Generalization performance of learning Machine

- Expected Risk:  $R(f(\vec{x}, \alpha)) = \int \frac{1}{2} |y - f(\vec{x}, \alpha)| dP(\vec{x}, y)$

- Empirical Risk:

$$R_{\text{emp}}(f(\vec{x}, \alpha)) = \frac{1}{2l} \sum_{i=1}^l |y_i - f(\vec{x}_i, \alpha)|$$

$$\underbrace{R(f(\vec{x}, \alpha))}_{\text{true error}} \leq \underbrace{R_{\text{emp}}(f)}_{\text{train error}} + \sqrt{\frac{h(\log(2l/h) + 1) - \log(\eta/14)}{l}}$$

برآورده  $1 - \eta \sqrt{2l/h}$  است که این فرمولی است که نسبت به  $1 - \eta$  کمتر است  
که  $1 - \eta$  است که می خواهیم این را بروز کرد

  $\eta = 0.05$ . این نسبتی را داریم که  $1 - \eta$  بود.  $1 - \eta$  کمتر است که  $1 - \eta^2$

طبق فرمول صفحه ۷۱، True error خطای است که از داده‌گی ندیده (بست) مرسی.  
 حس آندر: طبق معقول هم انتظار داریم که خطای داده‌گی سک از Train بیشتر باشد، اما این خوب است آگر بتوانیم نک حدیاب برای خطای داده‌گی ندیده روی چالش تجربه نماییم، که این -  
 سه بعده است generalization آن ربطاً خواهد بود.  
 هرچه این حدیاب عدد کوچک‌تر باشد بتر است. در مقاله Vapnik سان راده این حدیاب  
 ساختراز دوفاکتور است:

confidence term (۲)

Empirical (۱)

از پایه این تعبیر نماینند و وجود دارد confidence term

که این چنان  $\leftarrow h$  (۱) VC-dimension خواهد بود.

که تعداد داده‌گی  $\leftarrow L$  (۲) train

$\leftarrow (1-\eta)$  (۳) (این) نتایجی است که بحاجی گوید این نامساوی باحتمال

برقرار خواهد بود.

غیراندیش  $h$ ،  $L$ ،  $\eta$  چه ترتیبی دارند؟

به صورت آنودی حس درینم که هرچه تعداد داده‌گی train (L) بیشتر باشد، بتر است.  
 حس توآینم شان بدشم که وقتی  $L \rightarrow \infty$  حس را در دو این -

خوب است.

اگر ها زیاد شود، بعین داریم ضریب محقق را با احتمال برآورد و از خطای درجی کوچک (دست آزادی را با احتمال خوب) در اساس فرمول سه‌جنبه‌ای کوچک که مقدار confidence term با احراش ها زیاد نمی‌شود.

بعین داریم حدسایی خطای True error، احراش‌ها حقیقی داشن اصلان مطابق باشند.

$$\frac{(h(\log 2\ell/h)+1) - \log(7/14)}{2}$$

confidence term

بعین خاطر نویسی می‌شود، اما جایی که عده است کوچک باشد و از محقق که یا خانواده‌های انسانی استفاده کنیم که ظرفیت بسیاری داشته باشند و اصولاً در حیاتی دلخواه و علوم discriminator هم اصل parsimony را داریم که آن هم به عنین است و محقق کند، در واقع حقیقتی کوچک است.

حق خواهد داشت که این توانی که حقیقی توانی دارد را درست نمودن عمل خوب است.

خطیب‌نامه SVM برچی کردم که دو طبقه داریم و حق خواهیم آورنا را از هم جدا کنیم و خلاصه نمایم که خطای این کار خوب است پس به نسبال حدکاسته حقیقی کردم و ۱۰٪ حق خواهیم داشت که همان خطا را با صفتی که حق تواند دو طبقه را بالغه با خطای صفر (Training error) جدا کند آن را استفاده کنیم که margin بزرگ تولید کند.

SVM در فرم اولیه برای جدا کردن کلاس‌های دوگانه (binary classification) استفاده می‌شود اما خواهیم دید که حق تواند SVM را بهتر از دو کلاس هم قسم دارد.

Train کردن یک SVM معادل حل یک quadratic programming prob با سوابع.

اسه که حق تواند با خودی ساده و نصادر خوش را تواند linear constraints

## SVM overview

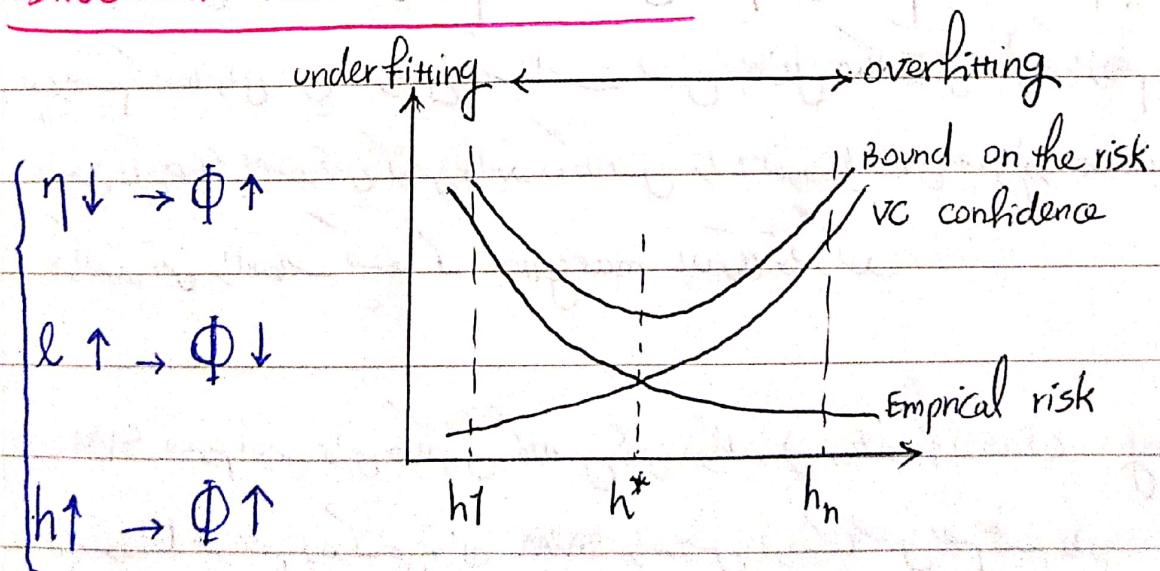
- SVMs perform structural risk minimization to achieve good generalization performance.

- The optimization criterion is the margin or separation between classes.

- Training is equivalent to solving a quadratic programming problem with linear constraints.

- Primarily two-class classifiers but can be extended to multiple classes.

## Structural risk minimization



$$R(\alpha) \leq R_{\text{emp}}(\alpha) + \sqrt{\frac{(h(\log 2l/h) + 1) - \log(1/4)}{l}}$$

- اگر جزو دھم با اختیال ہے برائیم کے درباری خطا صبی، تہذیبی رسم خون گست اسی اسی  
 ۱-۷  $\rightarrow$  probability  $\leftarrow (\eta \rightarrow \phi) \rightarrow (\text{خون} \rightarrow \text{کسر اسی})$ .

## Two approaches

goal: To find a trained machine in the series whose sum of empirical risk and VC confidence is minimal.

→ Neural Network:

- Fix the VC confidence and minimize the empirical risk. May be find High capacity mappings.  
(very non-linear  $\rightarrow$  overfitting)

→ Support vector machine :

- Fix the empirical risk and minimize the VC confidence.  
→ Equivalent to maximizing the margin.

پس اینجا هدف سیگردن یک آبرصدود و مسایی feature space که بتواند نتایج training صور مخلوط را جذب کند.

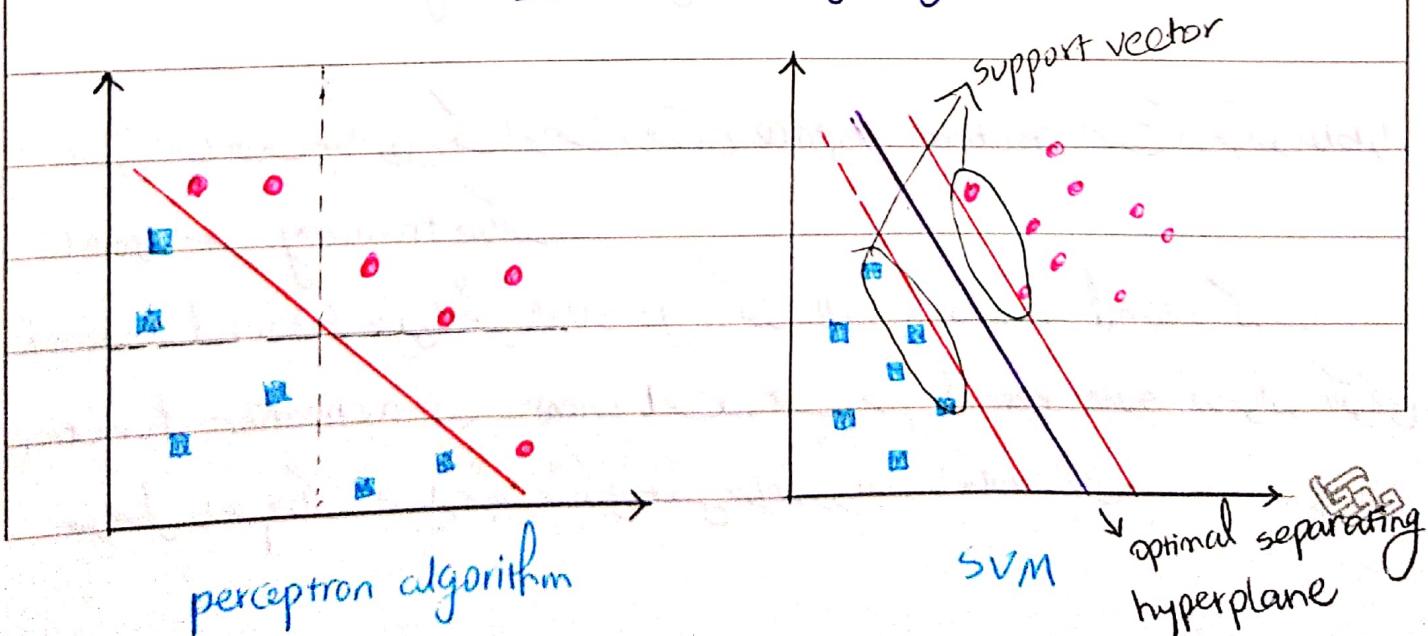
بررسی SVM، برای کمینه کردن هسته میان فضای margin و برای مکث کردن آن از نظر میان فضای confidence term کم استفاده کنیم.

## the two class problem

several decision boundaries can separate these two classes.

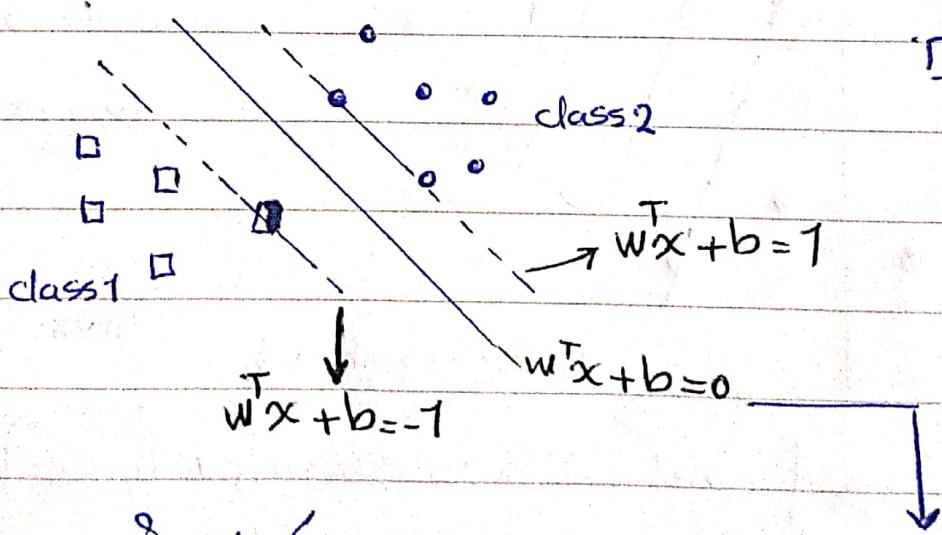
perceptron algorithm learns any separating hyperplane.

- SVM learns the best separating hyperplane.



بايد دنبال حالت  $\hat{y} = \text{sum}$  که در خطا margin داشت از آن خطای کمترین آنرا عبور حس کند.  
به عنوان classifier را در آنست.

از راهنمای معرفتی بجز حرف کنند و صفتی حواری support vector بیان می‌کند. حالا روشی به اینکه label، نسبت دسقیلی است و توانیم مکارله‌ی دو صفت را بصورت زیر در نظر بگیریم.



بازی هر چیزی × مادر آن را این صفتیک خواهد سد.

w & b are unknown.

$$w^T x + b = w_1 x_1 + w_2 x_2 + \dots + w_N x_N + b$$

حی توان تاں دارک خاصیتی دو صورتی موازی (m) گرا بر اس با :

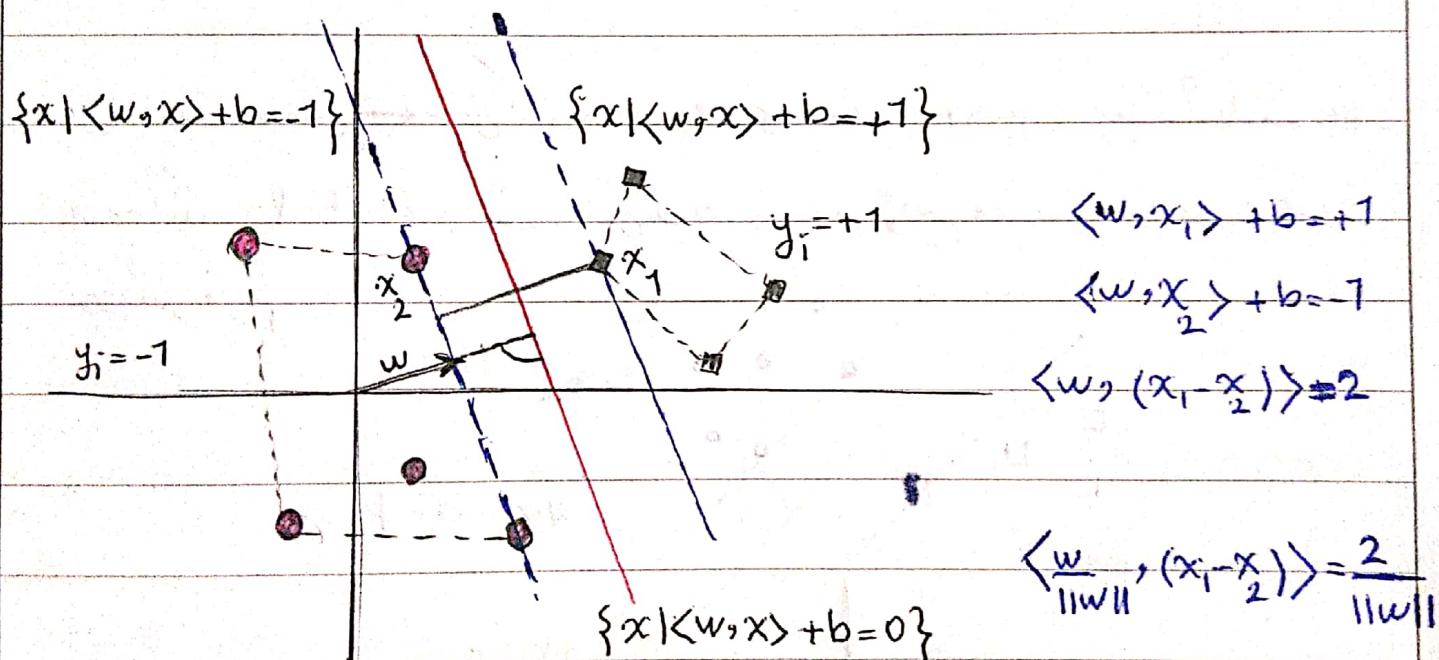
$$m = \frac{2}{||w||} \rightarrow \text{نرم} \frac{2}{||w||} \rightarrow \text{نرم} \frac{2}{||w||}$$

قادر به  $w$  و  $b$  حاصل کردن که وقتی  $m$  را حساب کنیم از همانجا دستور داده شد.

\*We should maximize the margin ( $m$ )

(Convex optimization)  $\Leftrightarrow$  optimization

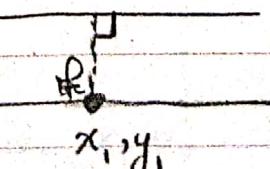
## Decision Boundary



HW:  $\rightarrow$  خروجی خاصیتی دو خط معانی را ثابت کند.

$$1) ax + by + c = 0 \leftarrow$$

$$2) ax + by + c' = 0 \leftarrow$$



$$\frac{|c - c'|}{\sqrt{a^2 + b^2}}$$

$$\frac{|c - c'|}{\sqrt{a^2 + b^2}}$$

minimize,  $C\|w\|$   $\Leftrightarrow$  minimize  $\frac{2}{\|w\|}$   $\Leftrightarrow$  maximize  $\frac{1}{\|w\|}$

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

$$\leftarrow C\|w\| \Leftrightarrow \|w\|^2, \|w\| \Leftrightarrow$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1, i=1, \dots, l$$

(l: number of training samples)

↑  
constraint

\* The above is an optimization problem with a convex quadratic objective and only linear constraints. Its solution gives us the optimal margin classifier. This optimization problem can be solved using commercial quadratic programming (QP) code.

\* Lagrange duality will lead us to this optimization problem's dual form, which will play a key role in allowing us to use kernels to get optimal margin classifiers to work efficiently in very high dimensional spaces.

The dual form will also allow us to derive an efficient algorithm for solving the above optimization problem that will typically do much better than generic QP software.

- \* درای نسبتی outlier را همچنان خلائق وجود دارد، مناسخ است می -
- \* درای نسبتی  $\mu$  کوئی underlying distribution نهان نموده باشیم، حالا آنر تقطیعی (استدای) باشیم که در آن outlier خواهد، حتی کوئی distribution نهاد.
- \* راه دستگیر اس آنر تقطیعی ب طرز معنی دار، آمارو<sup>ج</sup> (statistics) توزیع را جایگزین، پس
- \* مناسخ تقطیعی ب طرز معنی داری عبارتی، var، ... را جایگزین outlier است.

من می‌خواهم برای اینجا حالتی را در لگراژیان (lagrangian) optimization از استفاده از مکالمه (conjugate gradient) برای حل کنید. این مکالمه را می‌توانم با حل مسئله دوگانه (dual) و مسئله اصلی (primal) مرتبط داشتم.

\* The solution to the dual problem provides a lower bound to the solution of primal (minimization) problem.

\* In primal dual solution, minimization variable values are chosen such that they satisfy the constraints of the dual problem. This is called dual solution.

\* In dual solution, primal solution can be obtained by substituting the dual values in the primal objective function.

\* However in general the optimal values of the primal and dual problems need not be equal.

↳ their difference called the duality gap

رسالی duality gap و convex optimization کے مطابق اگر حجوم ختم کرنے والے KKT

الآن نعود إلى المقدمة فيsvm، حيث هي primal problem، بينما المقابل هو dual problem.

- The primal (against dual) optimization problem:  
( $w$  is primal variable)

$$\begin{aligned} & \min_w f(w) \\ & \text{s.t. } g_i(w) \leq 0, i=1, \dots, K \\ & h_i(w) = 0, i=1, \dots, t \end{aligned}$$

- The lagrangian dual problem is defined as:

$$\begin{aligned} \max L(w, \alpha, \beta) &= f(w) + \sum_{i=1}^K \alpha_i g_i(w) + \sum_{i=1}^t \beta_i h_i(w) \\ \text{s.t. } \alpha_i &\geq 0, i=1, \dots, K \end{aligned}$$

The  $\alpha_i$ 's and  $\beta_i$ 's are the lagrangian multipliers or dual variables.

consider the quality  $\Theta_P(w) = \max_{\substack{\alpha, \beta: \alpha_i \geq 0}} L(w, \alpha, \beta)$

لأن  $\lambda^*$  هي حلٌّ دُوالٌ، فالحلٌّ البدائي  $\bar{x} \leftarrow d^* = P^*$  صحيح.

kernelized SVM که در فرآیند اصلی (primal) مسأله ایجاد نمایشگذاری می‌کند.

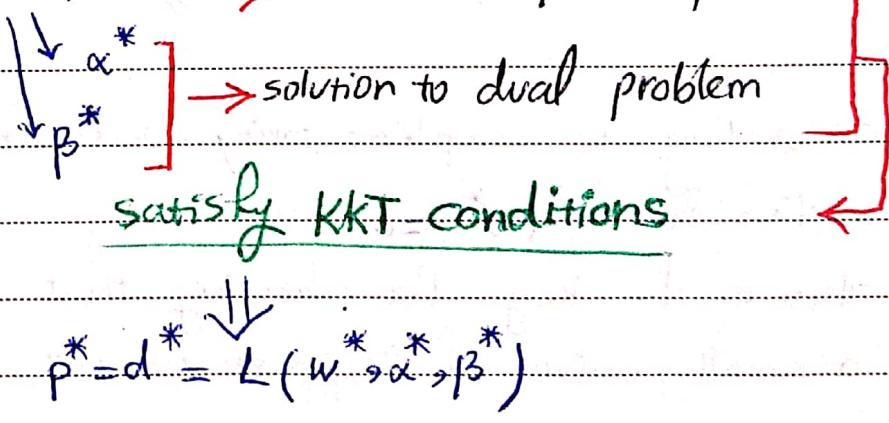
non-linear راچو ایت جای خوار از یعنی حسابات در رابطه Kernel trick برای بزرگی classification

## Conditions:

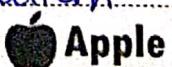
- Conditions:

  - + suppose  $f$  and  $g_i$ 's functions are convex and the  $h_i$ 's are affine
  - + and the constraints  $g_i$  are (strictly) feasible (there exists some  $w$  so that  $g_i(w) < 0$  for all  $i$ )

+ under above assumptions  $\rightarrow w^* \rightarrow$  solution to primal problem]



سیستم خارجی از نیم  
خطای بایک سیل alline حی تو این متعاضر که دریک کره پرالنر سیستم اندروید واصل مشخص از هم  
دارند را در بازه بیشترهای دیگری بیاوریم که از کردی اول کوچکتر بازیگر کسر اس و سنت خاصه شوند و  
که هم از نیم pattern



## Karush-Kuhn-Tucker (KKT) conditions

theorem:

suppose the problem :

$$\min_w f(w)$$

$$\text{s.t } g_i(w) \leq 0, i=1, \dots, k$$

$$h_i(w) = 0, i=1, \dots, t$$

has a local (max) min at  $w=w^*$ , and that a constraint qualification (to be specified) is satisfied at  $w^*$ . Then there are  $\alpha_1^*, \alpha_2^*, \dots, \alpha_K^*$  such that:

$$\alpha_i^* g_i(w^*) = 0, i=1, \dots, k$$

$$g_i(w^*) \leq 0, i=1, \dots, k$$

$$\alpha_i^* > 0, i=1, \dots, k$$

KKT

complementary condition.

according to complementary condition, if  $\alpha_i^* > 0$ , then it implies that  $g_i(w^*) = 0$ .

this is the key for showing that the SVM has only a small number of "support vectors".

-  $\alpha_i = 0$ , Train data b[er] w[ith] c[on]s[traint]  $\alpha_i > 0$ , support vector, b[ut] \*

-  $\alpha_i > 0$ , support vector,  $\rightarrow$  ای داده

اگر داده داری داشتی ممکن است  $\alpha_i > 0$  باشد، این داده را داده داری می‌نامیم.

•  $\alpha_i > 0$ , Train b[er] w[ith] c[on]s[traint]

The KKT conditions are necessary for a local max (min). They don't guarantee that a point satisfying them is actually a local max (min).

other KKT conditions:

$$\frac{\partial}{\partial w_i} L(w^*, \alpha^*, \beta^*) = 0, i=1, \dots, n \rightarrow \text{feature space dimensionality}$$

$$\frac{\partial}{\partial \beta_i} L(w^*, \alpha^*, \beta^*) = 0, i=1, \dots, t$$

### Lagrangian duality (max SVM margin)

optimal problem:

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

$$\text{s.t. } y^{(i)} (w^T x^{(i)} + b) \geq 1, i=1, \dots, l$$

we can write the constraints as:

$$g_i(w) = -y^{(i)} (w^T x^{(i)} + b) + 1 \leq 0$$

- \* From the KKT dual complementarity condition, we will have  $\alpha_i > 0$  for the training examples that have functional margin exactly equal to one.



$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^m \alpha_i [y^{(i)} (w^T x^{(i)} + b) - 1]$$



$$\nabla_w L(w, b, \alpha) = w - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} = 0$$



$$w = \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \rightarrow \frac{\partial}{\partial b} L(w, b, \alpha) = \sum_{i=1}^m \alpha_i y^{(i)} = 0$$

$$L(w, \alpha, b) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} - b \sum_{i=1}^m \alpha_i y^{(i)}$$

inner product of  $x^{(i)}$  &  $x^{(j)}$ 

zero

$$\alpha_i > 0, \quad \frac{\partial}{\partial b} L(w, b, \alpha) = \sum_{i=1}^m \alpha_i y^{(i)} = 0$$

Therefore, we obtain the following dual opt problem:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)}, x^{(j)})$$

$$\text{s.t } \alpha_i > 0, \quad i = 1, \dots, m$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

~~$w = \sum_{i=1}^l \alpha_i y^{(i)} x^{(i)}$~~



Year: ..... Month: ..... Day: ..... ( )

Subject: .....

To find the optimal  $w$ 's as a function of  $\alpha$ 's. Having found  $w^*$ , by considering the primal problem, it is also straightforward to find the optimal value for term  $b$  as:

$$b^* = \frac{\max_{i:y^{(i)}=1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=-1} w^{*T} x^{(i)}}{2}$$