- Support Vector Machines (SVMs).

Maria-Florina Balcan

03/25/2015

Support Vector Machines (SVMs).

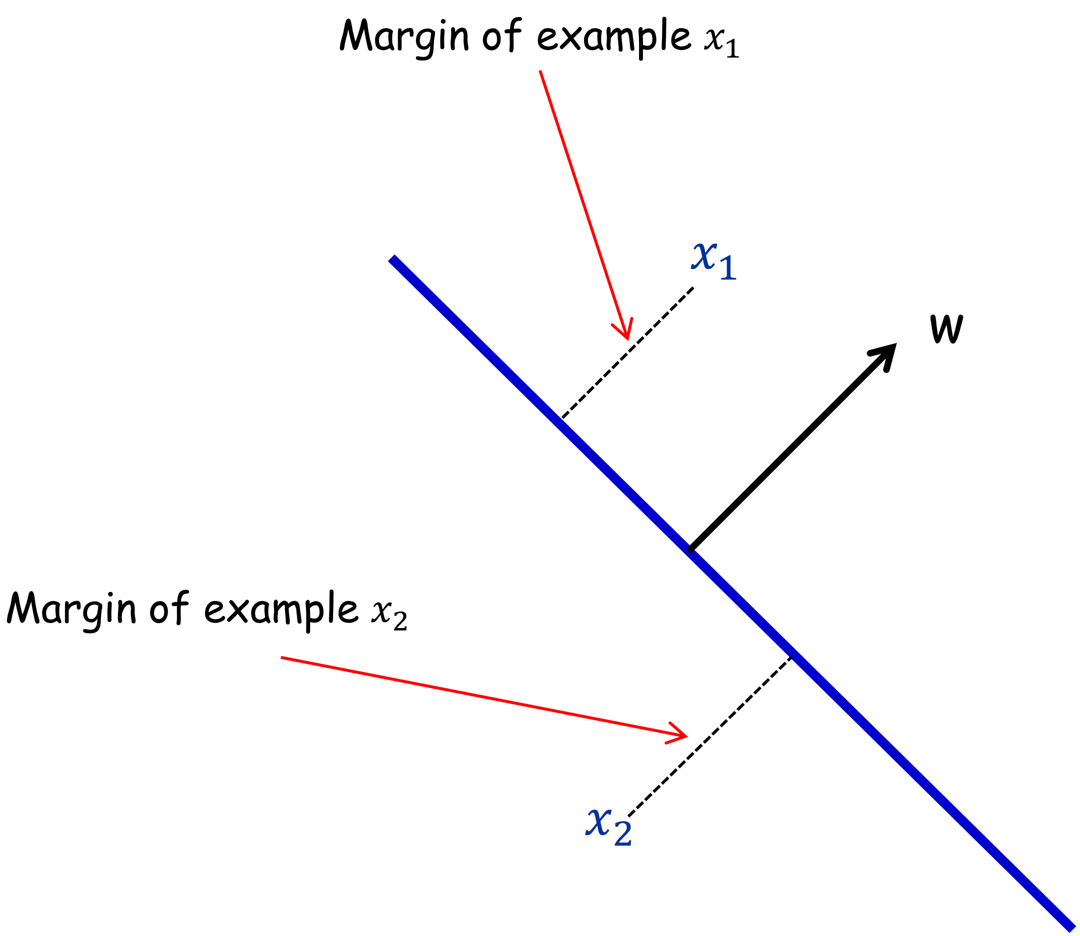
One of the most theoretically well motivated and practically most effective classification algorithms in machine learning.

Directly motivated by Margins and Kernels!

Geometric Margin

WLOG homogeneous linear separators $[w_0 = 0]$.

Definition: The margin of example x w.r.t. a linear sep. w is the distance from x to the plane $w \cdot x = 0$.



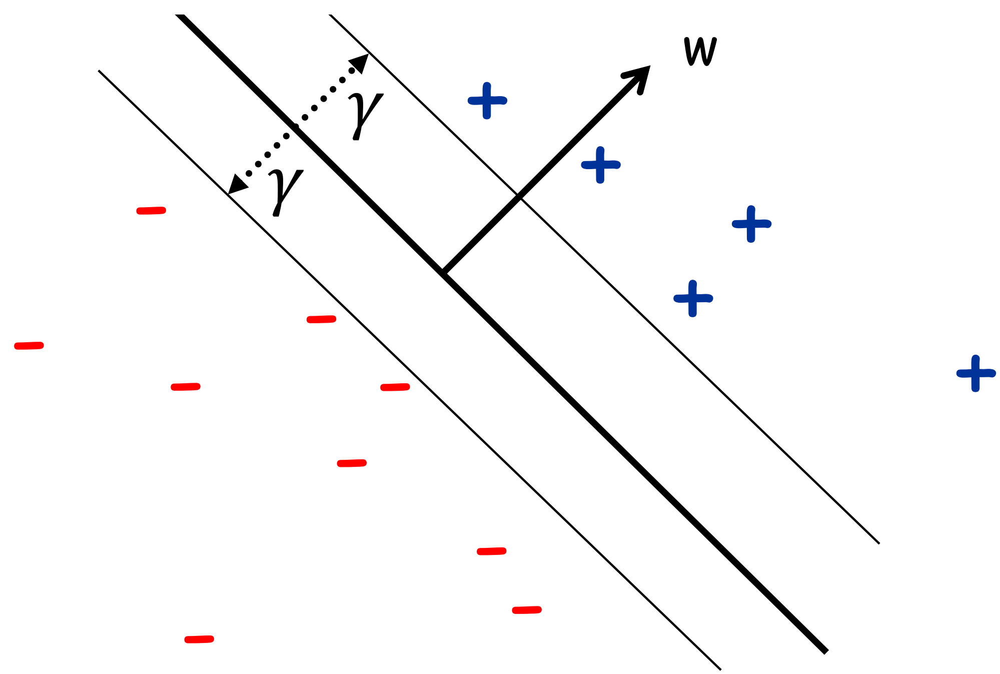
If $||w|| = 1$, margin of x w.r.t. w is $|x \cdot w|$.

Geometric Margin

Definition: The margin of example x w.r.t. a linear sep. w is the distance from x to the plane $w \cdot x = 0$.

Definition: The margin γ_w of a set of examples S wrt a linear separator w is the smallest margin over points $x \in S$.

Definition: The margin γ of a set of examples S is the maximum γ_w over all linear separators w.



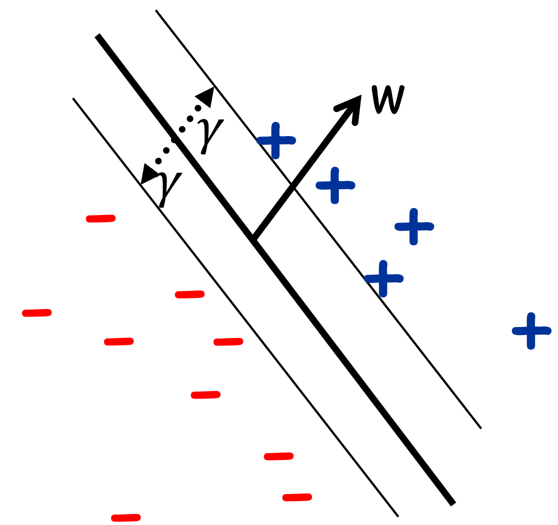
Margin Important Theme in ML

Both sample complexity and algorithmic implications.

Sample/Mistake Bound complexity:

- If large margin, # mistakes Peceptron makes is small (independent on the dim of the space)!

- If large margin γ and if alg. produces a large margin classifier, then amount of data needed depends only on R/γ [Bartlett & Shawe-Taylor '99].



Algorithmic Implications

Suggests searching for a large margin classifier… SVMs

Support Vector Machines (SVMs)

Directly optimize for the maximum margin separator: SVMs

First, assume we know a lower bound on the margin γ

假设已知γ

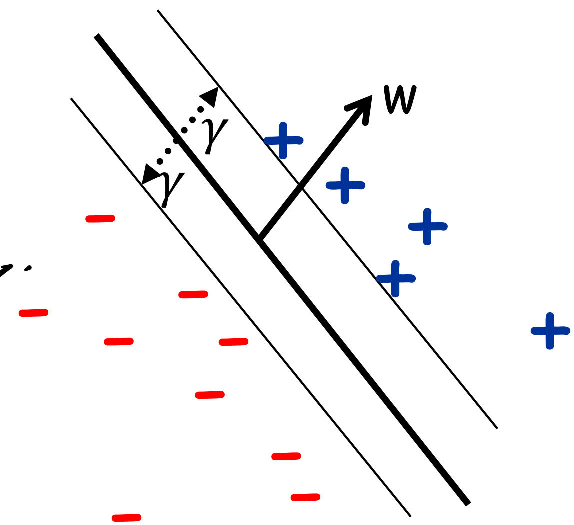
Input: γ, $S=\{(x_1, y_1), \ldots, (x_m, y_m)\}$; $\gamma_i = \frac{y_i w^T x_i}{\|w\|}$

Find: some w where:

用平方是为了便用优化.

- $\|w\|^2 = 1$
- For all i, $\underbrace{y_i w \cdot x_i}_{\text{置信度}} \geq \gamma$

Output: w, a separator of margin γ over S



Realizable case, where the data is linearly separable by margin γ

Support Vector Machines (SVMs)

Directly optimize for the maximum margin separator: SVMs

E.g., search for the best possible γ → 若γ未知

加上约束条件.

Input: $S=\{(x_1, y_1), \ldots, (x_m, y_m)\}$;

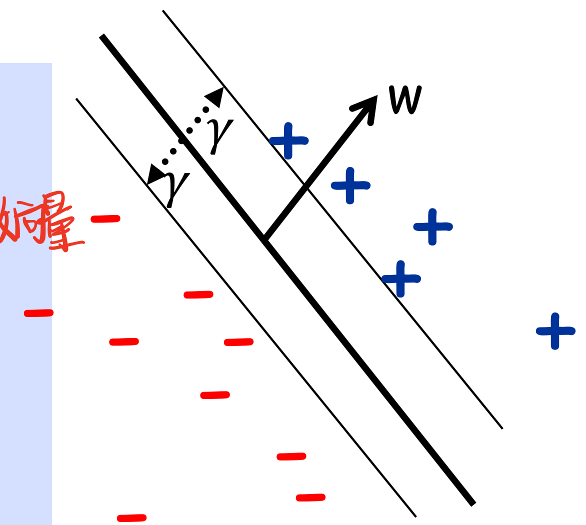
Find: some w and maximum γ where:

- $||w||^2 = 1$　　对于 $||w||=1$: non-convex. 所以加上平方 ⇒ 线性规化
- For all i, $y_i w \cdot x_i \geq \gamma$

$\min \underline{c}^T x$ 常数向量

s.t. $Ax=b$

Output: maximum margin separator over S



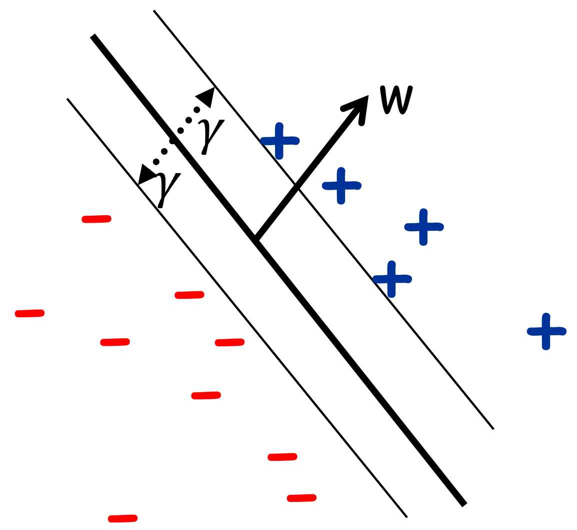
Support Vector Machines (SVMs)

Directly optimize for the maximum margin separator: SVMs

Input: S={(x_1, y_1), …,(x_m, y_m)};

Maximize γ under the constraint:

- $||w||^2 = 1$
- For all i, $y_i w \cdot x_i \geq \gamma$



Support Vector Machines (SVMs)

Directly optimize for the maximum margin separator: SVMs

Input: $S=\{(x_1, y_1), \ldots, (x_m, y_m)\}$;

Maximize γ under the constraint:

- $||w||^2 = 1$
- For all i, $y_i w \cdot x_i \geq \gamma$

objective function

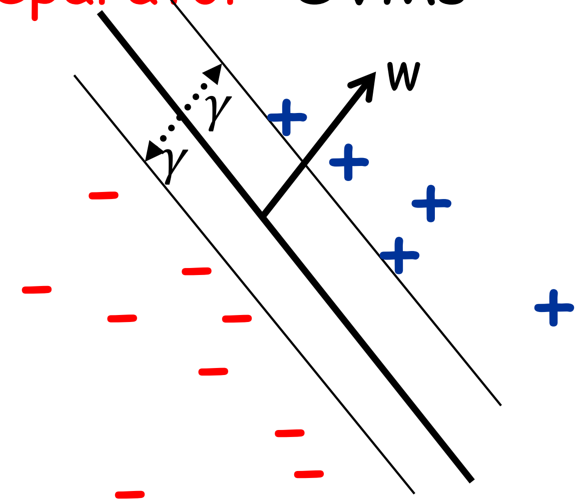
constraints

This is a **constrained optimization** problem.

- Famous example of constrained optimization: **linear programming**, where objective fn is linear, constraints are linear (in)equalities

Support Vector Machines (SVMs)

Directly optimize for the maximum margin separator: SVMs



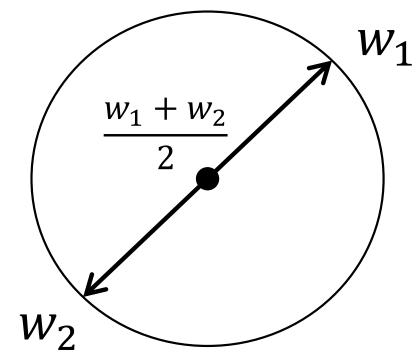
Input: S={(x_1, y_1), …,(x_m, y_m)};

Maximize γ under the constraint:

- $||w||^2 = 1$
- For all i, $y_i w \cdot x_i \geq \gamma$

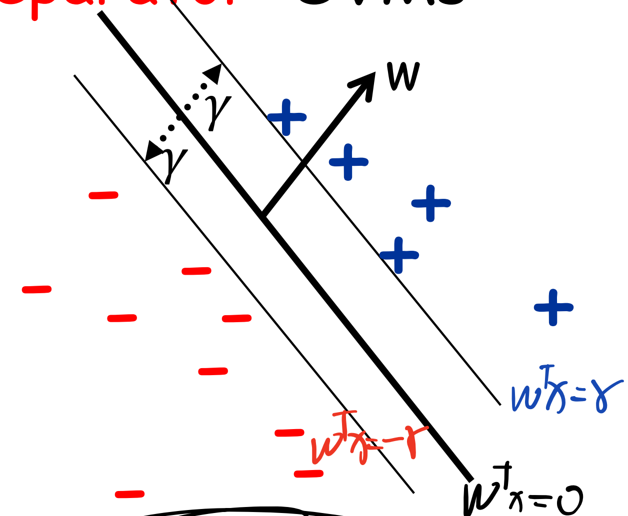
This constraint is non-linear.

In fact, it's even non-convex



Support Vector Machines (SVMs)

Directly optimize for the maximum margin separator: SVMs

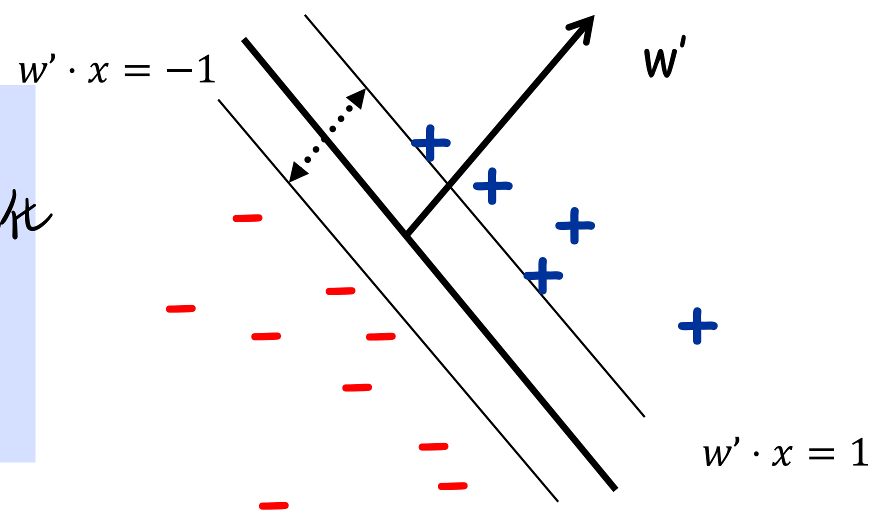


Input: S={(x_1, y_1), …,(x_m, y_m)};

Maximize γ under the constraint:

- $||w||^2 = 1$ / 令 $w' = \frac{w}{\gamma} \Rightarrow ||w'|| = \frac{||w||}{\gamma} = \frac{1}{\gamma}$
- For all i, $y_i w \cdot x_i \geq \gamma$

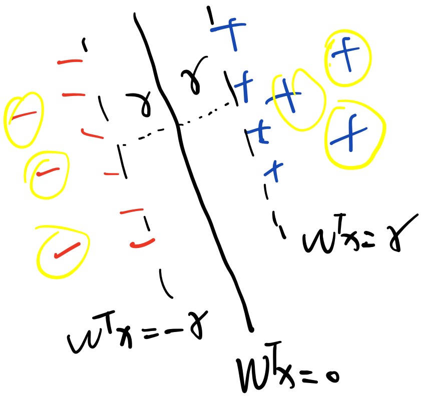
$w' = w/\gamma$, then max γ is equiv. to minimizing $||w'||^2$ (since $||w'||^2 = 1/\gamma^2$).
So, dividing both sides by γ and writing in terms of w' we get:



Input: S={(x_1, y_1), …,(x_m, y_m)};

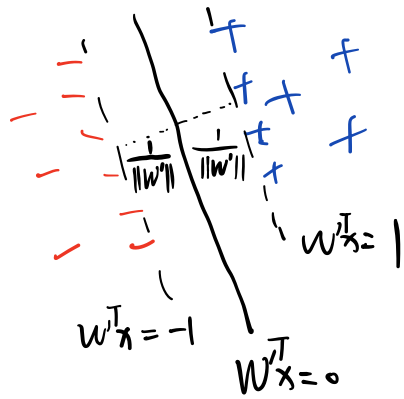
Minimize $||w'||^2$ → 最终是梯度求解,加耗可以优化 under the constraint:

- For all i, $y_i w' \cdot x_i \geq 1$



$w' = \frac{w}{\gamma}$

$\Rightarrow$



$\Rightarrow$ 对于任意点，需满足 $w'^Tx \geq 1$ 或 $w'^Tx \leq -1$ $\Rightarrow |w'^Tx| \geq 1$

实际上这些点，对边界、间隔的确定没有任何贡献.

靠近边界的点、称 support vectors.

Support Vector Machines (SVMs)

Directly optimize for the maximum margin separator: SVMs

Input: S={$(x_1, y_1), \ldots, (x_m, y_m)$};

$\text{argmin}_w \, ||w||^2$ s.t.:

- For all i, $y_i w \cdot x_i \geq 1$

This is a **constrained optimization** problem.

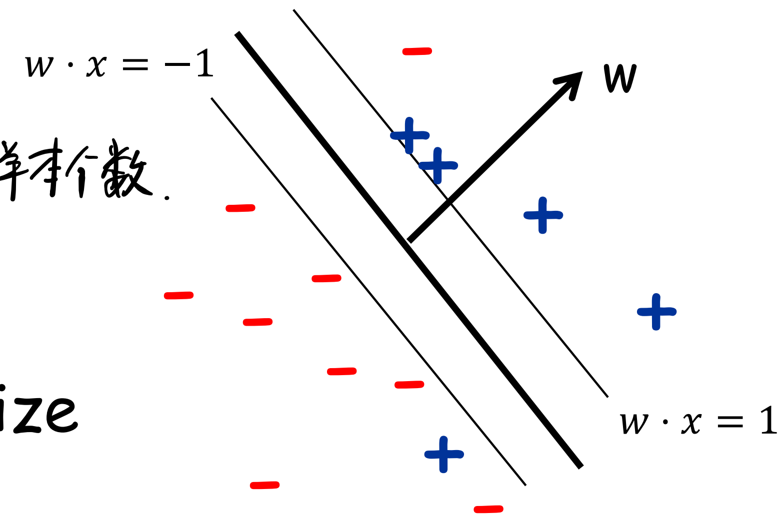
- The objective is convex (quadratic)
- All constraints are linear
- Can solve efficiently (in poly time) using standard **quadratic programing** (QP) software

二次规化：是可求解析解的．但不这样求(慢)，用对偶的解法．

Support Vector Machines (SVMs)

有噪声时.

Question: what if data isn't perfectly linearly separable?



Issue 1: now have two objectives

- maximize margin
- minimize # of misclassifications. 最小化分类错误的样本个数.

Ans 1: Let's optimize their sum: minimize

$$||w||^2 + C(\text{\# misclassifications})$$

where C is some tradeoff constant.

Issue 2: This is computationally hard (NP-hard). 

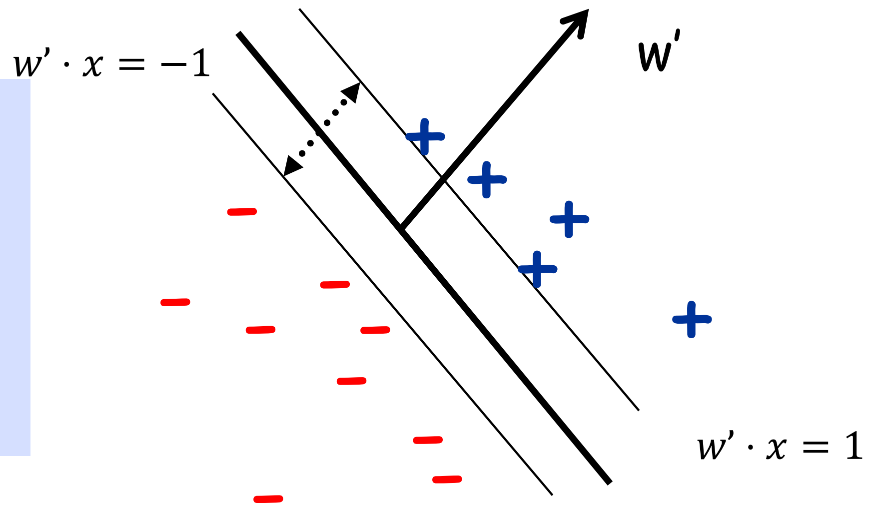
[even if didn't care about margin and minimized # mistakes]

NP-hard [Guruswami-Raghavendra'06]

Support Vector Machines (SVMs)

Question: what if data isn't perfectly linearly separable?

Replace "# mistakes" with upper bound called "hinge loss"



<u>Input</u>: S={(x_1, y_1), …,(x_m, y_m)};

Minimize $||w'||^2$ under the constraint:

- For all i, $y_i w' \cdot x_i \geq 1$

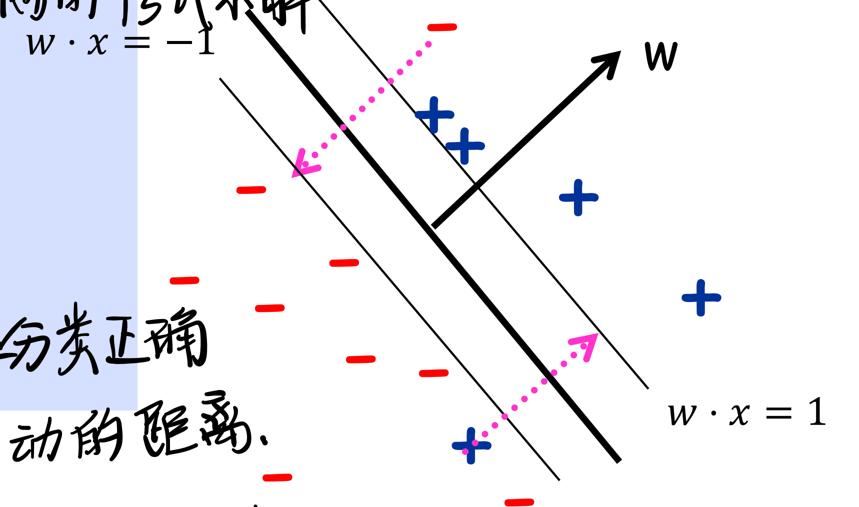
<u>Input</u>: S={(x_1, y_1), …,(x_m, y_m)}; 一般会用对偶的形式求解

<u>Find</u> $\text{argmin}_{w,\xi_1,\ldots,\xi_m} ||w||^2 + C \sum_i \xi_i$ s.t.:

- For all i, $y_i w \cdot x_i \geq 1 - \xi_i$

  $\xi_i \geq 0$ 表示使第i个样本分类正确

ξ_i are "slack variables" 所需要移动的距离.



使对每一个点都有 $y_i w x_i + \xi_i \geq 1$，只需在优化时加上 $C \Sigma \xi_i$ 保证距离不算太大。

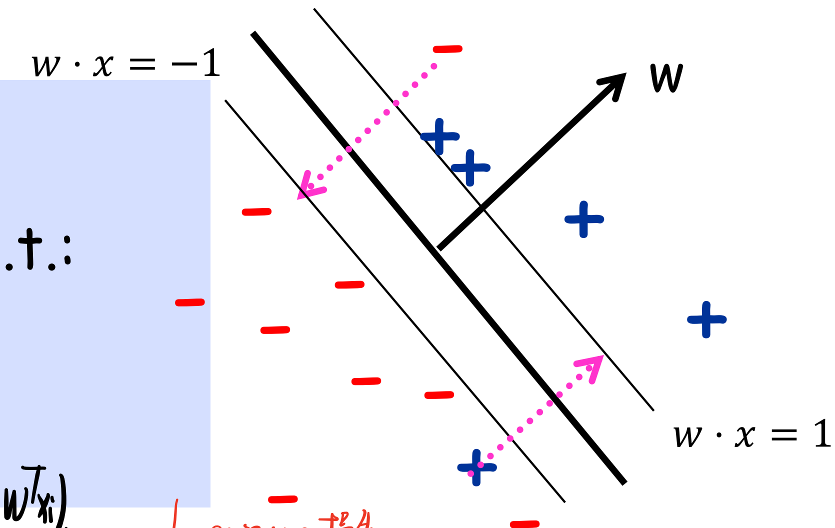
$\begin{cases} y_i w x_i \geq 1 \Rightarrow \xi_i = 0 \\ y_i w x_i < 1 \Rightarrow \xi_i > 0 \end{cases}$

Support Vector Machines (SVMs)

$\Rightarrow \xi_i = \max(0, 1 - y_i w^T x_i) = (1 - y_i w^T x_i)_+$ → 表示取大于0的部分.

Question: what if data isn't perfectly linearly separable?

Replace "# mistakes" with upper bound called "hinge loss"



<u>Input</u>: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$;

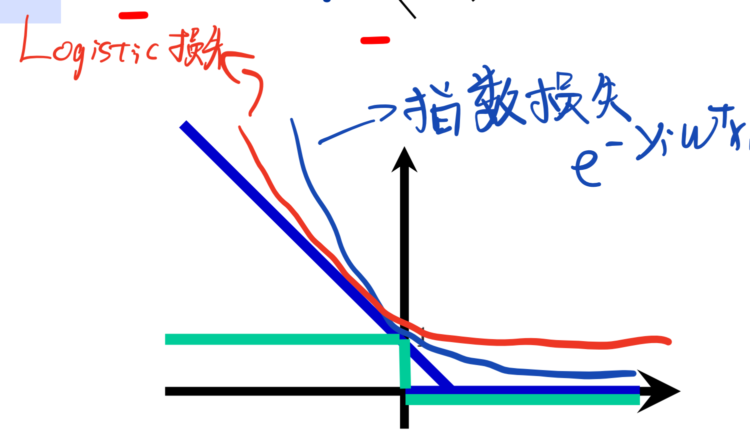
<u>Find</u> $\text{argmin}_{w, \xi_1, \ldots, \xi_m} ||w||^2 + C \sum_i \xi_i$ s.t.:

- For all i, $y_i w \cdot x_i \geq 1 - \xi_i$
  $\xi_i \geq 0$

$\Rightarrow$ 无约束的优化版本：$\underset{w}{\text{argmin}} \, ||w||^2 + C \sum (1 - y_i w^T x_i)_+$

ξ_i are "slack variables"

C controls the relative weighting between the twin goals of making the $||w||^2$ small (margin is large) and ensuring that most examples have functional margin ≥ 1.



$l(w, x, y) = \max(0, 1 - y\, w \cdot x)$

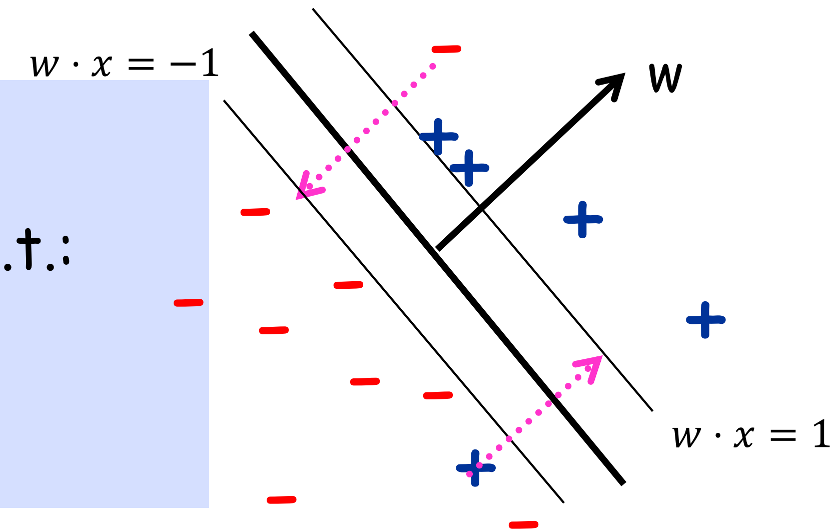
convex + C · convex ⇒ 是 convex.

将损失换成 Logistic Loss：$\text{argmin} \, \frac{1}{2}||w||^2 + C \sum (1 - e^{-y_i w^T x_i})$

↳作为正则化项.

Support Vector Machines (SVMs)

Question: what if data isn't perfectly linearly separable?
Replace "# mistakes" with upper bound called "hinge loss"

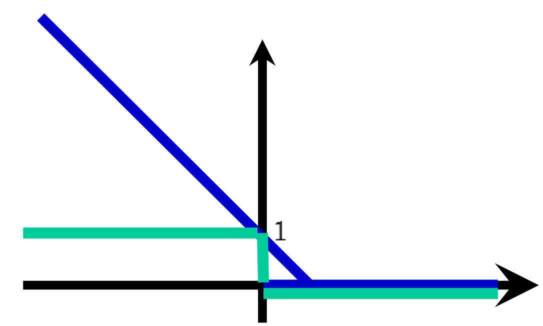


Input: S={(x_1, y_1), …,(x_m, y_m)};

Find $\text{argmin}_{w,\xi_1,\dots,\xi_m} ||w||^2 + C\sum_i \xi_i$ s.t.:

- For all i, $y_i w \cdot x_i \geq 1 - \xi_i$
  $\xi_i \geq 0$

Total amount have to move the points to get them on the correct side of the lines $w \cdot x = +1/-1$, where the distance between the lines $w \cdot x = 0$ and $w \cdot x = 1$ counts as "1 unit".



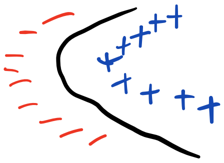
$l(w, x, y) = \max(0, 1 - y\, w \cdot x)$

What if the data is far from being linearly separable? 对于非线性的边界

Example:  vs 

No good linear separator in pixel representation.

SVM philosophy: "Use a Kernel"



$$\min f(x) \quad s.t.\ \underset{\lambda}{g(x)=0},\quad \underset{\nu}{h(x)\le 0}$$

拉格朗日：$L(x,\lambda,\nu) = f(x)+\lambda g(x)+\nu h(x)$

$$\Rightarrow \begin{cases} \min\limits_{x} \max\limits_{\lambda,\nu} L(x,\lambda,\nu) = p^* & \text{primal 原始} \\ \max\limits_{\lambda,\nu} \min\limits_{x} L(x,\lambda,\nu) = d^* & \text{dual 对偶} \end{cases} \Rightarrow p^* \ge d^*$$

convex opt $\Rightarrow p^* = d^*$

原始：$\min\limits_{w,b} \frac{1}{2}\|w\|^2$ s.t. $\underset{\alpha_i}{\underline{y_i w^T x_i \ge 1}}$ $\forall x_i \in S$ → 假设了完全线性可分.

(不考虑噪声)

$\hookrightarrow g(x) = y_i w^T x_i - 1 \ge 0$

$$L(w, \alpha_i) = \frac{1}{2}\|w\|^2 + \Sigma\alpha_i \underbrace{(y_i(w^T x_i + \overset{\text{截距}}{b}) - 1)}_{g(x)}$$

$$\begin{cases} \text{stationary}: & \frac{\partial L}{\partial w} = 0 \quad \frac{\partial L}{\partial \alpha} = 0 \quad \frac{\partial L}{\partial b} = 0 \\ \text{complementary}: & \alpha_i (y_i(w^T x_i + b) - 1) = 0 \\ \text{primal}: & y_i(w^T x_i + b) = 0 \\ \text{dual}: & \alpha_i \ge 0 \end{cases}$$

$$\frac{\partial L}{\partial w} = w - \Sigma \alpha_i y_i x_i = 0 \Rightarrow w = \Sigma \alpha_i y_i x_i$$

$$\Rightarrow \frac{1}{2}\|w\|^2 = \frac{1}{2}\sum_i\sum_j \alpha_i \alpha_j y_i y_j \underbrace{x_i^T x_j}_{\text{内积}}$$

可以在这里套用 kernel function. (对偶)

$$\frac{\partial L}{\partial b} = \Sigma \alpha_i y_i = 0$$

假设有m个点 $S=\{(x_1,y_1),\cdots,(x_m,y_m)\}$ $\Rightarrow$ 有m个不等式 $\Rightarrow$ m个 α_i x_i

$$\Rightarrow L(w,b,\alpha,\lambda)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\xi_i-\sum_{i=1}^{m}\alpha_i\left(y_i(w^Tx_i+b)-1+\xi_i\right)-\sum_{i=1}^{m}\xi_i\lambda_i$$

Stationary
$$\begin{cases}\dfrac{\partial L}{\partial w}=0 & \Rightarrow w=\sum_{i=1}^{m}y_i\alpha_i x_i\\[2ex] \dfrac{\partial L}{\partial b}=0 & \Rightarrow \sum_{i=1}^{m}\alpha_i y_i=0\\[2ex] \dfrac{\partial L}{\partial \xi_i}=0 & \Rightarrow C=\alpha_i+\lambda_i\end{cases}$$

Complementary
$$\begin{cases}\alpha_i\left(y_i(w^Tx_i+b)-1+\xi_i\right)=0\\ \xi_i\lambda_i=0 \quad \forall x_i\in S\\ \alpha_i\lambda_i\geq 0\end{cases}$$

$$y_i(w^Tx_i+b)\begin{cases}>1 & \xi_i=0 \quad \alpha_i=0\\[2ex] <1 & \xi_i>0 \quad \alpha_i=C\\[2ex] =1 & \xi_i=0 \quad 0\leq\alpha_i\leq C\end{cases}$$

Support Vector Machines (SVMs)

Input: S={(x_1, y_1), …,(x_m, y_m)};

Find $\text{argmin}_{w,\xi_1,\dots,\xi_m} ||w||^2 + C\sum_i \xi_i$ s.t.:

- For all i, $y_i w \cdot x_i \geq 1 - \xi_i$

  $\xi_i \geq 0$

Primal form

Which is equivalent to:

(线性可分时)

Input: S={(x_1, y_1), …,(x_m, y_m)};

Find $\text{argmin}_{\alpha} \frac{1}{2}\sum_i \sum_j y_i y_j \alpha_i \alpha_j \boxed{x_i \cdot x_j} - \sum_i \alpha_i$ s.t.:

内积

- For all i, $0 \leq \alpha_i \leq C_i$ → 对偶的原始约束.

  $\sum_i y_i \alpha_i = 0$ → 截距求导的约束

Lagrangian Dual

有了引那么提一个有误差的估计 ⇒ 会有新的约束：$\alpha_i \leq C$

SVMs (Lagrangian Dual)

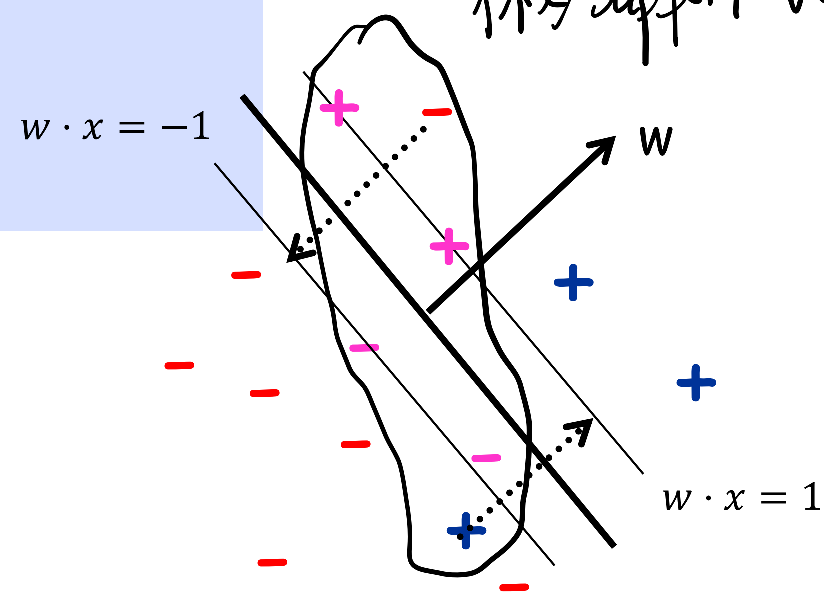
Input: $S=\{(x_1, y_1), \ldots, (x_m, y_m)\}$;

Find $\arg\min_\alpha \frac{1}{2}\sum_i \sum_j y_i y_j \alpha_i \alpha_j x_i \cdot x_j - \sum_i \alpha_i$ s.t.:

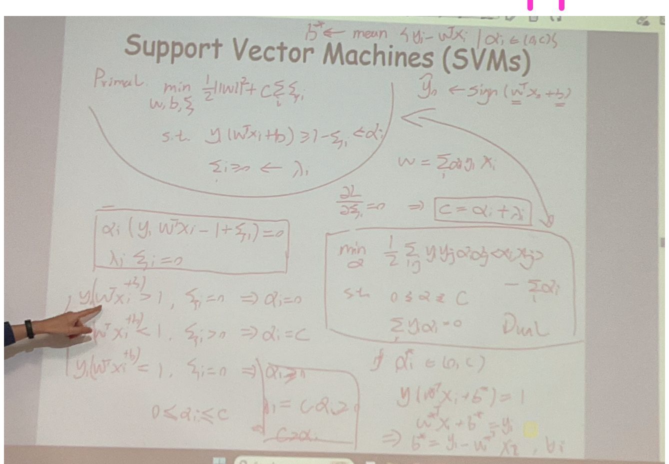
- For all i, $0 \leq \alpha_i \leq C_i$

$$\sum_i y_i \alpha_i = 0$$

只有这些点、的 $\alpha_i \neq 0$ 称为 Support Vector



- Final classifier is: $w = \sum_i \alpha_i y_i x_i$
- The points x_i for which $\alpha_i \neq 0$ are called the "support vectors"



$$w^T x_0 = \sum \alpha_i y_i (x_i^T x_0) \Rightarrow \sum \alpha_i y_i k(x_i, x_0)$$

要记录哪个 $\alpha_i \neq 0$

Kernelizing the Dual SVMs

<u>Input</u>: $S=\{(x_1,y_1), \ldots,(x_m,y_m)\}$;

<u>Find</u> $\quad \mathrm{argmin}_\alpha \frac{1}{2}\sum_i \sum_j y_i y_j\, \alpha_i \alpha_j x_i \cdot x_j - \sum_i \alpha_i$ s.t.:

- For all i, $\quad 0 \le \alpha_i \le C_i$

$$\sum_i y_i \alpha_i = 0$$

Replace $x_i \cdot x_j$ with $K(x_i, x_j)$.

对对偶问题优化求解

↳ 串行最小化优化问题 => $O(m)$ 线性复杂度.

- Final classifier is: $w = \sum_i \alpha_i y_i x_i$
- The points x_i for which $\alpha_i \neq 0$ are called the "support vectors"
- With a kernel, classify x using $\sum_i \alpha_i y_i K(x, x_i)$

Support Vector Machines (SVMs).

One of the most theoretically well motivated and practically most effective classification algorithms in machine learning.

Directly motivated by Margins and Kernels!

What you should know

- The importance of margins in machine learning.
- The primal form of the SVM optimization problem
- The dual form of the SVM optimization problem.
- Kernelizing SVM.
- Think about how it's related to Regularized Logistic Regression.