03/23/15

SUPPORT VECTOR MACHINES (SVM)

(1) Invented by Vapnik & Chervonenkis in '63.
"Modern" form by Cortes & Vapnik in '93 & '95.

→ Intuition: Don't solve a harder problem as an intermediate step towards your goal.

If you care about low-loss (misclassification rate), then optimize for that.

→ Goal: Classification

(1) Generative approach: estimate $P(\vec{X}|Y)$, $P(Y)$ $\Rightarrow$ Bayes Rule $\Rightarrow P(Y|\vec{X}) \Rightarrow \hat{y}_{MAP}$
(NB)

(2) Discriminative approach: estimate $P(Y|\vec{X})$ $\Rightarrow \hat{y}_{MAP} = \arg\max P(Y|\vec{X})$
(LR)

(3) Discriminative #2: estimate $\text{Score}(x) = w^T x$ (or $w^T x + b$)
(SVM)
$\hat{y} = \text{sign}(w^T x + b)$

Notation Change: → $y \in \{+1, -1\}$ (rather than $\{0, 1\}$)

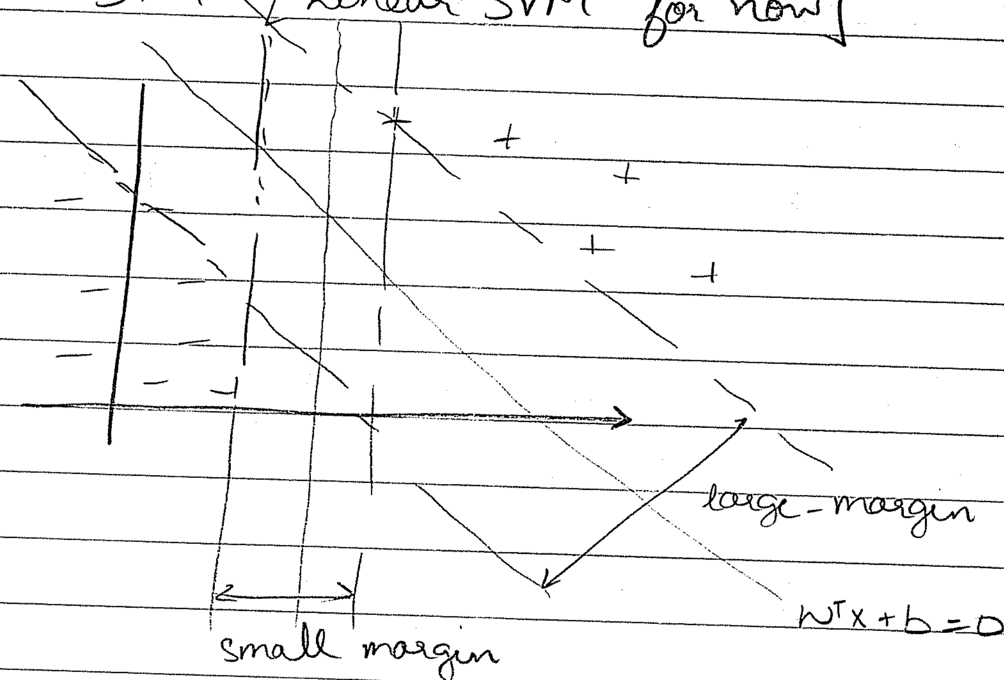
→ $\vec{w}$ no longer has bias weight included

So $\vec{w}^T \vec{x} + b \geq 0$ rather than $\vec{w}^T \vec{x} \geq 0$

$\vec{w} \in \mathbb{R}^d$ $\qquad$ $\vec{w} \in \mathbb{R}^{d+1}$

$\vec{x}$

② SVM [Linear SVM for now]



→ Many different $(\vec{w}, b)$ work well (actually perfectly) on training data. Which one should we choose?

→ Ans: The one with the largest margin!
Why?
→ Trust me, I'm Vapnik
→ Math will be easy & elegant.
→ There will be bounds on generalization!
→ Many different interpretations lead to this; Has to be correct!

Mathematically (or pseudo-mathematically for now)

$$\max_{\vec{w}, b} \quad \text{Margin}$$

$$\text{s.t.} \quad \text{Correct Classification on training data}$$

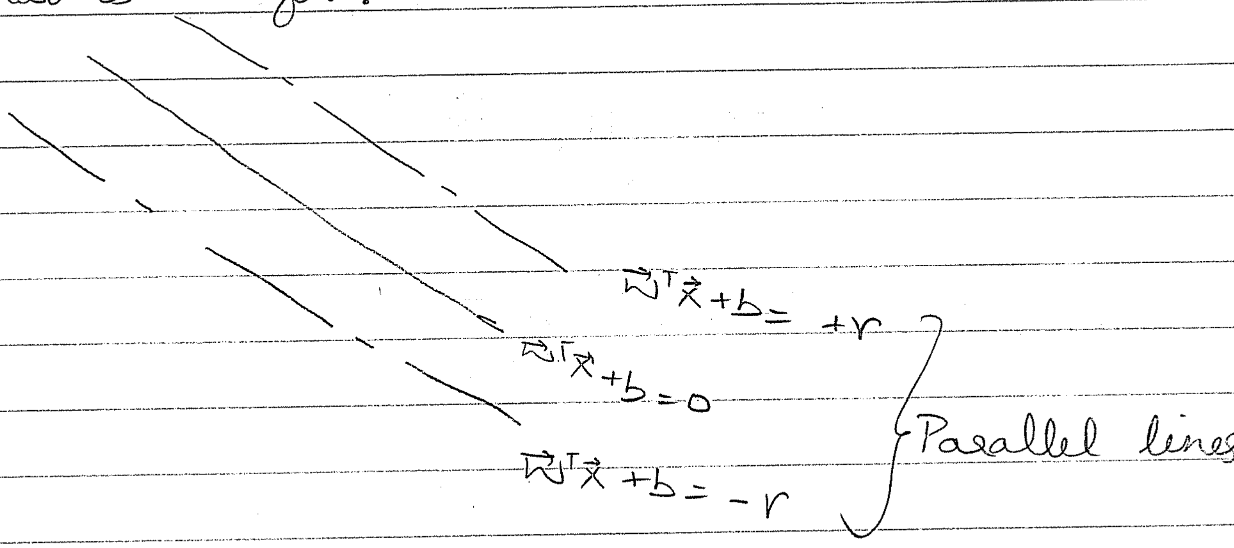
What is Correct Classification?

$$\Rightarrow \left\{ \begin{array}{ll} \vec{w}^T \vec{x}_i + b \geq 0 & \forall\ y_i = +1 \\ \vec{w}^T \vec{x}_i + b < 0 & \forall\ y_i = -1 \end{array} \right\}$$

together

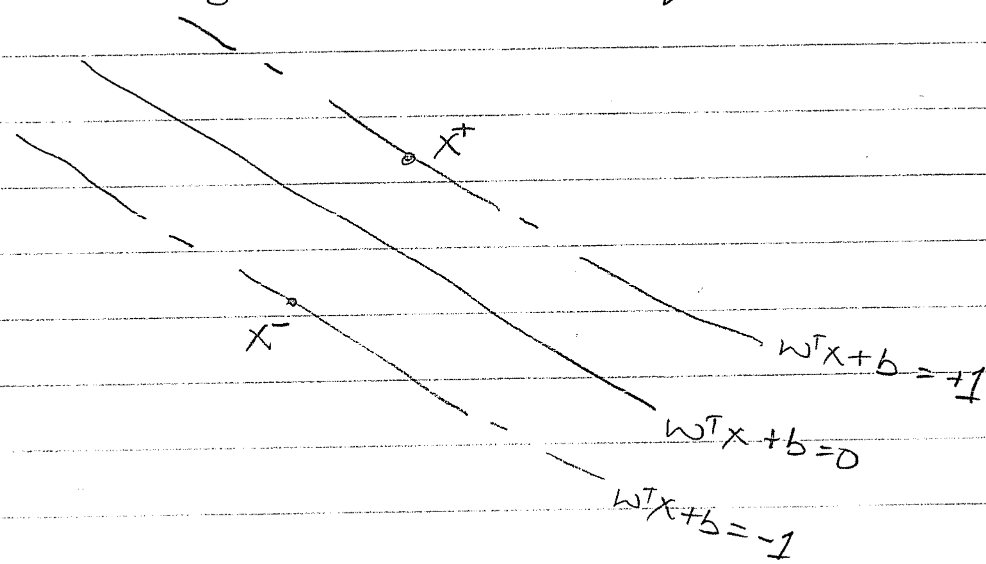
$$y_i (\vec{w}^T \vec{x}_i + b) \geq 0 \quad \forall i$$

What is Margin?



→ Notice that we can always scale both $\vec{w}$ & b to get arbitrary $\pm r$. So lets fix $r = +1$

Now



Let x^+ be a pt on $w^T x + b = +1$
" x^- be a pt on $w^T x + b = -1$

Let line from x^+ to x^- be orthogonal to our hyperplane $w^T x + b = 0$

$$\Rightarrow \quad \|x^+ - x^-\| = \text{Margin}$$

We know $(x^+ - x^-) \perp$ hyperplane

$$\Rightarrow \quad (x^+ - x^-) \underbrace{\parallel}_{\text{parallel}} \vec{w}$$

$$\Rightarrow \quad (x^+ - x^-) = \underbrace{\lambda}_{\text{scalar}} \vec{w}$$

Let's left & right multiply by $\vec{w}^T$

$$w^T (x^+ - x^-) = \lambda w^T w$$

$$\underbrace{w^T x^+}_{(1-b)} - \underbrace{w^T x^-}_{(-1-b)} = \lambda w^T w$$

$$(1 - b) - (-1 - b) = \lambda w^T w$$

$$\Rightarrow \quad \lambda = \frac{2}{w^T w} \quad \Rightarrow \quad \boxed{\begin{aligned} \text{Margin} &= \|x^+ - x^-\| \\ &= \frac{2}{w^T w} \|\vec{w}\| = \boxed{\frac{2}{\|\vec{w}\|}} \end{aligned}}$$

Finally Linear SVM with a "hard margin"

$$\max_{\vec{w},b} \frac{2}{\|w\|} \Big\} \text{ Not very nice to optimize}$$

s.t $y_i(\vec{w}^T\vec{x}_i + b) \geq 0$

2 changes

$$\approx \begin{cases} \min_{\vec{w},b} \frac{1}{2}\|w\|^2 \equiv \frac{1}{2}w^Tw \\ \text{s.t} \quad y_i(\vec{w}^T\vec{x}_i + b) \geq +1 \quad \forall i \end{cases}$$

why?

Very well studied optimization problem

→ Quadratic Program (QP)

→ $w^Tw \equiv$ quadratic in $\vec{w}$ & b

→ Convex

→ Hessian $= \frac{1}{2}\begin{bmatrix} I_{d \times d} & 0 \\ 0 & 0 \end{bmatrix} \geq 0$ (PSD)

→ $d+1$ variables, N constraints

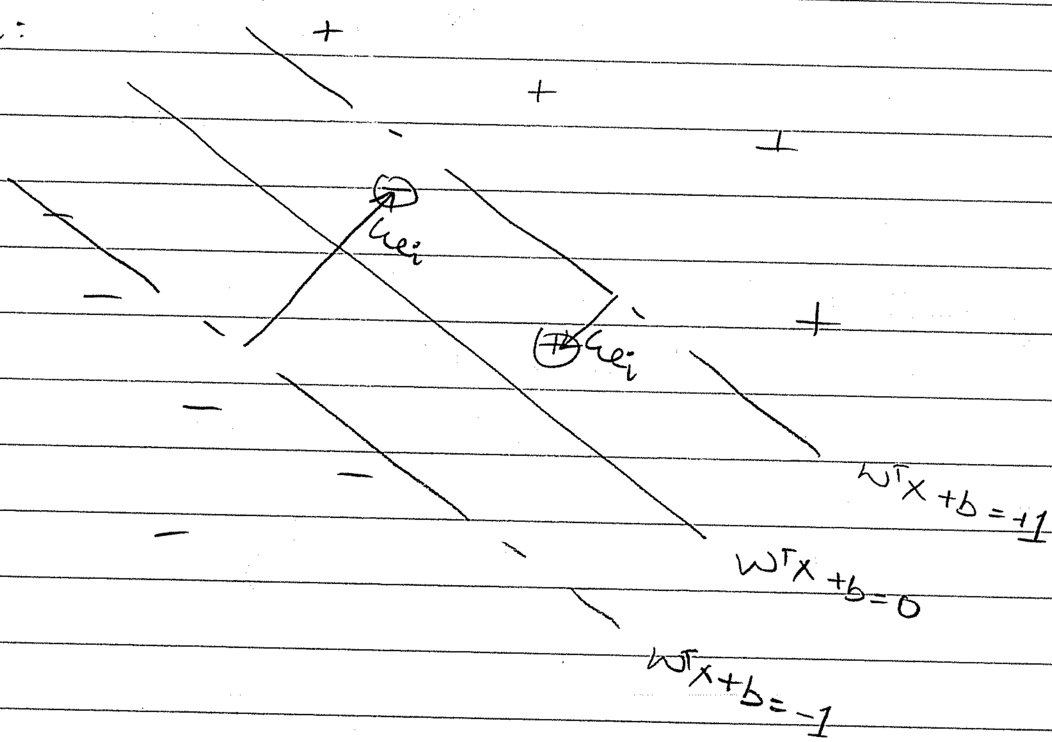
→ Standard QP solvers will solve this optimally for you.

(3) *Soft-Margin Linear SVM

→ What if data is not linearly separable?
→ QP will be infeasible ⇒ no solution

→ How can we allow for some mistakes?
→ Idea: allow some "slack"
allow every data-point to violate constraint by some "slack" ϵ_i

Picture:



$$\min_{\vec{w}, b, \epsilon_i} \quad \frac{1}{2} w^T w + \underbrace{C \sum_{i=1}^{N} \epsilon_i}_{\text{penalize the violations / slacks}}$$

s.t.

$$\vec{w}^T \vec{x}_i + b \geq 1 - \epsilon_i \qquad \forall i$$

ϵ_i → violation in margin

$$\epsilon_i \geq 0$$

0 → can't set slack to negative to game the objective

→ Still a convex QP

→ $\underbrace{d}_{\vec{w}}+\underbrace{1}_{b}+\underbrace{N}_{\{\xi_i\}}$ variables

→ 2N constraints

$C \equiv$ trade-off parameter

$C=0 \Rightarrow w^* = \vec{0} \quad b^*=0$ (Opt soln)

$C=\infty \Rightarrow$ Hard-Margin SVM
(might be infeasible for linearly non-separable data)

4 Hinge-Loss

Hinge Function is defined as:

$h(z) = \max\{0, 1-z\}$

In SVM QP

$$\left\{\begin{array}{l} \xi_i \geq 1 - y_i(w^T x_i + b) \\ \xi_i \geq 0 \end{array}\right\}$$

$$\Rightarrow \xi_i \geq \max\{0, 1-y_i(w^T x_i+b)\}$$

Almost like $h(y_i(w^Tx_i+b))$ but inequality

Let's fix that

Let $s_i \equiv$ score by SVM of ith data-point
$= w^T x_i + b$

$$\varepsilon_{e_i} \geq \max\{0, 1 - y_i s_i\}$$

Claim: Let $(w^*, b^*, \varepsilon_{e_i}^*)$ be QP optimum.
$$\varepsilon_{e_i}^* = \max\{0, 1 - y_i s_i^*\}$$

Proof: By contradiction. Assume equality doesn't hold

$$\Rightarrow \varepsilon_{e_i}^* > \max\{0, 1 - y_i(w_i^{*T} x_i + b^*)\}$$

Define $\varepsilon_{e_i}^{new} = \max\{0, 1 - y_i(w_i^{*T} x_i + b)\}$

by hinge loss construction
$$\begin{cases} \varepsilon_{e_i}^{new} \geq 0 \\ \varepsilon_{e_i}^{new} \geq 1 - y_i(w_i^{*T} x_i + b) \end{cases}$$

$\Rightarrow (w^*, b^*, \varepsilon_{e_i}^{new})$ is a FEASIBLE soln to SVM QP

Also $\{w^*, b^*, \varepsilon_{e_i}^{new}\}$ reduces objective by
$$C \sum_{i=1}^{N} (\varepsilon_{e_i}^* - \varepsilon_{e_i}^{new}) > 0$$

$\Rightarrow (w^*, b^*, \varepsilon_e^*)$ is NOT OPT.

$\Rightarrow$ contradiction

$\Rightarrow$ QED

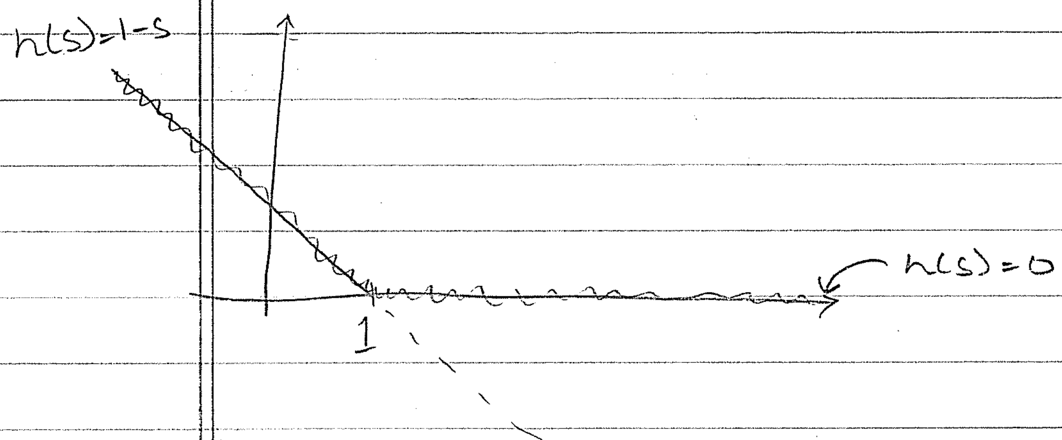
So $\xi_i^* = \max\{0, 1 - y_i(\vec{w}^{*T} x_i + b^*)\}$

$\Rightarrow$ SVM QP can be re-written as

$$\left[\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{N} h\left(y_i(\vec{w}^T \vec{x}_i + b)\right)\right]$$

Very cool! Begining to look like things we recognize LOSS + Regularization

Unconstrained QP where
$h(z) = \max\{0, 1-z\}$ Hinge-Loss!



(5) Comparison between LR & SVM

$\rightarrow$ SVM: $\min_{\vec{w}}$ hinge-loss $+ \frac{1}{2} w^T w$

$\rightarrow$ LR: $\max_{w} \log P(y|x,w) - \lambda w^T w$

Can we compare the first terms?

$\rightarrow$ Consider just 1 training pt $(\vec{x}_1, y_1 = +1)$

[Datasets are never this small, but just for illustration]

LR	SVM
$P(Y=1 \mid X, w) = \frac{1}{1+e^{-w^T x - b}}$	
$\vec{w}, b \sim N(\vec{0}, t^2 I)$	max Margin s.t Correct Class

$$\hat{w}_{MAP}, \hat{b}_{MAP} = \underset{w,b}{\operatorname{argmax}} \log P(Y=1 \mid \vec{x}, w, b) + \log P(w, b)$$

$$= \underset{\vec{w}, \vec{b}}{\operatorname{argmax}} \log\left[\frac{1}{1+e^{-w^T x - b}}\right] - \frac{1}{\sqrt{2\pi t^2}} e^{-\frac{w^T w}{2t^2}}$$

$$= \underset{\vec{w}, b}{\operatorname{argmin}} \frac{1}{2} w^T w + \lambda \underbrace{\log(1+e^{-w^T x + b})}_{\text{log-loss}}$$

$$\equiv \min_{w,b} \frac{1}{2} w^T w + \text{hinge}(w^T x + b)$$