

# Ch 15 SVM

Friday, March 27, 2020 3:05 PM

## Support Vector Machines (SVM)

Mostly following ch 15 in Shalev-Schwartz & Ben-David

SVM use linear classifiers,

well-understood theory and good practical performance

- First analyzed in the simpler separable case by Vapnik & Chervonenkis 1964

- Margin-based analysis + popularization + name due to

Corinna Cortes & Vladimir Vapnik 1995

now head of google research

and Boser, Guyon, Vapnik '92

Can be used w/ Kernels ... to be discussed next chapter

Setup:  $S = \{z_i\}_{i=1}^m$ ,  $z_i = (\vec{x}_i, y_i)$ ,  $y_i \in \{\pm 1\}$   
binary classification

Assuming data is (linearly) separable (for now)

$h(x) = \langle w, x \rangle + b$  is our model for  $\text{fl}$  (aka  $L_d$  in ch. 9)

Assume  $\exists (w, b)$  s.t.  $y_i = \text{sign}(\langle w, x_i \rangle + b) \forall i \in [m]$   
i.e.  $y_i (\langle w, x_i \rangle + b) > 0$

As before, often work w/ "homogeneous" case ( $b=0$ )

by including a new dimension,  $w \leftarrow \begin{bmatrix} w \\ b \end{bmatrix} \quad x \leftarrow \begin{bmatrix} x \\ 1 \end{bmatrix}$

In that case, in ch 9 we derived the ERM problem  
for half space classifiers as the linear program (LP)

Find  $w$  st.  $y_i \langle w, x_i \rangle \geq 1 \quad \forall i \in [m]$

(or...  $\min_w \frac{1}{2} \|w\|^2$  st.  $y_i \langle w, x_i \rangle \geq 1 \quad \forall i \in [m]$  )

and also  $\text{VCdim}(\text{fl} = \text{Sign} \circ L_d) = d+1$

**Goal of SVM:** - remove dependence on ambient dimension of

- extend to non-separable case

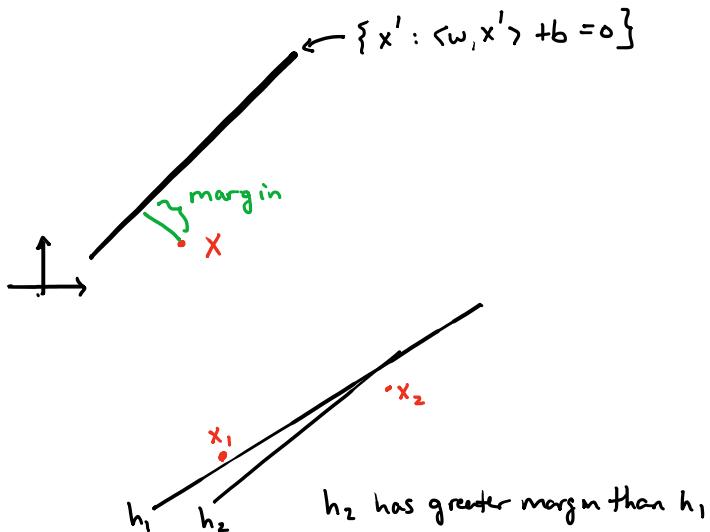
... related issue:

in separable case, which  $w$  to choose (since ERM not unique)

Main idea: maximize the margin

Def the (geometric) margin between a point  $x$  and the hyperplane defined by  $\{x' : \langle w, x' \rangle + b = 0\}$  is the distance\* between them, and the margin between  $\{x_i\}_{i \in [m]}$  and the hyperplane is the minimum of the individual margins

\* in the appropriate Hilbert space norm



**Fact (Claim 15.1)** Let  $h = \{v : \langle w, v \rangle + b = 0\}$ ,

then if  $\|w\|=1$ ,  $\text{dist}(x, h) = |\langle w, x \rangle + b|$

Proof  $\text{dist}^2(x, h) = \min \|x - v\|^2$  s.t.  $\langle w, v \rangle + b = 0$

use Lagrange multipliers:

$$\mathcal{L}(v, \lambda) = \|x - v\|^2 + \lambda (\langle w, v \rangle + b)$$

$$\min \mathcal{L}(v, \lambda) \Rightarrow 0 = 2(v - x) + \lambda w$$

$$\text{so } v = x - \frac{\lambda}{2}w, \text{ find } \lambda \text{ by plugging into equality}$$

$$\langle w, x - \frac{\lambda}{2}w \rangle + b = 0$$

$$\Rightarrow \langle w, x \rangle - \frac{\lambda}{2} + b = 0, \quad \frac{\lambda}{2} = \langle w, x \rangle + b$$

$$\text{so } v = x - (\langle w, x \rangle + b)w$$

thus  $\|x - v\| = |\langle w, x \rangle + b| \cdot \|w\|^{-1}$  □  
 Thus, if we want the maximum margin linear classifier,

solve

$$\begin{array}{ll} \max_{(w,b) \in \mathbb{R}^{d+1}} & \min_{i \in [m]} |\langle w, x_i \rangle + b| \\ \text{s.t. } \|w\| = 1 & \text{margin} \\ \uparrow \text{wlog} & \text{s.t. } y_i(\langle w, x_i \rangle + b) > 0 \\ & \text{correct classification} \\ & \swarrow \text{can merge (Exercise 15.1)} \end{array}$$

$$\max_{(w,b)} \min_{i \in [m]} y_i(\langle w, x_i \rangle + b)$$

or, even nicer, due to strict (since strongly) convexity, min is unique

**HARD-SVM**

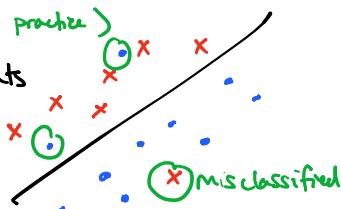
$$(w_0, b_0) = \arg \min \|w\|^2 \text{ s.t. } y_i(\langle w, x_i \rangle + b) \geq 0$$

Convex quadratic (theoretical...  
and  $w \leftarrow w_0/\|w_0\|$  and  $b \leftarrow b_0/\|w_0\|$  not used in practice)

You can analyze via Rademacher complexity  
but we're going to proceed to "Soft"-SVM

No longer assuming data is separable: soft-SVM  
(the version you use in practice)

Allow some misclassified data points



SOFT-SVM

$$\begin{array}{ll} \min_{w, b, \xi} & \lambda \|w\|^2 + \frac{1}{m} \sum \xi_i \\ \text{s.t. } & y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \end{array}$$

$\xi_i > 0$  means  $(x_i, y_i)$  misclassified

so penalize  $\frac{1}{m} \sum \xi_i$  in objective

slack variable

$$\xi_i \geq 0$$

$y_i \in \{-1, 1\}$

and parameter  $\lambda > 0$  describes how much we're willing

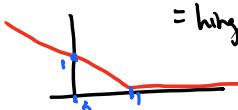
to tolerate misclassifications (since all else being equal,  
small  $\|w\|$  is good: smaller R.C.)

Can rewrite Soft-SVM in 2 ways

(1) in terms of sample risk of a loss function

$$\text{hinge}((w, b), (x, y)) = \max(0, 1 - y(\langle w, x \rangle + b))$$

$$= \text{hinge}(y(\langle w, x \rangle + b))$$



(we've seen before:  
a convex surrogate  
for 0-1 loss)

$$\text{and } \hat{L}_S^{\text{hinge}}((w, b)) = \frac{1}{m} \sum_{i=1}^m \text{hinge}((w, b), (x_i, y_i))$$

then can write Soft-SVM as

$$\min_{w, b} \hat{L}_S^{\text{hinge}}((w, b)) + \lambda \|w\|^2$$

ERM      RLM

*proof of equivalence ... straightforward:*  
*constraint  $y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$*   
*is really an equality*  $\square$

## ② via duality

(SOFT-)  
SVM  
DUAL

$$\max_{d \in \mathbb{R}^m} \sum_{i=1}^m d_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m d_i d_j y_i y_j \langle x_i, x_j \rangle$$

*or in matrix notation*

$$d^T d - \frac{1}{2} d^T G d,$$

$$G = \begin{matrix} \diag(d) \\ \vdots \\ \diag(d) \end{matrix} \quad G \geq 0 \quad \text{So convex}$$

### Comments

- In optimization, sometimes primal or dual is easier to solve. Some solvers solve both simultaneously (usually at the end you can convert a dual sol'n into a primal sol'n)

in our case,  $w = \sum_{i=1}^m d_i y_i x_i$  (see also Thm 15.8)

Must  $d_i = 0 \dots$   
 the  $x_i$  corresponding to  $d_i > 0$   
 are "Support vectors"

- The dual problem only involves  $\vec{x}_i$  in that it needs  $\langle x_i, x_j \rangle \forall i, j \in [m]$  and nothing more.  
 This is the basis of the **Kernel trick**  
 that we'll discuss in the next chapter
- There are many algorithms to solve Soft SVM since it's a nicely structured convex problem
  - Book shows a SGD (Subgradient) algo.
  - libLinear is well-known (partly for its examples)
  - Scikit Learn, etc., have implementations

## Soft-SVM Sample Complexity

Fact:  $L^{\text{hinge}}$ , for  $b=0$ , (ie.,  $f(w) = \lfloor 1 - y \langle w, x \rangle \rfloor_+$ )

$$\begin{aligned} & \text{is } \|x\| \text{-Lipschitz} \\ (\text{proof}) \quad & \delta f(w) = \begin{cases} -y \cdot x & \text{if... either way,} \\ 0 & \text{else... } |y|=1, \\ & \Rightarrow \|\delta f(w)\| \leq \|x\| \end{cases} \end{aligned}$$

Thus, letting  $b=0$  wlog,  
we can apply the results from last chapter on  
RLM, using our reformulation of soft-SVM as

$$\min_w \lambda \|w\|^2 + L^{\text{hinge}}(w)$$

**Corollary 15.7** Let  $D$  be a distribution over  $X \times Y$  with  $\cancel{Y = \{0, 1\}}$   $\tilde{Y} = \{\pm 1\}$   
and assume  $\|x\| \leq \rho \forall x \in X$ . Then if  $A(S)$  is the soft-SVM algo.,

$$\textcircled{1} \quad \forall u \in \mathbb{R}^d, \quad \mathbb{E}_{S \sim D^m} L_D^{\text{hinge}}(A(S)) \leq L_D^{\text{hinge}}(u) + \lambda \|u\|^2 + \frac{2\rho^2}{\lambda m}$$

$$\textcircled{2} \quad \forall u \in \mathbb{R}^d, \quad \mathbb{E}_{S \sim D^m} L_D^{0-1}(A(S)) \leq L_D^{\text{hinge}}(u) + \lambda \|u\|^2 + \frac{2\rho^2}{\lambda m}$$

(follows immediately since  $L^{\text{hinge}} \geq L^{0-1}$ )

$$\textcircled{3} \quad \forall B > 0, \text{ choosing } \lambda = \sqrt{\frac{2\rho^2}{B^2 m}} \text{ then}$$

$$\mathbb{E}_S L_D^{0-1}(A(S)) \leq \mathbb{E}_S L_D^{\text{hinge}}(A(S)) \leq \min_{\|u\| \leq B} L_D^{\text{hinge}}(u) + \sqrt{\frac{8\rho^2 B^2}{m}}$$

Proof  $\textcircled{1}$  via Corollary 13.8,  $\textcircled{3}$  easy or via Cor. 13.9

Discussion:

Good result if  $\min_{\|u\| \leq B} L_D^{\text{hinge}}(u)$  is small

Sometimes true and may improve on VC-dim. based analysis