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 (now head of google research) & Vladimir Vapnik 1995

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 $\mathcal{H}$ (aka L_d in ch. 9)

Assume $\exists (w,b)$ s.t. $y_i = \text{sign}(\langle w, x_i \rangle + b_i)$ (ϕ in ch 9.) $\forall i \in [m]$

i.e. $y_i(\langle w, x_i \rangle + b_i) > 0$

As before, often work w/ "homogeneous" case ($b=0$)
by including a new dimension, $w \leftarrow \begin{bmatrix} w \\ b \end{bmatrix} \begin{matrix} \uparrow d \\ \downarrow 1 \end{matrix}$ $x \leftarrow \begin{bmatrix} x \\ 1 \end{bmatrix}$

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

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

(or... $\min_w o^T w$ s.t. $y_i \langle w, x_i \rangle \geq 1$ $\forall i = [m]$)

and also $VCdim(\mathcal{H} = \text{sign} \circ L_d) = d+1$

Goal of SVM: - remove dependence on ambient dimension d

- extend to non-separable case

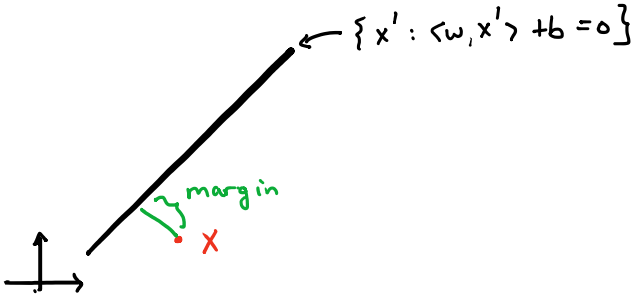
... related issue:

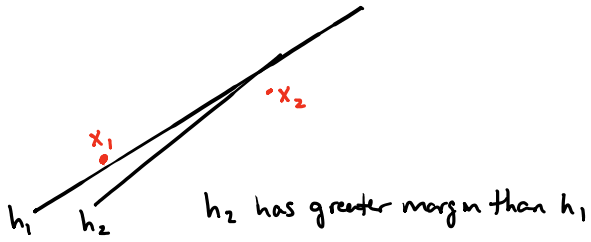
in separable case, which w to choose (since ERM not unique)

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





h_2 has greater margin than h_1

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:

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

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

so $v = x - \lambda/2 \, w$, find λ by plugging into equality

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

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

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

thus $\|x - v\| = |\langle w, x\rangle + b| \cdot \|w\|$ ($\|w\| = 1$) □

Thus, if we want the maximum margin linear classifier,

solve

$$\max_{(w,b)\in\mathbb{R}^{d+1},\ \|w\|=1} \ \min_{i\in[m]} |\langle w, x_i\rangle + b_i| \quad \text{s.t. } y_i(\langle w, x_i\rangle + b_i) > 0$$

(not good; margin; correct classification; $\|w\|=1$ ↑ wlog)

can merge (Exercise 15.1)

$$\max_{\substack{(w,b)\\ \|w\|=1}} \ \min_{i\in[m]} y_i(\langle w, x_i\rangle + b_i)$$

or, even nicer,

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

HARD-SVM: $(w_0, b_0) = \operatorname{argmin} \|w\|^2$ s.t. $y_i(\langle w, x_i\rangle + b_i) \geq 0$

(Convex quadratic)

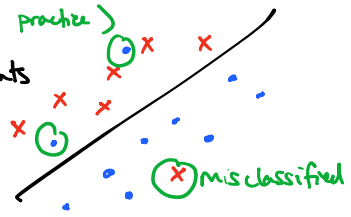
and $w \leftarrow w_0/\|w_0\|$ and $b \leftarrow b_0/\|w_0\|$

(theoretical... 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

$$\min_{w,b,\xi} \ \lambda\|w\|^2 + \frac{1}{m}\sum_i \xi_i \quad \text{s.t. } (\forall i\in[m])\ y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i$$

(slack variable)

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

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

$\xi_i \geq 0$

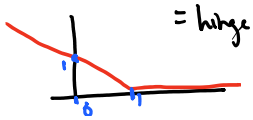
$(y_i \in \{\pm 1\})$

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

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



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

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

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 □

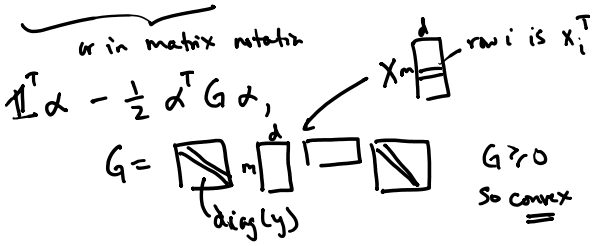
② via duality

(SOFT-) SVM DUAL

$$\max_{\alpha \in \mathbb{R}^m, \alpha \geq 0} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j\rangle$$

or in matrix notation

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



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 soln into a primal soln)
  - in our case, $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ (see also Thm 15.8)
  - Most $\alpha_i = 0$ ...
  - the x_i corresponding to $\alpha_i > 0$ 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^{hinge}, for $b=0$, (i.e., $f(w) = \lfloor 1 - y\langle w, x\rangle \rfloor_+$)

is $\|x\|$-Lipschitz

(*pf* $\partial f(w) = \begin{cases} -y\cdot\vec{x} & \text{if} \dots \\ 0 & \text{else} \dots \end{cases}$ either way, $|y|=1$, so $\|\partial f(w)\| \le \|x\|$)

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

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

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

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

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

(follows immediately since $\ell^{\text{hinge}} \ge \ell^{0\text{-}1}$)

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

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

Pf ① via Corollary 13.8, ③ easy or via Cor. 13.9

Discussion:

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

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