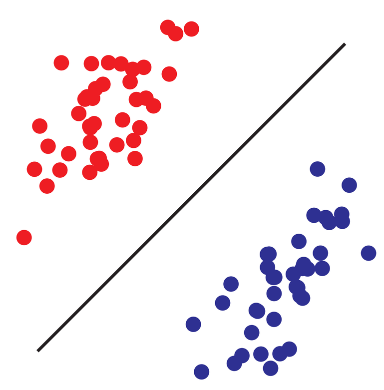
A New Perspective on Learning Linear Separators with Large L_qL_p Margins

Maria-Florina Balcan and Christopher Berlind

Georgia Tech College of Computing

Margins

Intuition: Learning should be easy when data is far from the decision boundary

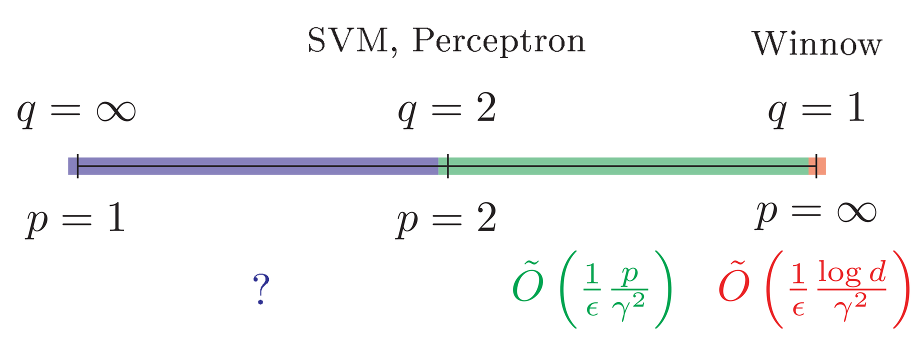


Definition: The L_qL_p margin of w w.r.t. D is

$$\gamma_{q,p}(w) = \inf_{x\sim D} \frac{|w \cdot x|}{\|w\|_q \|x\|_p}$$

where $1 \le p, q \le \infty$ and $1/p + 1/q = 1$.

The Margin Spectrum:



[Grove et al., 2001; Servedio, 2002]

Contributions

- Sample complexity bound covering the entire spectrum of margins
- Sufficient condition on data under which large margins lead to fast learning
- Upper and lower bounds for a family of problems showing a concrete advantage for $p = 1$
- Experimental confirmation that the theoretical results are relevant in practice

Discussion

- Important to consider entire margin spectrum
- Performance depends on both γ and $\|\mathbf{X}\|_{2,p}$
- Non-realizable case: L_q-norm regularization
- Relative sparsity of data and weight vector

Open questions:

- Algorithms that adaptively choose optimal p
- Generalization to multiple kernel learning
- Use $\|\mathbf{X}\|_{2,p}$ to aid feature selection

Experiments

L_qL_p SVM: Given a set of n training examples, we can efficiently solve the convex program

$$\min_w \quad \|w\|_q + C\sum_{i=1}^n \xi_i$$
$$\text{s.t.} \quad \frac{y^i(w\cdot x^i)}{\|x^i\|_p} \ge 1 - \xi_i, \quad 1 \le i \le n.$$

Equivalent to minimizing the hinge loss of an L_p-normalized data set using L_q-norm regularization.

Empirical advantage for $p = 1$ on synthetic data and on several real data sets from the UCI repository.

Generalization Bound

Theorem 1. *Let $\|X\|_p = \left(\sup_{x\sim D}\|x\|_p\right)$ and*

$$\|\mathbf{X}\|_{2,p} := \left(\sum_{i=1}^d\left(\sum_{j=1}^n |x_i^j|^2\right)^{p/2}\right)^{1/p}$$

If there are constants $C = C(d,p)$ and $0 \le \alpha < 1$ such that $\|\mathbf{X}\|_{2,p} \le Cn^\alpha \|X\|_p$ for any data set from D, then

$$\tilde{O}\left(\frac{1}{\epsilon}\left(\frac{C\sqrt{p}}{\gamma_{q,p}}\right)^{\frac{1}{1-\alpha}}\right)$$

*samples suffices to achieve error ϵ for **any** $1 \le p < \infty$.*

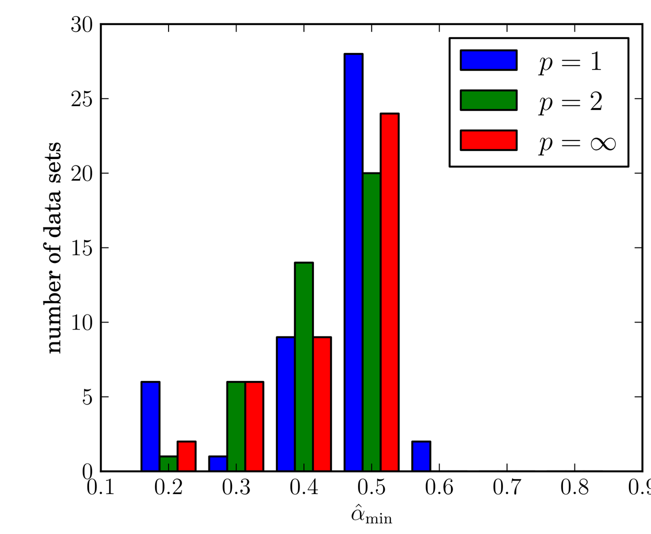
Proof summary:

- Novel bound on fat-shattering dimension
- Use Khintchine inequality and bound on $\|\mathbf{X}\|_{2,p}$
- Apply standard generalization error bound

Bounding the $L_{2,p}$-norm

Theoretically:
When $C = 1$, can always use $\alpha = 1/2$ when $p \ge 2$, but may need as much as $\alpha = 1/p$ when $p < 2$.

In reality:
For almost all data sets tested, we can bound $\|\mathbf{X}\|_{2,p}$ with $C = 1$ and $\alpha \le 1/2$, regardless of p.



Histogram of α estimates on 47 real data sets

Example 1: Unhelpful Margins

Basis vectors with $w^* \in \{-1,1\}^d$.

w^*: ++-+++-+--+--++---

D:
00010000000000000 +
00000000000100000 0 −
000000000100000000 −
⋮

p	1	2	∞
$\gamma_{q,p}(w^*)$	1	$1/\sqrt{d}$	$1/d$
$\|\mathbf{X}\|_{2,p}$	$\sqrt{dn}$	$\sqrt{n}$	$\sqrt{n/d}$
s.c.	$\tilde{O}(d/\epsilon)$	$\tilde{O}(d/\epsilon)$	$\tilde{O}(d/\epsilon)$

Also have lower bound of $\tilde{\Omega}(d)$.

Example 2: Helpful Margins

Divide the d coordinates evenly into k blocks.

w^*: ++++++------++++++

D:
111111000000000000 +
000000000000111111 +
000000111111000000 −
⋮

p	1	2	∞
$\gamma_{q,p}(w^*)$	1	$1/\sqrt{k}$	$1/k$
$\|\mathbf{X}\|_{2,p}$	$\sqrt{kn}$	$\sqrt{n}$	$\sqrt{n/k}$
s.c.	$\tilde{O}(k/\epsilon)$	$\tilde{O}(k/\epsilon)$	$\tilde{O}(k/\epsilon)$

Significant improvement when $k = o(d)$.

Making the Case for $L_\infty L_1$ Margins

Divide the d coordinates evenly into k blocks.

Distribution D randomly picks a block and either
- sets to 1 a single variable in the block or
- sets to 1 exactly $d/(2k)$ variables in the block.

Target w^* maximizes $L_\infty L_1$ margin.

w^*: ++++++------++++++

D:
001000000000000000 +
000000011001000000 −
000000000000110001 +
000000100000000000 −
⋮

Theorem 2. *If $k = O(d^{1/4})$ and $\epsilon = \Omega(d^{-1/4})$ in the above learning setting, then any algorithm restricted to using the large-margin class*

$$W_p = \{w \in \mathbb{R}^d : \gamma_{q,p}(w) \ge \gamma_{q,p}(w^*)\}$$

for a fixed p has sample complexity

$$p = 1: \quad \tilde{O}(\sqrt{d})$$
$$p > 1: \quad \tilde{\Omega}(d).$$

Proof summary:

- W_1: ++++++------ (unanimous on each block)
- W_2: +++-++-+---- (some dissenters allowed)
- W_∞: ++--+--+++-- (any $w \in \{-1,1\}^d$)
- Bound covering number for each W_p
- Apply distribution-specific s.c. bounds

Synthetic and Real Data Results

Synthetic data (top):
Blocks, Gaussian

Real data (bottom):
Fertility, SPECTF, CNAE-9