1. project proposal due on Mon 16 on gradescope

2. HW2:

problem 1. 2. normalized margin $= y \cdot \bar{f}_t(x)$ ($\bar{f}_t$: f_t)

problem 4: $F_m \to F_n$

σ: $\sigma(W_1 \cdot X)$

$\sigma([v_1 \cdots v_d]) = (\sigma(v_1) \cdots \sigma(v_d))$

$\langle W_1, X \rangle \to W_1 \cdot X$

l_2 − SVM.

$$\min_{w} \|w\|_2 \to \frac{1}{2}\|w\|_2^2$$

$$\text{s.t. } \forall i. \quad y_i \cdot \langle w, x_i \rangle \geq 1.$$

$\Updownarrow$ α $\hat{w}$ $\quad w = \alpha \cdot \hat{w}$ $\quad \alpha > 0$. $\|\hat{w}\| = 1$.

$$\min_{\alpha, \hat{w}: \alpha > 0, \|\hat{w}\|=1} \alpha$$

$$\text{s.t. } \forall i \quad y_i \, \alpha \cdot \langle \hat{w}, x_i \rangle \geq 1.$$

$\Updownarrow$

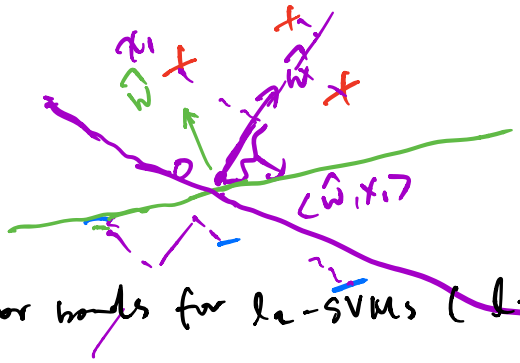
$$\max_{\alpha, \hat{w}: \alpha > 0, \|\hat{w}\|=1} \frac{1}{\alpha}$$

$$\text{s.t. } \forall i \quad y_i \quad \langle \hat{w}, x_i \rangle \geq \frac{1}{\alpha}.$$

for any fixed $\hat{w}$. the optimal choice of α is such that

$$\frac{1}{\alpha} = \min_i y_i \langle \hat{w}, x_i \rangle$$

$$\Updownarrow \quad \max_{\hat{w}: \|\hat{w}\|=1} \min_i y_i \langle \hat{w}, x_i \rangle$$



Generalization error bounds for ℓ_2-SVMs (ℓ_2-bounded linear predictors)

Thm: fix $B_2, R_2 > 0$, $S = (x_1, y_1) \dots (x_m, y_m) \overset{iid}{\sim} D$.

D supported on $\{x \in \mathbb{R}^d : \|x\|_2 \le R_2\} \times \{\pm 1\}$. Fix $\theta \in (0, B_2 R_2]$

Then, w.p. $1-\delta$, for all $w : \|w\|_2 \le B_2$:

$$P_D(y\langle w, x\rangle \le 0) \le P_S(y\langle w, x\rangle \le \theta) + O\left(\frac{B_2 R_2}{\theta}\sqrt{\frac{\ln^2/\delta}{m}}\right)$$

Pf: same strategy as the ℓ_1/ℓ_∞ margin bound:

- introduce ramp loss
- uniform concentration of ramp losses
- contraction ineq. of Rademacher complexity

Pf comes down to show: given $S = (x_1, y_1) \dots (x_m, y_m)$.

s.t. $\forall i$. $\|x_i\|_2 \le R_2$.

$$\mathcal{G} = \{ m_w : \|w\|_2 \le B_2 \}$$

$$m_w(x, y) = y\langle w, x\rangle.$$

Bound $\quad \mathrm{Rad}_S(\mathcal{G}) = \frac{1}{m} \mathbb{E}_{\sigma \sim U(\pm 1)^m} \sup_{w : \|w\|_2 \le B_2} \sum_{i=1}^m \sigma_i y_i \langle w, x_i \rangle$

$$= \frac{1}{m} \mathbb{E}_{\sigma \sim U(\pm 1)^m} \sup_{w : \|w\|_2 \le B_2} \sum_{i=1}^m \sigma_i \langle w, x_i \rangle$$

$$= \frac{1}{m} \mathbb{E}_{\sigma \sim U(\pm 1)^m} \sup_{w : \|w\|_2 \le B_2} \left\langle w, \sum_{i=1}^m \sigma_i x_i \right\rangle$$

Cauchy Schwarz: $\langle \alpha, \beta \rangle \le \|\alpha\|_2 \cdot \|\beta\|_2$

$$\le \|w\|_2 \left\| \sum_{i=1}^m \sigma_i x_i \right\|_2$$

$$\le B_2 \left\| \sum_{i=1}^m \sigma_i x_i \right\|_2$$

$(\mathbb{E} Z)^2 \le \mathbb{E}[Z^2]$.

$$\le \frac{B_2}{m} \mathbb{E}_\sigma \left\| \sum_{i=1}^m \sigma_i x_i \right\|_2 \qquad Z$$

$$\le \frac{B_2}{m} \cdot \sqrt{\mathbb{E}_\sigma \left\| \sum_{i=1}^m \sigma_i x_i \right\|_2^2}$$

$$= \frac{B_2}{m} \cdot \sqrt{\mathbb{E}_\sigma \left[\left\langle \sum_{i=1}^m \sigma_i x_i, \sum_{j=1}^m \sigma_j x_j \right\rangle \right]}$$

Linearity of inner product

$$= \frac{B_2}{m} \cdot \sqrt{\mathbb{E}_\sigma \left[\sum_{i=1}^m \sum_{j=1}^m \sigma_i \sigma_j \langle x_i, x_j \rangle \right]}$$

$$= \frac{B_2}{m} \cdot \sqrt{\sum_i \sum_j \mathbb{E}_\sigma[\sigma_i \sigma_j] \langle x_i, x_j \rangle}$$

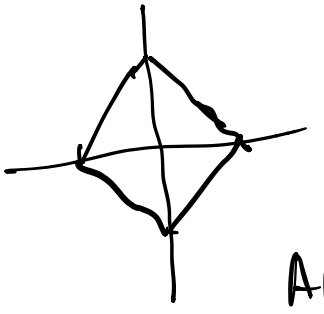
$$\begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$$

$$= \frac{B_2}{m} \cdot \sqrt{\sum_{i=1}^{m} \underbrace{\langle x_i, x_i \rangle}_{= \|x_i\|_2^2 \le R_2^2}}$$

$$\le \frac{B_2}{m} \cdot \sqrt{m \cdot R_2^2} = \frac{B_2 R_2}{\sqrt{m}} \quad \text{✗}$$

comparision b/w ℓ_1/ℓ_∞ & ℓ_2/ℓ_2 margin bounds

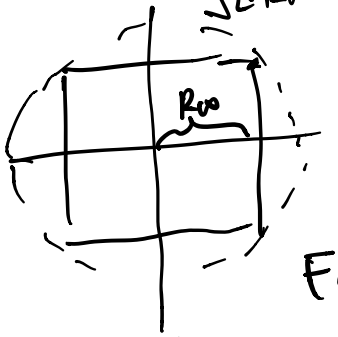
	constraints on x	constraint on w	error bound
ℓ_1/ℓ_∞	$\|x\|_\infty \le R_\infty$ ($\max_i \|x_i\|$)	$\|w\|_1 \le B_1$	$\tilde{O}\left(\frac{B_1 R_\infty}{\theta}\sqrt{\frac{1}{m}}\right)$
ℓ_2/ℓ_2	$\|x\|_2 \le R_2$	$\|w\|_2 \le B_2$	$\tilde{O}\left(\frac{B_2 R_2}{\theta}\sqrt{\frac{1}{m}}\right)$



incomparable in general

Applying ℓ_2/ℓ_2 generalization bound to ℓ_1/ℓ_∞ setting:

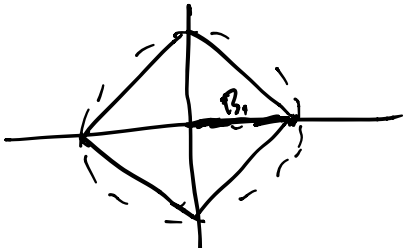
$\|x\|_\infty \le R_\infty \Rightarrow \|x\|_2 \le R_2$?



$R_2 = \sqrt{d} \cdot R_\infty$.

$B_2 = B_1$

Fact $\|w\|_2 \le \|w\|_1$



will get a bound in terms of

$R_2 B_2 = \sqrt{d} \cdot R_\infty \cdot B_1$ a factor of $\sqrt{d}$

worse than the original ℓ_1/ℓ_∞ bound.

Applying ℓ_1/ℓ_∞ bound to the ℓ_2/ℓ_2 setting.

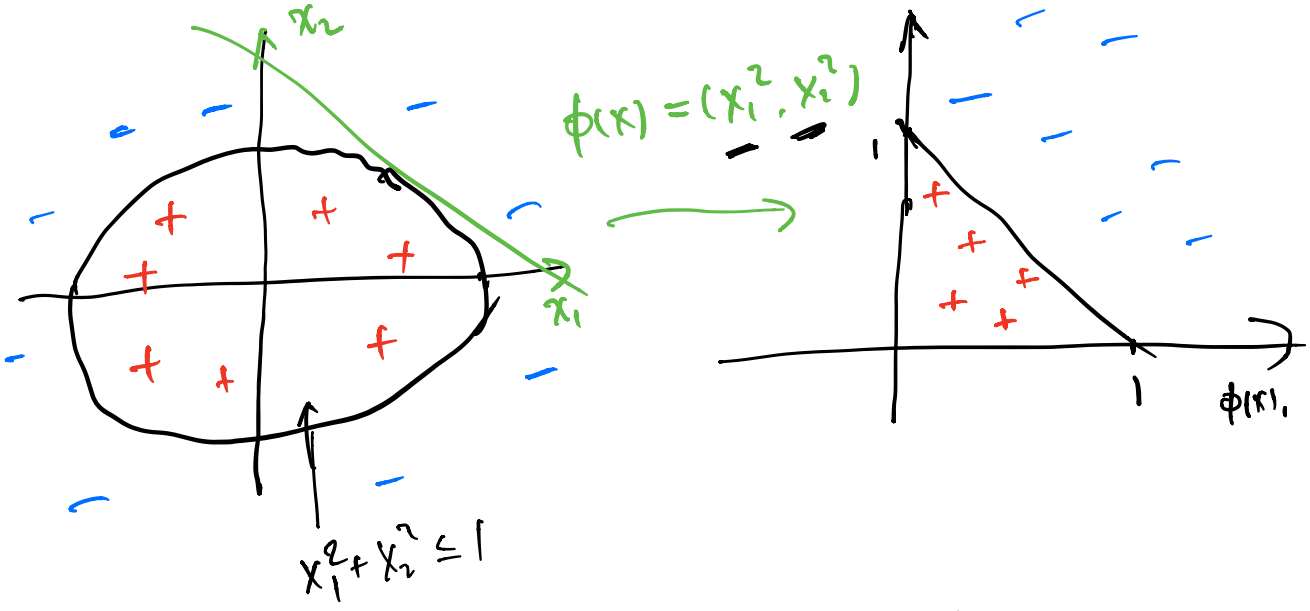
(will be $\sqrt{d}$ factor worse. (left as exercise).

Coping w/ linear non separability in SVMs.

Ideas :

1. introduce nonlinear feature maps (basis fns)
2. relax the SVM formulation : allowing some samples to be incorrectly classified.

1.



- define $\phi: \mathbb{R}^d \to \mathbb{R}^m$ $\quad (x_i, y_i) \to (\phi(x_i), y_i)$
- Solve SVM on $(\phi(x_i), y_i)$'s. $\longrightarrow$ $\hat{w} \in \mathbb{R}^m$
- Final predictor : $\text{sign}(\langle \phi(x), \hat{w} \rangle)$
- there are SVM training algs that has time complexity independent of m.

$= k(x,y)$: kernel

(If $\langle \phi(x), \phi(y) \rangle$ can be evaluated in time independent of m). — Kernel trick

2. Soft margin SVMs.

$$\min_{w,(\xi_1,\dots,\xi_m)\geq 0} \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^{m} \xi_i$$

s.t. $\forall i.\quad y_i \langle w, x_i \rangle \geq 1 - \xi_i$

$\lambda \downarrow$: more likely to classify more correctly.

$\|w\|$ will likely to be large.

fix any w: the optimal choices of ξ_i's are:

$$\left.\begin{array}{l} \xi_i \geq 1 - y_i\langle w, x_i\rangle \\ \xi_i \geq 0 \end{array}\right] \Leftrightarrow \xi_i \geq \max(0, 1 - y_i\langle w, x_i\rangle)$$

the optimal choices of ξ_i's $= \max(0, 1 - y_i\langle w, x_i\rangle)$.

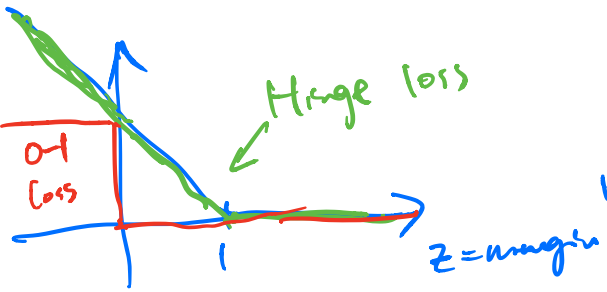
$$\min_w \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^{m} \max(0, 1 - y_i\langle w, x_i\rangle)$$

$\frac{\lambda}{2}\|w\|_2^2$: complexity

$\sum_{i=1}^{m} \max(0, 1 - y_i\langle w, x_i\rangle)$: empirical risk

$= \phi(y_i\langle w, x_i\rangle)$

where $\phi(z) = \max(0, 1-z)$.



Regularized loss minimization:

$$\min_w \quad \lambda \cdot R(w) + \sum_{i=1}^{m} \phi(f_w(x_i), y_i)$$

$R(w)$: free to change.

Notable examples:

$R(w) = \frac{1}{2}\|w\|_2^2$. $\phi(f_w(x_i), y_i) = (f_w(x_i) - y_i)^2$

$f_w(x) = \langle w, x \rangle$.

$\lambda = 0$: Ordinary least squares.

$\lambda > 0$: ridge regression.

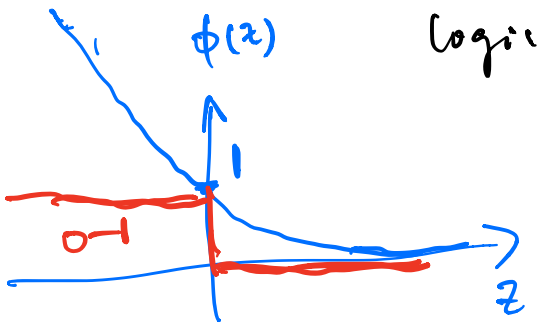
$R = \|w\|_1$: lasso

$R(w) = \frac{1}{2}\|w\|_2^2$. $\phi(f_w(x_i), y_i) = \log_2\left(1 + e^{-y_i f_w(x_i)}\right)$

$f_w(x) = \langle w, x \rangle$. logistic loss.

$\ln(1+x) \approx x$

e^{-z} $z > 0$ $z \le 0$



Logistic regression

Next class:

Regularized loss minimization's solution's generalization performance. from stability perspective