ML AND DATA ANALYSIS	RADEMACHER COMPLEXITY AND NORMS	Oct. 19. 2023

$$R(f) - \hat{R}(f) \leq \sup_{f \in F} (R(f) - \hat{R}(f)) \underset{1-\delta}{\leq} \mathbb{E}_{D_m} \sup_{f \in F} (R(f) - \hat{R}(f)) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

MC DIARMID INEQUALITY

RISK DUE TO THE DATA

WITH HYPOTHESIS:

$f \in F$

F INDP. FROM D_m

D_m is i.i.d.

$\ell \in [0,1]$

$$\leq \mathbb{E}_{D_m} \mathbb{E}_{\underline{\sigma}} \sup_{f \in F} \frac{2}{m} \sum_{i=1}^{m} \sigma_i \ell(f(x_i), y_i) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

vector of i.i.d. VAR. in [0,1] WITH EQUIPROBABILITY, RADEMACHER RANDOM QUANTITY

→ B.C.

$$= 1 - 2 \inf_{f \in F} \frac{1}{m} \sum_{i=1}^{m} \ell(f(\underline{x}_i), y_i)$$

→ KERNEL-RIDGE

$$\propto \|\underline{w}\|^2 = \underline{w}' I \underline{w}$$

HOW MUCH IS THE ABILITY OF MY SPACE OF FUNC TO FIT WHATEVER LABEL I PUT ON MY DATA (A MEASURE ON HOW MANY FUNC. THERE ARE IN MY SPACE).

⇓

THE COMPLEXITY OF MY MODEL SPACE

I DON'T HAVE INFINITE DATASETS SO I CANNOT COMPUTE THE EXPECTATION, SO I WANT TO COMPUTE IT FROM MY DATA:

$$R(f) \underset{1-\delta}{\leq} \hat{R}(f) + \sup_{f \in F} \frac{2}{m} \sum_{i=1}^{m} \sigma_i \ell(f(\underline{x}_i), y_i) + 4\sqrt{\frac{\ln(1/\delta)}{2m}} + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

TO GET THIS RESULT LET'S ANALYSE MORE THE SUPREMA:

$$g((\sigma_1, \underline{x}_1, y_1), \dots, (\sigma_m, \underline{x}_m, y_m)) = \sup_{f \in F} \frac{2}{m} \sum_{i=1}^{m} \sigma_i \ell(f(\underline{x}_i), y_i)$$

★ f^*

$$g'((\sigma_1, \underline{x}_1, y_1), \dots, (\sigma_{j-1}, \underline{x}_{j-1}, y_{j-1})(\sigma_j', \underline{x}_j', y_j')(\sigma_{j+1}, \underline{x}_{j+1}, y_{j+1}) \dots, (\sigma_m, \underline{x}_m, y_m)) =$$

♡

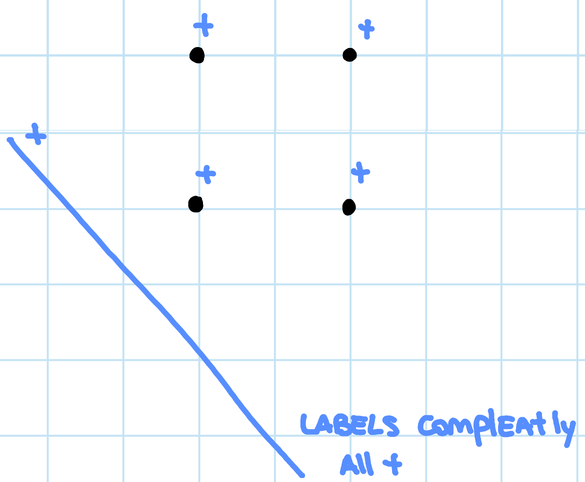
$$= \sup_{f \in F} \left[\frac{2}{m} \sum_{\substack{i=1 \\ i \neq j}}^{m} \sigma_i \ell(f(\underline{x}_i), y_i) + \frac{2}{m} \sigma_j' \ell(f(\underline{x}_j'), y_j') \right]$$

$f^{*'}$

$$|\star - B| \le \begin{cases} \star - B = \star(f^*) - B(f'^*) \le \star(f^*) - B(f^*) = \frac{2}{m}\left[\sigma_j \ell(f^*(x_j), y_j) - \sigma_j' \ell(f^*(x_j'), y_j')\right] \\ B - \star = B(f'^*) - \star(f^*) \le B(f'^*) - \star(f'^*) = \frac{2}{m}\left[\sigma_j' \ell(f'^*(x_j), y_j) - \sigma_j \ell(f'^*(x_j'), y_j')\right] \end{cases}$$

$\le \frac{4}{m}$ (WORST CASE) $\Longrightarrow$ $\frac{\star}{4}$ SATISFIES MCDIARMID INEQUALITY

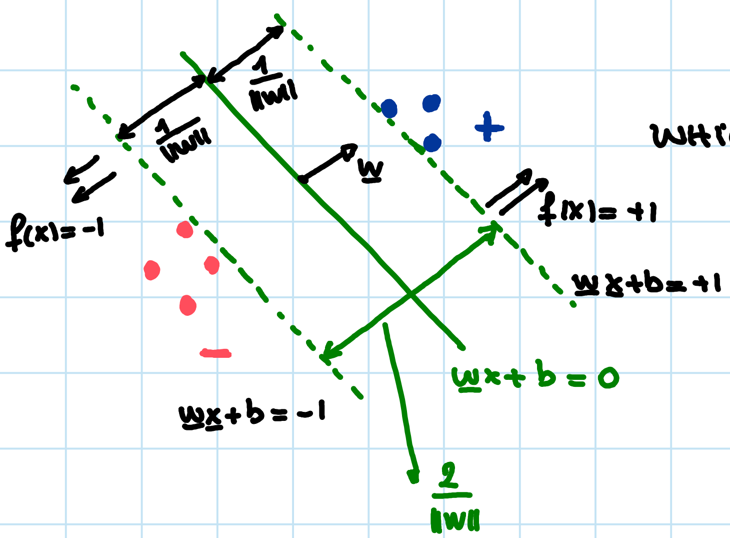
example:



1	2	3	4	
+	+	+	+	●
+	+	+	−	●
+	+	−	+	●
+	+	−	−	●
+	−	+	+	●
+	−	+	−	●
+	−	−	+	×
+	−	−	−	●
−	+	+	+	●
−	+	+	−	×
−	+	−	+	●
−	+	−	−	●
−	−	+	+	●
−	−	+	−	●
−	−	−	+	●
−	−	−	−	●

I'm green if I can classify.

THE RADEMACHER complexity IS ABLE TO MEASURE THE complexity OF A class OF FUNC. THAT CONTAINS INFINITE NUMBER OF FUNC.

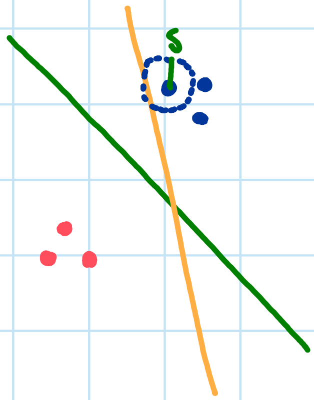


WHICH IS THE BEST LINEAR MODEL?

$$f(x) = \underline{w}\underline{x} + b \longrightarrow \begin{cases} \ge 0 \text{ if } f = +1 \\ \le 0 \text{ if } f = -1 \end{cases}$$

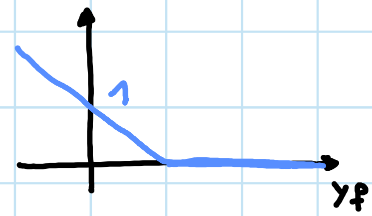
$$\arg\max_{\underline{w}} \frac{2}{\|\underline{w}\|} = \arg\max_{\underline{w}} \frac{1}{\|\underline{w}\|} = \arg\min_{\underline{w}} \|\underline{w}\| = \arg\min_{\underline{w}} \|\underline{w}\|^2$$

MINIMIZING THE $\|w\|$ IN B.C. IS EQUIVALENT TO MAXIMIZING THE MARGIN.



HOW TO SAY THAT ● IS MORE ROBUST OF THE ● ?

$$\ell(f(x), y) = \max[0, 1 - yf]$$



$$\tilde{\ell}(f(x), y) = \max_{\tilde{x} \,:\, \tilde{x} \in B(x)} \ell(f(x), y)$$

$$\|\tilde{x} - x\|^2 < \delta$$

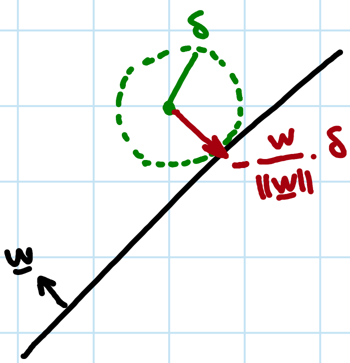
I WANT TO CLASSIFY CORRECTLY THE ENTIRE BALL, NOT JUST THE POINT.

I DON'T CARE ABOUT SEMPLICITY, I ASSUME THERE'S SOMEONE THAT WANTS TO INDUCE MISTAKES ON MY MODEL $\Rightarrow$ I WANT TO MINIMIZE THE RISK.

$$\min_{f \in F} \hat{R}(f) = \min_{w} \frac{1}{m} \sum_{i=1}^{m} \max_{\tilde{x} : \|\tilde{x} - x\|^2 < \delta} \max[0, 1 - y_i w \tilde{x}] =$$

LET'S SOLVE IT IN CLOSE FORM

ADVERSARIAL POINT



$$\min_{w} \sum_{i=1}^{m} \max\left[0, 1 - y_i w \left(x_i - \frac{w}{\|w\|} \delta\right)\right] =$$

$$= \min_{w} \sum_{i=1}^{m} \max[0, 1 - y_i w x_i + y_i \|w\| \delta]$$

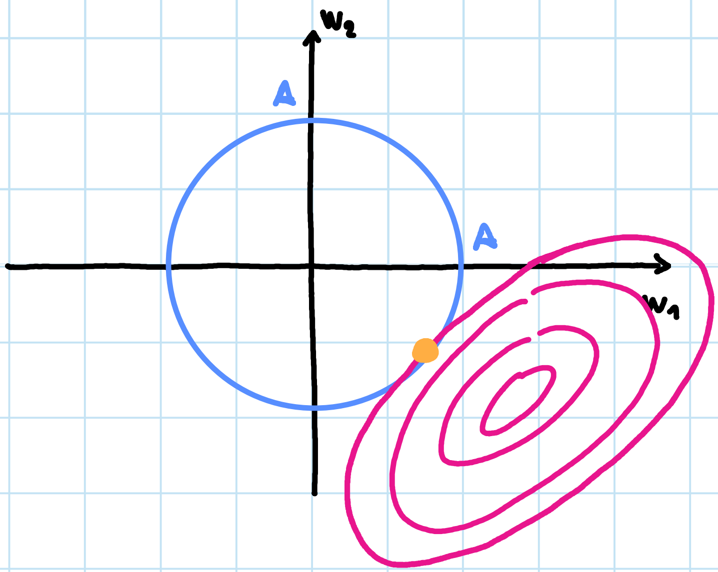
$$\leq \ell(f(x_i), y_i) + c\|w\|$$

OF COURSE I WOULD LIKE TO PUT THE SEPARATOR FURTHER FROM THE POINT, TO MAKE THE MODEL MORE ROBUST, BUT THE FARNESS DEPENDS ON THE MARGIN THAT DEPENDS ON $\|w\|$.

SO $\|\underline{w}\|^2$ IS DIFFERENTIABLE AND CONVEX.

LET'S ANALYSE THIS PROBLEM: $\min_{\underline{w}} \|X\underline{w} - y\|^2 + \lambda\|\underline{w}\|^2$ THAT IS EQUIVALENT TO

$\min_{\underline{w}} \|X\underline{w} - y\|^2$ S.T. $\|\underline{w}\|^2 \leq A$ (LAGRANGE MULTIPLIERS)



I WANT TO FIND THE MINIMUM OF THE PARABOLA, INSIDE THE DOMAIN.

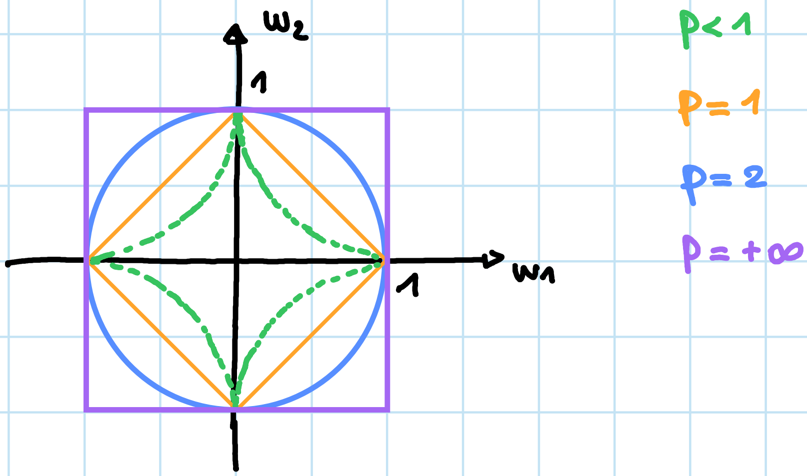
● IS THE SOLUTION

$$\|\underline{w}\|_p^p = \left(\sum_{i=1}^{m} |w_i|^p\right)^{1/p} \quad \text{NORM } p$$

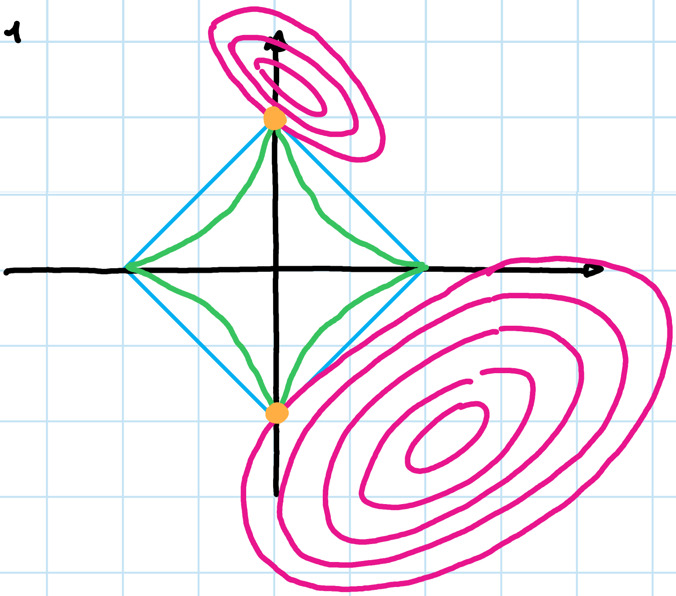
$$\|\underline{w}\|_1 = \sum_{i=1}^{m} |w_i| \quad \text{MANHATTAN NORM}$$

$$\|\underline{w}\|_2^2 = \sqrt{w_1^2 + w_2^2 + \dots + w_n^2} \quad \text{EUCLEADIAN NORM}$$

$$\|\underline{w}\|_\infty = \max(w_1, w_2, \dots, w_n) \quad \text{INFINITE NORM}$$



$p<1$



THE SOLUTION RELIES ON THE VERTICIES!

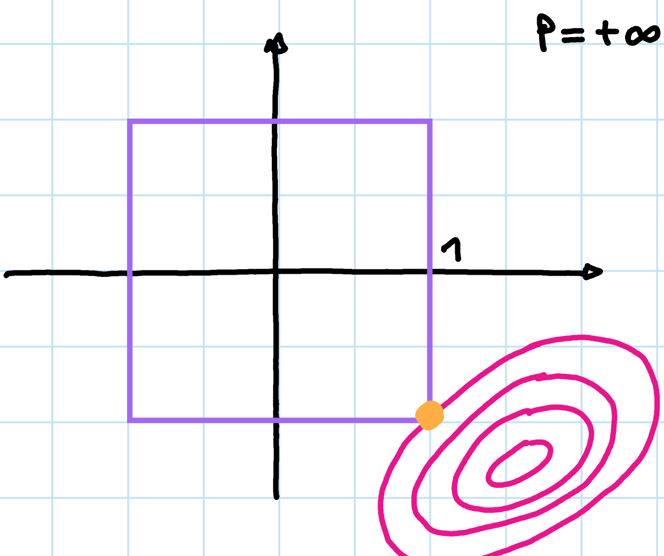
EVEN WITH $p=1$ THE SOLUTION MOST OF THE TIME LIES ON THE VERTICIES $\Rightarrow$ L1 ENFORCES SPARSITY IN ML MODELS

↓

FIRST ALGORITHM OF FEATURE REDUCTION (HOW TO REDUCE # OF FEATURES, ALSO FEATURE SELECTION)

PROBLEM: BY CHANGING FEW FEATURES, I CHANGE THE SOLUTION $\Rightarrow$ ML MODEL IS NOT ROBUST.

IN THE OPPOSITE WAY:



ALL THE ELEMENTS IN MY MODEL WILL BE DIFFERENT FROM 0, THE ML MODEL WILL TRY TO USE ALL THE FEATURES.

CHANGING THE NORM, CHANGES COMPLEATLY THE BEHAVIOUR.

RIDGE REGRESSION: $\|X\underline{w} - \underline{y}\|^2 + \lambda \|\underline{w}\|_2^2$

MSE

L2 NORM

LASSO : $\|X\underline{w} - \underline{y}\|^2 + \lambda \|\underline{w}\|_1$ completely a different behaviour.

L1 NORM