Stability or Robustness of Machine Learning

Focus on regression / classification.

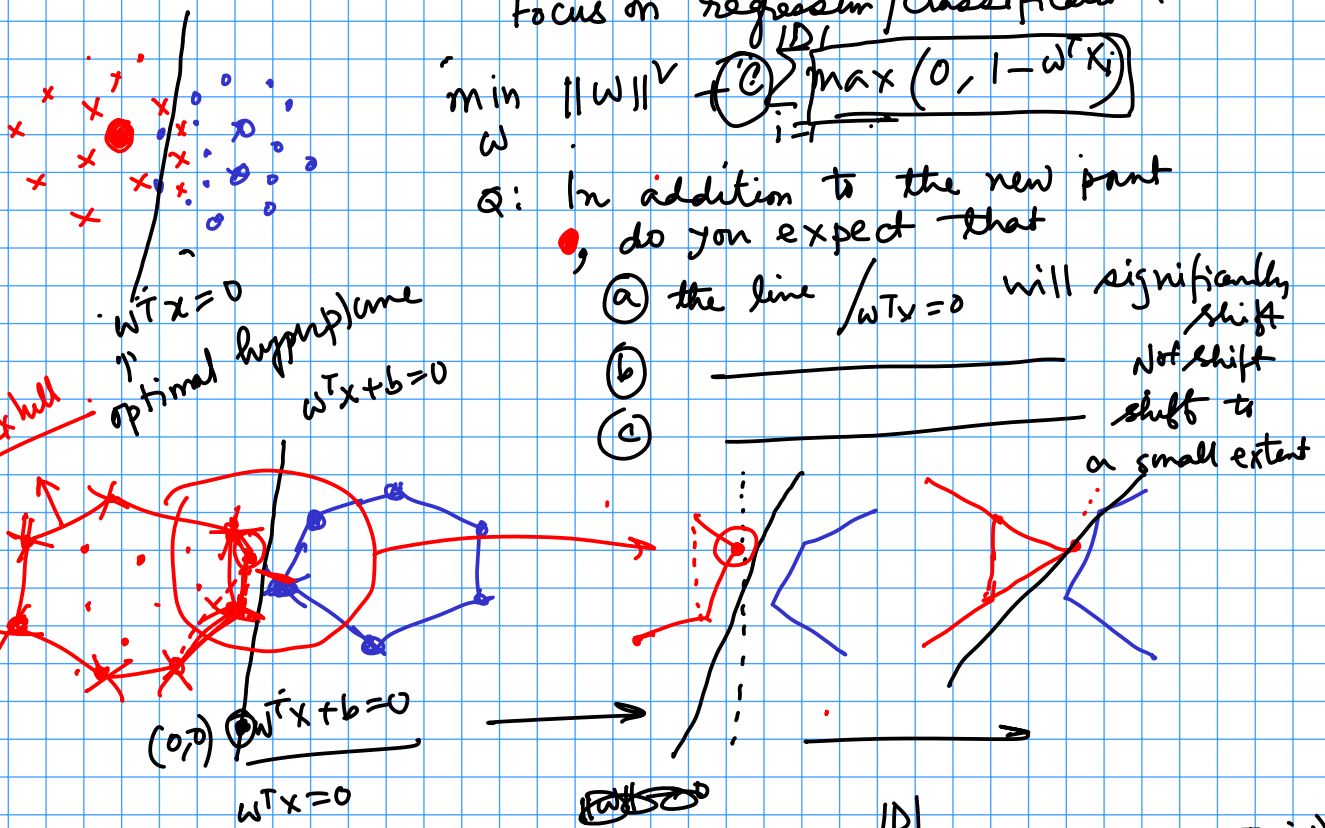
$$\min_w \|w\|^2 + C \sum_{i=1}^{|D|} \max(0, 1 - w^T x_i)$$

Q: In addition to the new point, do you expect that

(a) the line $w^T x = 0$ will significantly shift

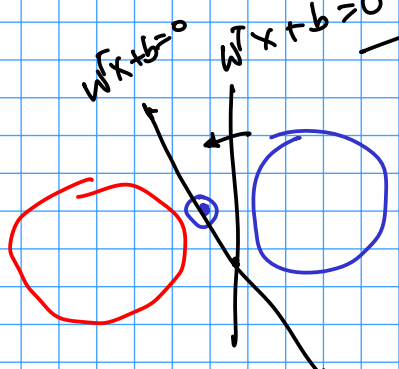
(b) Not shift

(c) shift to a small extent



$$\min_w \left(\|w\|^2 + b^2\right) + C \sum_{i=1}^{|D|} \max\left(0, 1 - y_i\left(w^T x_i + b\right)\right)$$

$$\min_w \|w\|^2 + C \sum_{i=1}^{|D|} \max\left(0, 1 - y_i\left(w^T x_i\right)\right)$$



$\|(w_1 - w_2)\| + |b_1 - b_2|$

$\|b_1 - b_2\|$

Robustness $\Rightarrow$ stability

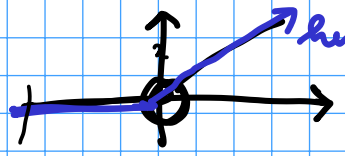
$\rightarrow$ additional point does not change learned pname to a large extent

$\rightarrow$ overfitting $\rightarrow$ model the noise as well.

$\longrightarrow$ Differential privacy will be breached

$A(I) \approx A(I \cup e')$ in prob. sense

* $a_+ = \max(0, a) = \text{ReLU}(a)$
* $a_+ \rightarrow$ hinge loss on a



$$\min_w \; \|w\|^2 + C \sum_{i=1}^{|D|} \left(1 - y_i \cdot w^T x_i\right)_+ = \max(0, 1 - y_i w^T x_i)$$

$\Updownarrow$

$$\min_w \; \lambda \|w\|^2 + \sum_{i=1}^{|D|} \left(1 - y_i w^T x_i\right)_+ \Rightarrow w^*(D, \lambda)$$

optimal

$\lambda = 1/C$

$D = \{(x_i, y_i)\}$

$L_{D,\lambda}(w)$

$L_{Dve}(w) - L_D(w)$

Soham De

$$w^*(D, \lambda) = \underset{w}{\text{argmin}} \; L_D(w)$$

$$w^*(Dve, \lambda) = \underset{w}{\text{argmin}} \; L_{Dve}(w).$$

(x, y)

If the learner/SVM is stable or Robust — what we expect

$$\|w^*(D, \lambda) - w^*(Dve, \lambda)\| \rightarrow$$

$$\lambda \|w^*(D, \lambda)\| \neq \lambda \|w^*(Dve, \lambda)\|^2$$

$$\lambda \|w\|^2 + \sum_{i=1}^{|D|} \ell(i, w)$$

$|D| \rightarrow 1000000$

30000

convex

classification: $\ell(i, w) = \max(0, 1 - y_i \cdot w^T x_i)$

Regression: $\ell(i, w) = (y_i - w^T x_i)^2$

If ℓ is convex

$$\Rightarrow \ell(i, w) = \ell(i, w') + \left(\frac{\partial \ell}{\partial w'}\right)^T (w - w') + (w - w')^T \left[\frac{\partial^2 \ell}{\partial w'^2}\right] (w - w')$$

$$\ell(i, w) \geq \ell(i, w') + \left(\frac{\partial \ell}{\partial w'}\right)^T (w - w')$$

In theory $\min_w \; \lambda \|w\|^2 + \frac{1}{|D|} \sum_{i=1}^{|D|} \ell(i, w)$ $\quad \lambda \simeq \lambda_c |D|$

In practice $\min_w \; \lambda \|w\|^2 + \frac{1}{|D|} \sum_{i=1}^{|D|} \ell(i, w)$

weight

Stability + Generalization by Olivier Bousquet 2002

← Understanding ML 141 – 144.