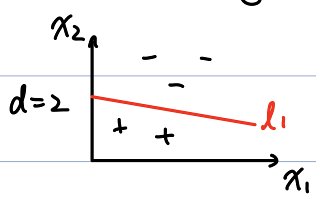
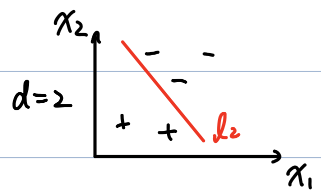
Lec 25 Support Vector Machine (SVM)

1. What we have for now:

↳ Binary linear classification

- Consider linearly seperatable dataset





- Both l_1 and l_2 achieve $E_{in} = 0$. But l_1 is more robust

  ↑ tolerote more noise

2. Classifier Margin:

↳ $\mathcal{D} = \{(\underline{x}_1, y_1) \cdots (\underline{x}_N, y_N)\}$ training set

$\underline{x}_n \in \mathbb{R}^d$, $y \in \{+1, -1\}$

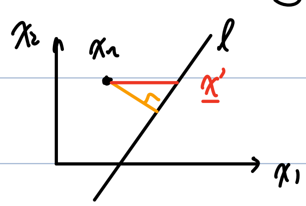
- Suppose the decision rule is:

  $y = \text{sign}(\underline{w}^T \underline{x} + b)$, $\underline{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$.

  // Note we do not use the argumented form

- Let l: $\underline{w}^T \underline{x} + b = 0$ be the decision boundary

- Distance from any $\underline{x}_n$ to l:



want to find distance.

$\because \dfrac{\underline{w}}{\|\underline{w}\|}$ is prependicular to l

$\therefore$ For any point $\underline{w}'$ on l,

$\text{dist}(\underline{x}_n, l) = $ length of projection of $(\underline{x}_n - \underline{x}')$ onto $\dfrac{\underline{w}}{\|\underline{w}\|}$

$$= \frac{|\underline{w}^T(\underline{x}_n - \underline{x}')|}{\|\underline{w}\|}$$

$\because$ $\underline{x}'$ is on the line ℓ : $\underline{w}^T \underline{x}' + b = 0$

$$\therefore \ \text{dist}(\underline{x}_n, \ell) = \frac{|\underline{w}^T \underline{x}_n + b|}{\|\underline{w}\|}$$

$\because$ ℓ correctly classify all points

$\Rightarrow y_n(\underline{w}^T \underline{x}_n + b) > 0 \quad \forall n$

$\because$ Also since $y_n \in \{+1, -1\}$

$\Rightarrow y_n(\underline{w}^T \underline{x}_n + b) = |\underline{w}^T \underline{x}_n + b|$

$$\therefore \ \boxed{\text{dist}(\underline{x}_n, \ell) = \frac{y_n(\underline{w}^T \underline{x}_n + b)}{\|\underline{w}\|}}$$

↳ <u>Margin of ℓ:</u>

$$\rho(\ell) = \min_{1 \le n \le N} \text{dist}(\underline{x}_n, \ell)$$

↳ <u>Goal of SVM:</u>

Find a classifier that maximize the margin.

<u>3. key Normalization:</u>

↳ Let ℓ: $\underline{w}^T \underline{x} + b = 0$ be a decision boundary that correctly classifies each point. i.e. $y_n(\underline{w}^T \underline{x}_n + b) > 0 \quad \forall n$

- Let $\delta = \min_{1 \le n \le N} y_n(\underline{w}^T \underline{x}_n + b) \quad \Rightarrow \delta > 0$

↳ We will always choose $(\underline{w}, b)$ such that $\delta = 1$

Note: this is loss without generality (WLOG)

↳ Suppose $\delta \neq 1$ for some $(\underline{w}, b)$

- Let $\underline{\tilde{w}} = \frac{\underline{w}}{\delta}$, $\hat{b} = \frac{b}{\delta}$,

  $\tilde{\ell} : \underline{\tilde{w}}^T \underline{x} + \hat{b} = 0$, the same boundary as ℓ.

- For $\tilde{\ell}$:

$$\min_{1 \le n \le N} (\underline{\tilde{w}}^T \underline{x} + \hat{b}) = \frac{1}{\delta} \min_{1 \le n \le N} (\underline{w}^T \underline{x} + b) = \frac{\delta}{\delta} = 1$$

- Then the margin:

$$\rho(\ell) = \min_{1 \le n \le N} \frac{y_n(\underline{w}^T \underline{x}_n + b)}{\| \underline{w} \|} = \frac{1}{\| \underline{w} \|}$$

4. Largest Margin Decision Boundary:

↳ $\max \frac{1}{\| \underline{w} \|}$ s.t. $\min_{1 \le n \le N} y_n(\underline{w}^T \underline{x}_n + b) = 1$