FOUNDATIONS OF MACHINE LEARNING [FOML]

Mohri et al.

[select portions]

$$\mathcal{D} = \left\{ (x^{(i)}, y^{(i)}) \;\middle|\; \begin{array}{l} x^{(i)} \in \mathcal{X},\ y^{(i)} \in \{-1, 1\} \\ i \in [N] \end{array} \right\}$$

$$\mathcal{D} \sim P^{(N)}$$

Assume that there exists w^* such that

$$\text{sign}\left((w^*)^T x^{(i)}\right) = y^{(i)} .$$

$i \in [N]$.

$w^{(1)} = 0$

$\|x^{(i)}\| \leq R$

$i = 1, \ldots, N$

$$\Rightarrow y^{(i)} \left(w^{*T} x^{(i)}\right) > 0$$

$$\gamma = \min_i \frac{y^{(i)} (w^*)^T x^{(i)}}{\|w^*\|} .$$

$w^{(k)}$ be the current estimate.

Let $(x^{(l)}, y^{(l)}) \in \mathcal{D}$ such that

$$\operatorname{sign}\left(\left(w^{(k)}\right)^T x^{(l)} \right) \neq y^{(l)}$$

$$\Rightarrow \quad y^{(l)}\left(w^{(k)T} x^{(l)} \right) < 0$$

$$w^{(k+1)} = \begin{cases} w^{(k)} + y^{(l)} x^{(l)} & [\text{Update}] \\ w^{(k)} & \text{otherwise} \end{cases}$$

On a linearly separable dataset the perceptron algorithm terminates after making at most $\frac{R^2}{\gamma^2}$ updates

$$\frac{R^2}{\gamma^2} = \frac{R^2 \|w^*\|^2}{\min_i \{y_i (w^{*T} x^{(i)})\}^2}$$

$\mathcal{D}$ be a Sample of size N.

Linearly separable [There exists $\gamma > 0$].

Let Perceptron Algorithm return a Classifier $h_{\mathcal{D}}^{(P)}$.

$$R(h_{\mathcal{D}}^{(P)}) = \underset{X,Y \sim \mathcal{P}}{P}(h_{\mathcal{D}}^{(P)}(X) \neq Y)$$

An algorithm A acting on a sample $\mathcal{D}$ of size m return a classifier $h_{\mathcal{D}}^{(A)}$.

The leave one out error of A on $\mathcal{D}$ is

$$\bar{R}_{\mathcal{D}}^{LOO}(A) = \frac{1}{m} \sum_{i=1}^{m} 1_{\{h_{\mathcal{D}^{(i)}}^{(A)}(x^{(i)}) \neq y^{(i)}\}}$$

$$\mathcal{D}^{(i)} : \mathcal{D} \setminus \{(x^{(i)}, y^{(i)})\}$$

$$E_{\mathcal{D} \sim p(m)} \bar{R}_{\mathcal{D}}^{LOO}(A)$$

Expected LOO error on a random sample of size m.

$$\mathbb{E}_{\mathcal{D}\sim p(m)} \bar{R}^{LOO}_{\mathcal{D}}(A)$$

$$= \mathbb{E}_{\mathcal{D}\sim p(m-1)} R\left(h^{(A)}_{\mathcal{D}}\right)$$

[Reading: Sec 5.2.4.
Lemma 5.3.
FOML]

Perceptron

$$\frac{\mathbb{E}_{\mathcal{D} \sim p(N)} R\left(h_{\mathcal{D}}^{(P)}\right) = \mathbb{E}_{\mathcal{D} \sim p(N+1)} \bar{R}_{\mathcal{D}}^{\text{LOO}}(P)}{}$$

$$\mathbb{E}_{\mathcal{D} \sim p(N)} R\left(h_{\mathcal{D}}^{(P)}\right)$$

$$\leq \mathbb{E}_{\mathcal{D} \sim p(N+1)} \frac{\min\left(M(\mathcal{D}), \frac{R^2(D)}{\gamma^2(\mathcal{D})}\right)}{N+1}$$

[Theorem 8.9 FOML]

Generalization error of SVM

$$\min_{\{w,b\}} \quad \frac{1}{2}\|w\|^2 \qquad w^* = \sum_{i=1}^{N} \lambda_i y_i x^{(i)}$$

$$y_i(w^T x^{(i)} + b) \geq 1 \qquad \lambda_i \geq 0$$

$$i = 1, \ldots, N \qquad SV = \{i \mid \lambda_i > 0\}$$

Let $b = 0$

Optimum attained at w^*

$$\min_{\{w,b\}} \quad \frac{1}{2}\|w\|^2$$

$$y_i(w^T x^{(i)}) \geq 1 \qquad i = 1, \ldots N$$

$$\Rightarrow \quad \max_{w,\gamma} \quad \gamma$$

$$y_i(w^T x^{(i)}) \geq \gamma \|w\| \qquad i \in [N]$$

optimum (γ^*, w^*)

$$\gamma^* = \frac{1}{\|w^*\|}$$

$$h_{\theta}^{SVM}(x) = \text{sign}(w^{*T} x)$$

$$R(h_{\theta}^{SVM}) = P(\text{sign}(w^{*T} X) \neq Y)$$

$$(X, Y) \sim P$$

[Theorem 5.4 FOML]

$$\mathbb{E}_{\mathcal{D} \sim P^{(N)}} R(h_{\mathcal{D}}^{SVM}) \leq \mathbb{E}_{\mathcal{D} \sim P^{(N+1)}} \frac{|SV(\mathcal{D})|}{N+1}$$

$|SV(\mathcal{D})|$ = Number of Support vectors

Proof: Take $\mathcal{D} \sim P^{(N+1)}$ to be separable. solve SVM.

Argue $\bar{R}^{LOO}_{\mathcal{D}}(SVM) \leq \frac{|SV(\mathcal{D})|}{N+1}$

weak: Average generalization error.

Relate Training set error to generalization error.

If Training set error is 0 then is the generalization error 0?

$a \leq Z \leq b$ $\quad Z_i$, iid P $\quad E(Z) = \mu$

$$P\left(\left| \frac{1}{n}\sum_{i=1}^{n} z_i - \mu \right| > \epsilon \right) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

$$\therefore \frac{1}{n}\sum_{i=1}^{n} z_i - \Delta \leq \mu \leq \frac{1}{n}\sum_{i=1}^{n} z_i + \Delta$$

with prob $1-\delta$. $\qquad 2e^{-\frac{2n\Delta^2}{(b-a)^2}} \leq \delta$

A VC dimension approach

$$R(h) \leq R_{emp}(h) + \sqrt{\frac{V}{N}\left(\log N + \log \frac{1}{\delta}\right)}$$

with prob $1-\delta$.

$$R_{emp}(h) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}$$

Training set error

$V \rightarrow$ VC Dimension

[Reading: Burges Tutorial]

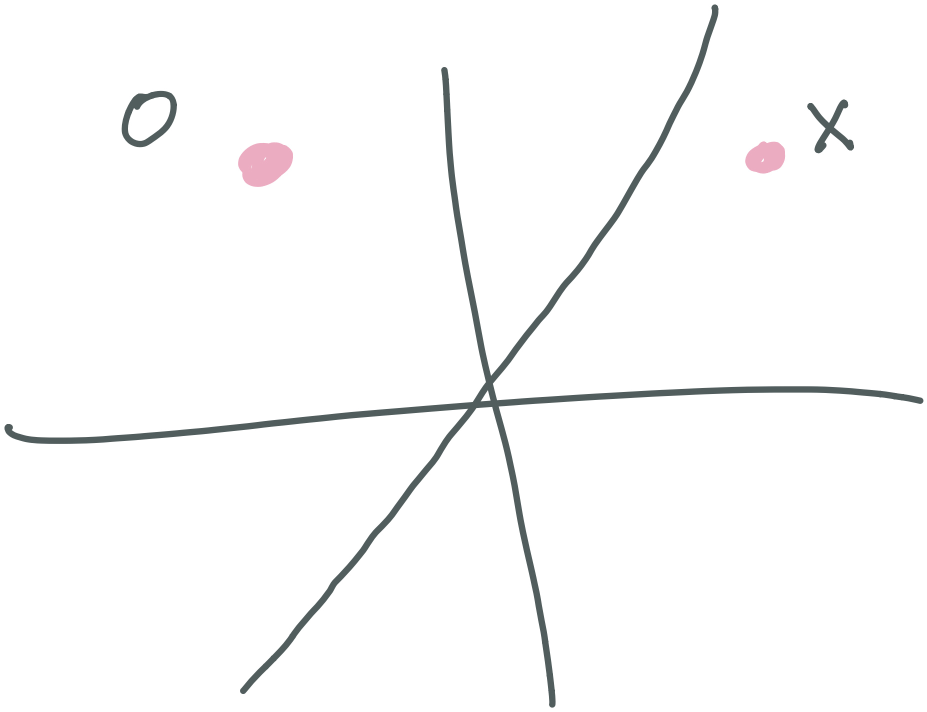
$H = \{h \mid h \text{ is a classifier}\}$

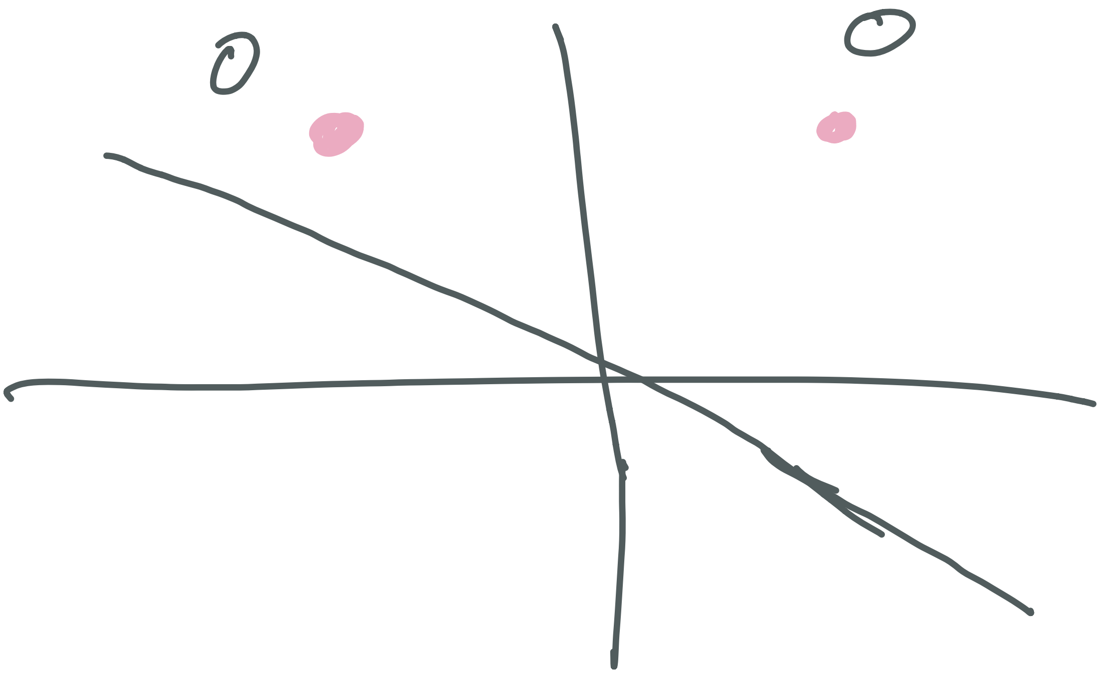
$V(H)$ = maximum number of points on which classifiers from H can learn all possible Labellings.

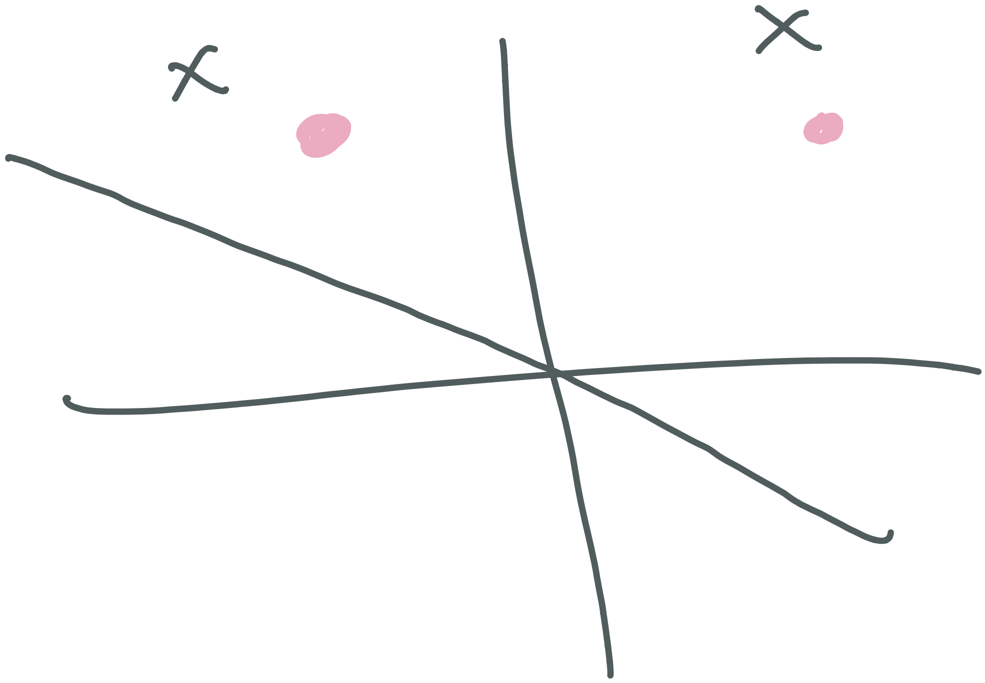
$H = \{h \mid h(x) = \text{sign}(w^T x),\ w \in \mathbb{R}^2\}$

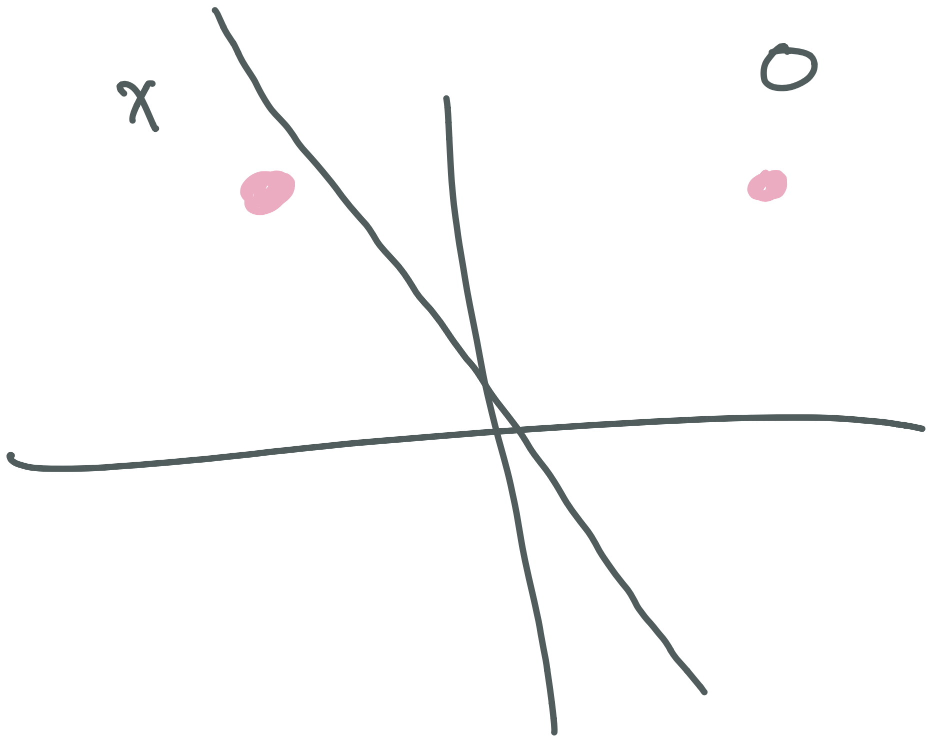
$x \in \mathbb{R}^2$

$\text{sign}(w^T x)$









Will also work for 3 points

But not for 4 points

If $\|w\| \leq B$. $\quad \|x\| \leq R \quad x \in \mathbb{R}^d$

$$H = \{h \mid h(x) = \text{sign}(w^T x + b)\}$$

$$V(H) \leq B^2 R^2$$

$$R(h) \leq R_{emp}(h) + O\left(\frac{RB}{\sqrt{n}}\right)$$

with prob $1-\delta$.

$$\min_{\|w\| \leq B} R_{emp}(h)$$

$$h(x) = \text{sign}(w^T x + b)$$

Motivates the following formulation

$$\min_{w,b} \quad C \cdot \sum_{i=1}^{N} \max(0, 1 - y_i (w^T x_i + b)) + \frac{1}{2} \|w\|^2$$