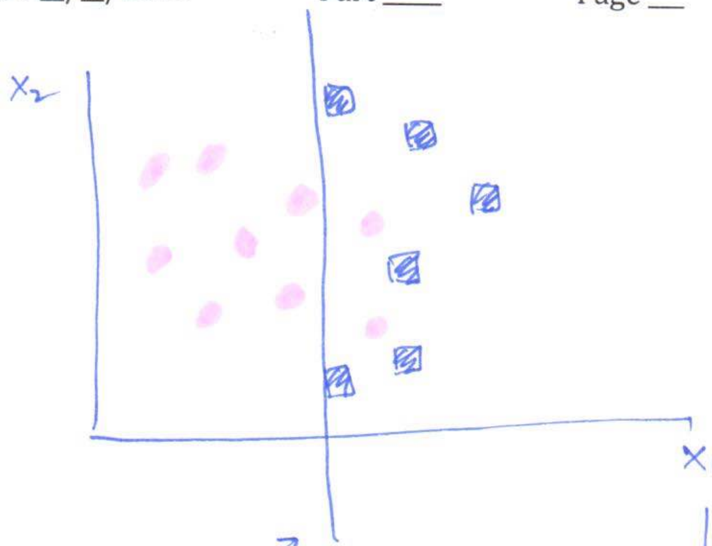
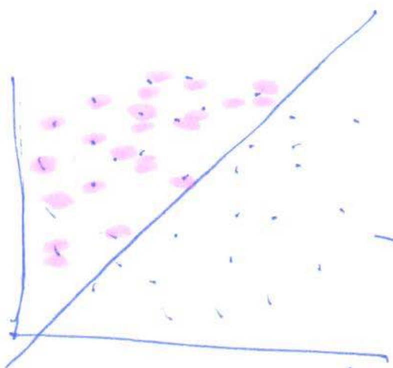


$f(\ )$

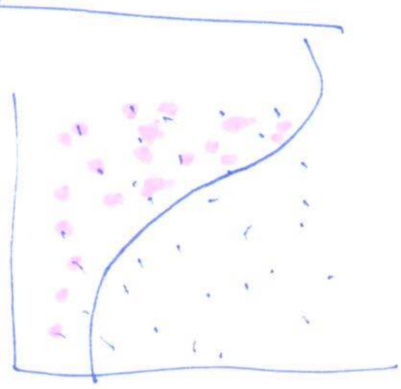
Inductive Bias.

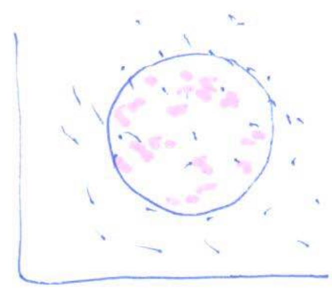
$f \rightarrow$ linear

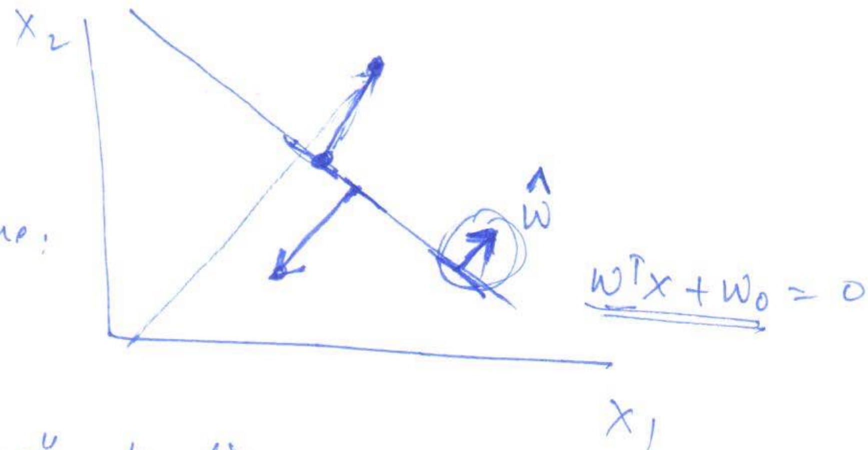
Linear Boundary
Discriminating Surface



$\rightarrow$ **Linearly Separable**







For a point on the line:

$w^T x + w_0 = 0$

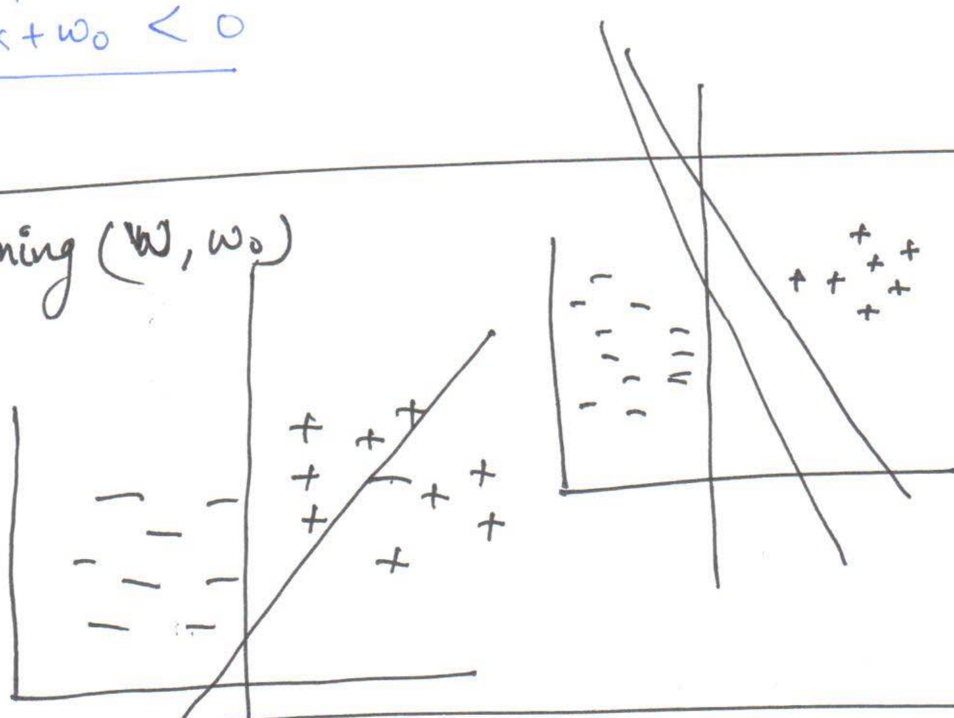
For a point "above" the line:

$w^T x + w_0 > 0$

For a point "below" the line:

$w^T x + w_0 < 0$

Learning (w, w_0)



Generalization (True) Risk vs. Empirical Risk.

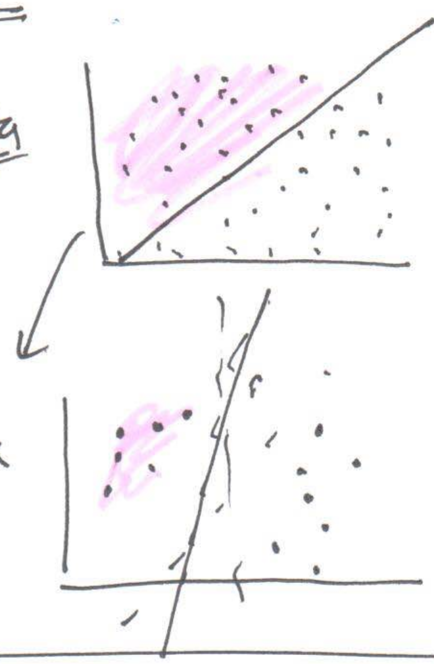
loss on future (unseen) data → True Risk

We do not have data to measure true risk.

We can measure risk on training data

Empirical risk.

All data



Overfitting ← Train data

Train data

Assume $\vec{x_i}$ doesn't have bias term

$$w, w_0$$

x_1	y_1	$\text{sign}(w^T x_1 + w_0)$
x_2	y_2	$\text{sign}(w^T x_2 + w_0)$
⋮	⋮	⋮
x_n	y_n	$\text{sign}(w^T x_n + w_0)$

---

x_i $\quad$ y_i $\quad$ If $(w^T x_i + w_0) \geq 0$ $\quad \tilde{y}_i = +1$

If $(w^T x_i + w_0) < 0$ $\quad \tilde{y}_i = -1$

If $y_i = +1$ and $\tilde{y}_i = +1$ ⇒ No mistake

$y_i = -1$ and $\tilde{y}_i = -1$ ⇒ no mistake

$y_i = \pm 1$ and $\tilde{y}_i = -1$ ⇒ Mistake
$\qquad\qquad\qquad\quad +1$ ⇒ Mistake

If $y_i \tilde{y}_i > 0 \Rightarrow$ No mistake

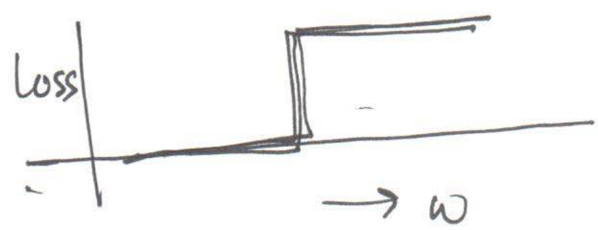
$y_i \tilde{y}_i < 0 \Rightarrow$ Mistake.

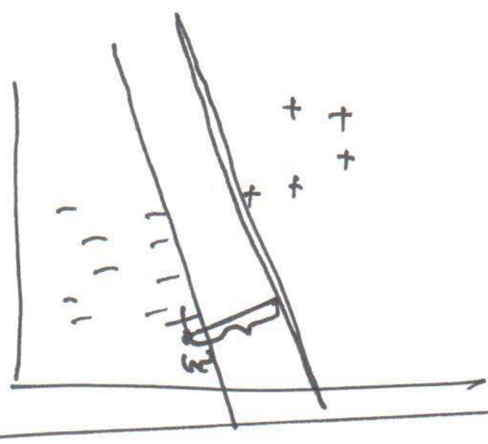
$y_i(\underline{w^T x_i + w_0}) > 0 \Rightarrow$ No mistake

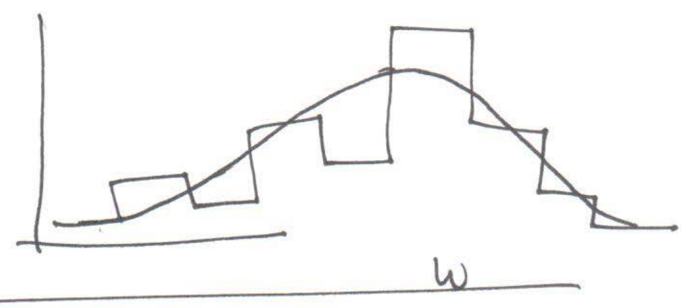
$y_i(w^T x_i + w_0) < 0 \Rightarrow$ Mistake.

$$\boxed{\sum_{i=1}^{n} \mathbb{I}\left(\underline{y_i(w^T x_i + w_0)} < 0\right)} \longrightarrow \text{0-1 loss}$$

$\boxed{\text{Find } \underline{w, w_0} \text{ that minimizes 0-1 loss}}$





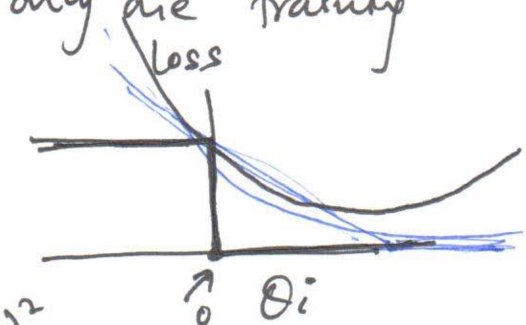


let us assume that we have only one training data point. x_i $y_i = +1$

$\underbrace{(w^T x_i + w_0)}_{\theta_i}$ If $\underline{\theta_i > 0}$

$(y_i - w^T x_i)^2 = (1-\theta_i)^2$

$(1 + \exp(-y_i w^T x_i))$

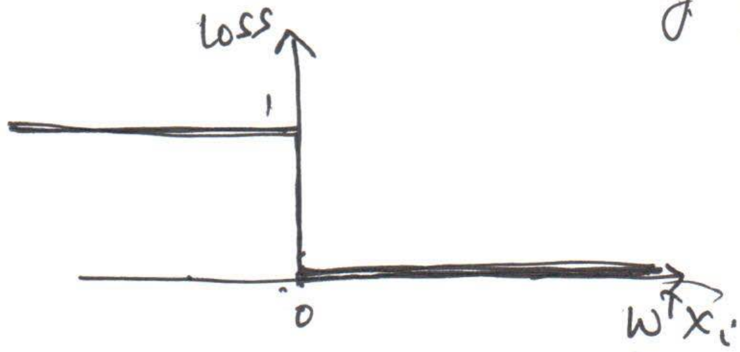


let us assume that we have 1 training point.

let $y_i = +1$

$\underline{w^T x_i}$

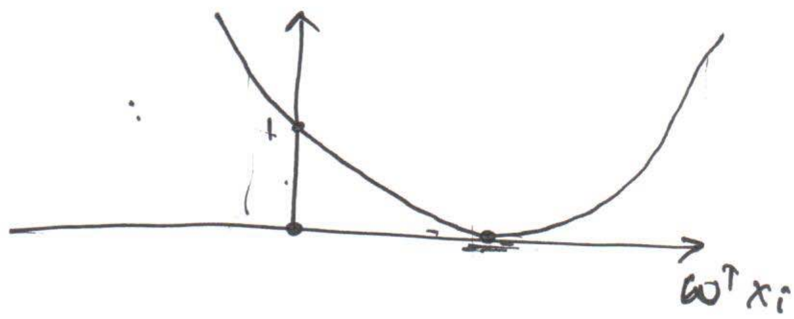
Contains the bias.

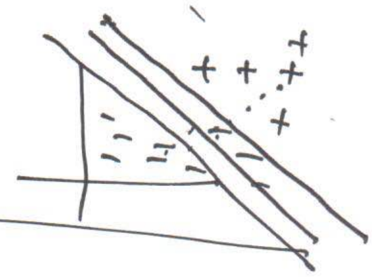


$(y_i - w^T x_i)^2$

SL — $(1 - \underline{w^T x_i})^2$

Perceptron

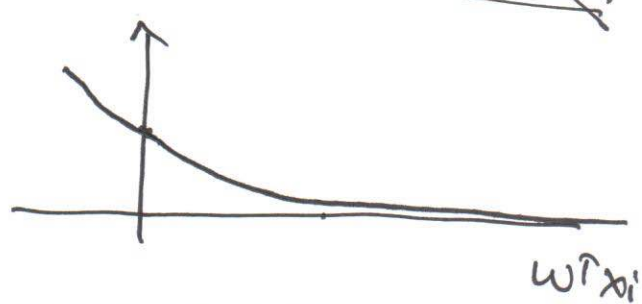




$\log(1 + \exp(-y_i w^T x_i))$

Log Loss. $\underline{\log(1 + \exp(-w^T x_i))}$

logistic Regression

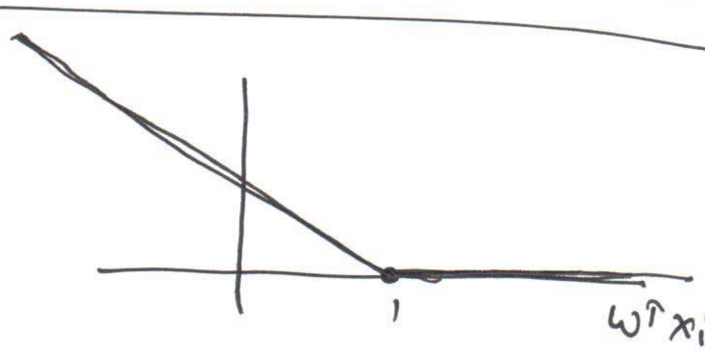


$\max(0, 1 - y_i w^T x_i)$

$\max(0, \underline{1 - w^T x_i})$

Hinge loss.

Support Vector Machine



Sigmoid. $\sigma(a) = \dfrac{1}{1+\exp(-a)} = \dfrac{1}{1+e^{-a}}$

$$\frac{d}{da}\sigma(a) = \frac{-1}{(1+e^{-a})^2} \cdot -e^{-a} = \frac{e^{-a}}{(1+e^{-a})^2}$$

$$= \frac{e^{-a}}{(1+e^{-a})} \cdot \frac{1}{(1+e^{-a})}$$

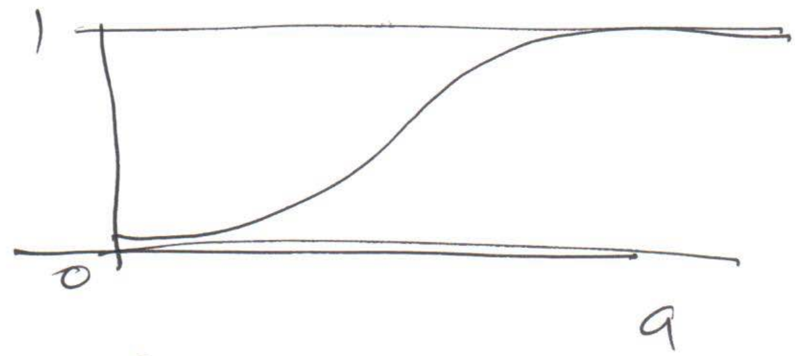
$$= \left(1 - \frac{1}{1+e^{-a}}\right)\left(\frac{1}{1+e^{-a}}\right)$$

$$= \sigma(a)\left(1 - \sigma(a)\right)$$

---

x_i

$$\sigma(\underline{w^T x_i}) = \frac{1}{1+\exp(-w^T x_i)}$$



$$P(y_i = +1) = \sigma(w^T x_i)$$

$$P(y_i = -1) = 1 - \sigma(w^T x_i)$$

---

x_1	y_1	p_1
x_2	y_2	p_2
x_3	y_3	p_3
$\vdots$	$\vdots$	$\vdots$
x_n	y_n	p_n

$$p_i = \frac{1}{1+\exp(-y_i w^T x_i)}$$

$$\cancel{\prod p_i} \cdot p_1 \times p_2 \times \cdots p_n$$

$$J(w) = -\sum \log \underline{p_i}$$

$$= \sum \log\left[1 + \exp(-y_i w^T x_i)\right]$$

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-w^T x_i)\right)$$

$$\nabla J(w) = \frac{1}{n} \sum_{i=1}^{\cancel{n}} \frac{d}{dw} \log\left(1 + \exp(-w^T x_i)\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\left(1 + \exp(-w^T x_i)\right)} \exp(-w^T x_i)\,(-x_i)$$

$$= \cancel{x_i} - \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\exp(-w^T x_i)}{1 + \exp(-w^T x_i)}\right] x_i$$

Set $\nabla J(w) = 0$

and solve for w 

Gradient descent.

$w \leftarrow w_{init}$.

until converged:

$$w \leftarrow w - \eta \nabla J(w)$$

too sensitive to η

Newton's Method.

$w \leftarrow w_{init}$

until converged

$$w \leftarrow w - \underline{\eta}\, \underline{H^{-1}}\, \nabla J(w)$$

Hessian Matrix