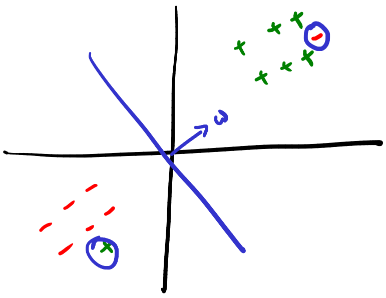
Perceptron mistakes $\leq \dfrac{R^2}{\gamma^2}$ $\leftarrow$ Radius, $\leftarrow$ Margin



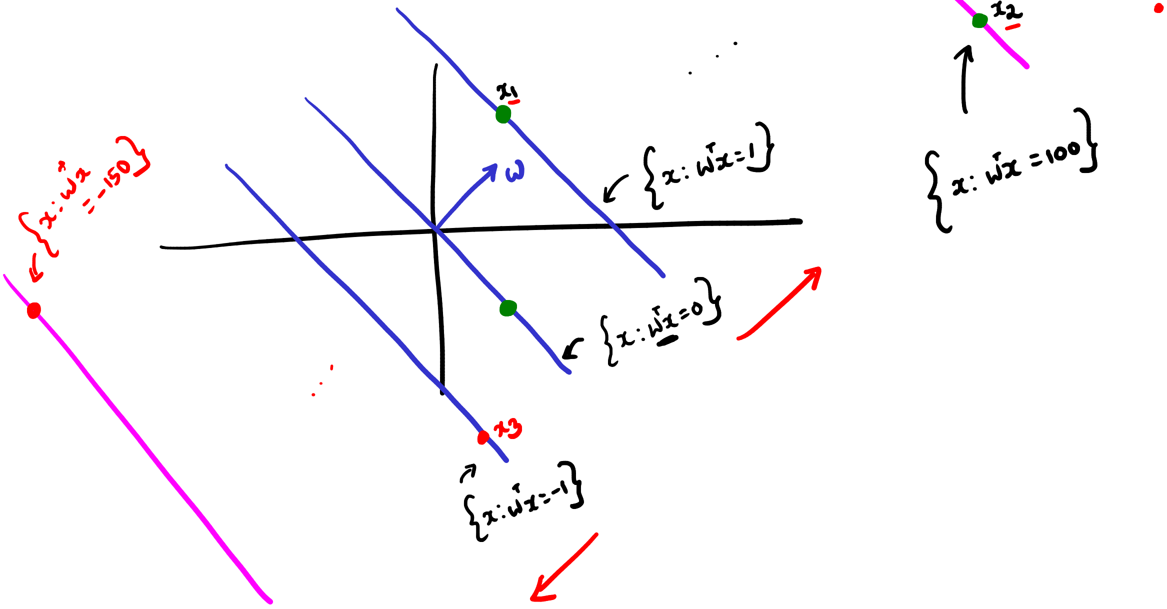
$$P(y=1 \mid x) = \begin{cases} 1 & \text{if } w^T x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Can we model probabilities differently?

- Start with a simple LINEAR model

$z = w^T x$ $\leftarrow$ feature $\quad x \in \mathbb{R}^d$

$z \in [-\infty, \infty]$




- Larger the score $(z = w^T x)$, more the probability of being +1

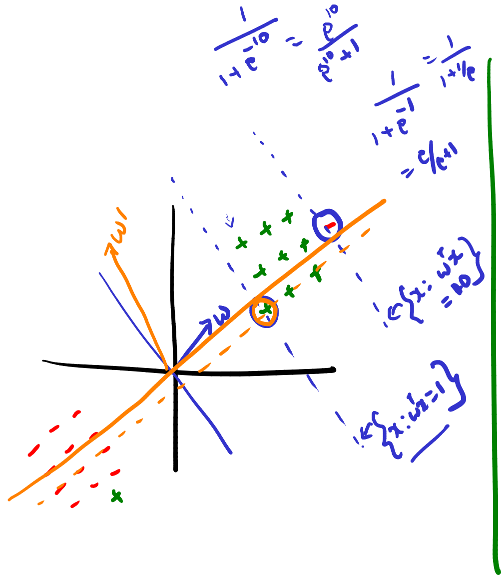
Score

$g(z) = 0.5$ if $z = 0$

LINK FUNCTION

$g(z) \to 1$ as $z \to \infty$

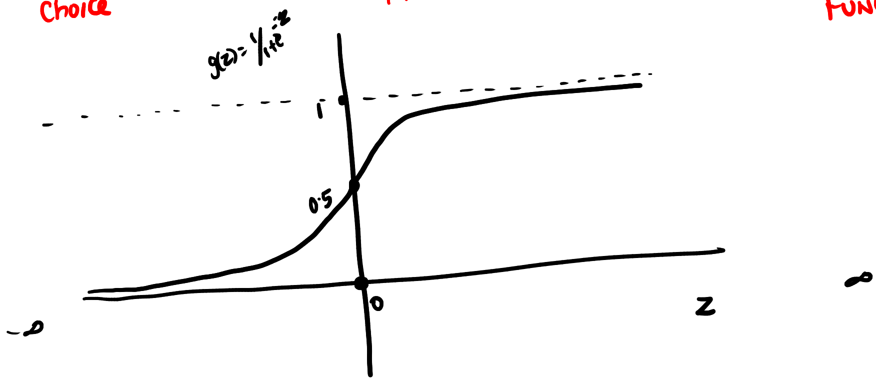
$g(z) \to 0$ as $z \to -\infty$




one popular choice

$$g(z) = \frac{1}{1+e^{-z}}$$

SIGMOID / LOGISTIC FUNCTION




---

MODEL: LOGISTIC REGRESSION

$$P(y=1/x) = \frac{1}{1+e^{-w^T x}} = g(w^T x)$$

Dataset: $\{(x_1,y_1) \ldots, (x_n,y_n)\}$ $\quad y_i \in \{0,1\}$.

How to find w: Maximum Likelihood.

$$\mathcal{L}(w; \text{Data}) = \prod_{i=1}^{n} \left(g(w^T x_i)\right)^{y_i} \left(1 - g(w^T x_i)\right)^{(1-y_i)}$$

$$\log \mathcal{L}(w; \text{Data}) = \sum_{i=1}^{n} y_i \log\left(g(w^T x_i)\right) + (1-y_i)\log\left(1 - g(w^T x_i)\right)$$

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

$$= \sum_{i=1}^{n} \underbrace{\left[(1-y_i)(-w^T x_i) - \log\left(1+e^{-w^T x_i}\right)\right]}$$

Goal:

$$\boxed{\max_{w} \underbrace{\sum_{i=1}^{n}\left[(1-y_i)(-w^T x_i) - \log\left(1+e^{-w^T x_i}\right)\right]}_{\log L(w)}}$$

- No closed form expression
- Can perform Gradient descent. [ascent]

$$\nabla \log L(w) = \sum_{i=1}^{n} \left[ (1-y_i)(-x_i) - \left( \frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \right)(-x_i) \right]$$

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

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

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

Gradient update rule

step-size

$$w_{t+1} = w_t + \eta_t \nabla \log L(w_t)$$

$$= w_t + \eta_t \left( \sum_{i=1}^{n} \overbrace{x_i}^{\in \mathbb{R}^d} \overbrace{\Big( \underbrace{y_i}_{\{0,1\}} - \underbrace{\frac{1}{1+e^{-w_t^T x_i}}}_{g(w_t^T x_i)} \Big)}^{\theta_i} \right)$$

$x_{test} \in \mathbb{R}^d$

$$\boxed{\hat{y}_{test} = \text{sign}\left( \hat{w}^T x_{test} \right)}$$

## KERNEL VERSION

- Can argue $w^* = \sum_{i=1}^{n} \alpha_i x_i$ [Formal theorem is called the Representor theorem].

## REGULARIZED VERSION

CROSS VALIDATE

Hyper parameter.

$$\min_{w} \sum_{i=1}^{n} \left[ \log\left(1 + e^{-w^T x_i}\right) + w^T x_i (1-y_i) \right] + \underbrace{\frac{\lambda}{2} \|w\|^2}_{\text{Regularizer}}$$