

Classification



Thus far, regression:

predict a continuous value given some inputs

Given $x \in \mathbb{R}^d$, predict $y = f(x)$

$y \in \mathbb{R}$

continuous, 1, 0.1, π --

Reading Your Brain, Simple Example

$$f(x_1) = \text{"Person"}$$
$$f(x_2) = \text{"Animal"}$$

Encoding:

[Mitchell et al.] let $\begin{cases} \text{"Person"} = 1 \\ \text{"Animal"} = 0 \end{cases}$

Pairwise classification accuracy: 85%

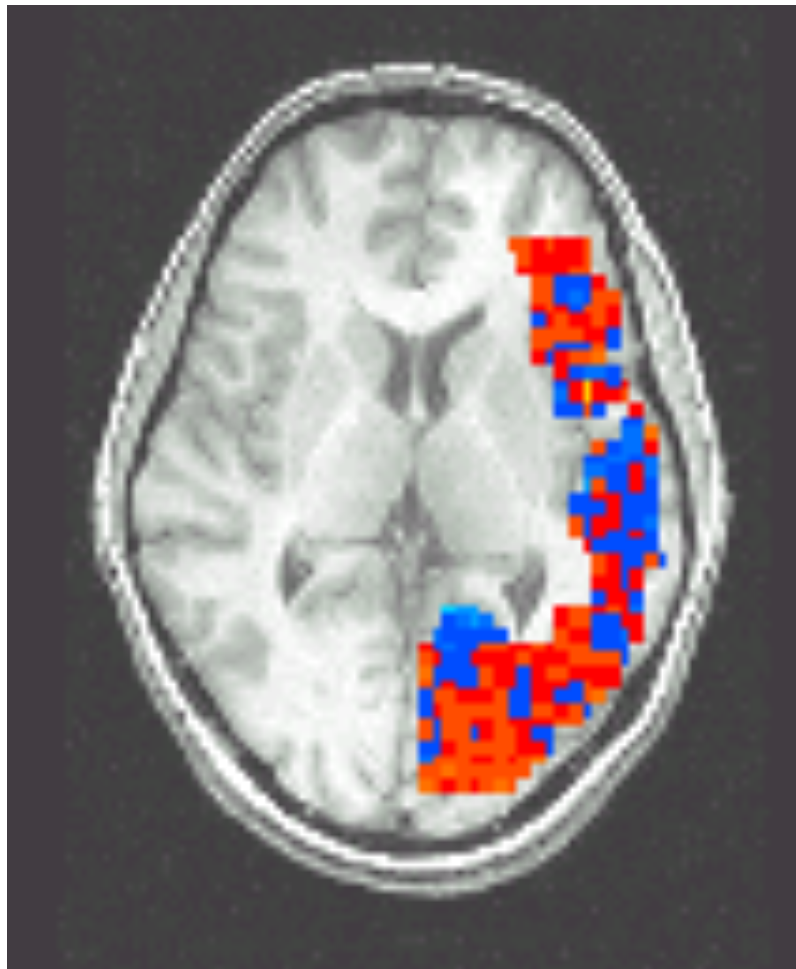
$$f(x_1) = 1$$

$$f(x_2) = 0$$

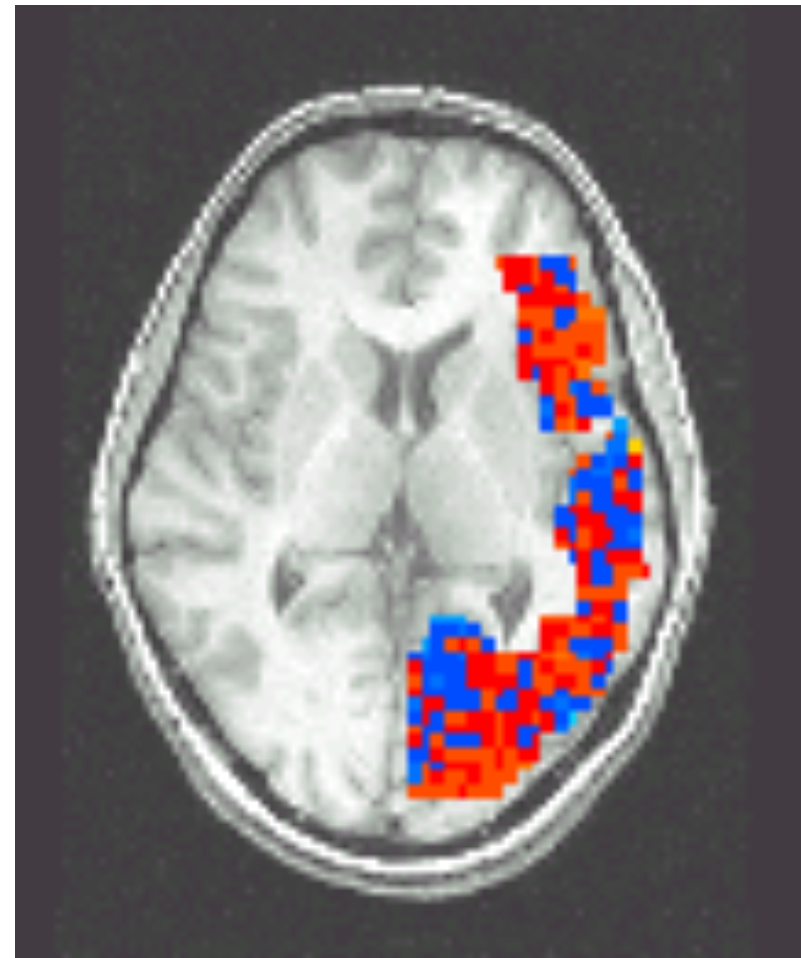
Person



Animal



x_1



x_2

Classification

- Learn $f: X \rightarrow Y$

- X - features

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

- Y - target classes

$$\{0, 1\}$$

binary

$$\{1, \dots, K\}$$

multi-class

- Loss Function

$$l(f(x), y) = \mathbb{1}\{f(x) \neq y\}$$

1 wrong
0 correct

- Expected loss of f :

Performance measure

$$\mathbb{E}_{X,Y} [l(f(x), y)] = \mathbb{E}_{X,Y} [\mathbb{1}\{f(x) \neq y\}] = \mathbb{E}_X [\mathbb{E}_{Y|X} [\mathbb{1}\{f(x) \neq y\} | X=x]]$$

joint distribution
over X, Y

$$\mathbb{E}_{Y|X} [\mathbb{1}\{f(x) \neq y\} | X=x] = \sum_{i=1}^K P(Y=i | X=x) \cdot \mathbb{1}\{f(x) \neq i\}$$

$$= \sum_{i \neq f(x)} P(Y=i | X=x)$$

$$= 1 - P(Y=f(x) | X=x)$$

$$\sum_{i=1}^K P(Y=i | X=x) = 1$$

Classification

- Learn $f: X \rightarrow Y$
 - X - features
 - Y - target classes

- **Loss Function** $\ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}$

- **Expected loss of f :**

$$\mathbb{E}_{XY}[\mathbf{1}\{f(X) \neq Y\}] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x]]$$

$$\begin{aligned}\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x] &= \sum_i P(Y = i|X = x)\mathbf{1}\{f(x) \neq i\} = \sum_{i \neq f(x)} P(Y = i|X = x) \\ &= 1 - P(Y = f(x)|X = x)\end{aligned}$$

- **Suppose you knew $P(Y|X)$ exactly, how should you classify?**

Classification

- Learn $f: X \rightarrow Y$
 - X - features
 - Y - target classes

- **Loss Function** $\ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}$

- **Expected loss of f :**

$$\mathbb{E}_{XY}[\mathbf{1}\{f(X) \neq Y\}] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x]]$$

min
f

$$\mathbb{E}_{Y|X}[\mathbf{1}\{f(x) \neq Y\}|X = x] = \sum_i P(Y = i|X = x)\mathbf{1}\{f(x) \neq i\} = \sum_{i \neq f(x)} P(Y = i|X = x) = 1 - P(Y = f(x)|X = x)$$

- **Suppose you knew $P(Y|X)$ exactly, how should you classify?**
 - **Bayes-Optimal classifier:**

$$f(x) = \arg \max_y \mathbb{P}(Y = y|X = x)$$

Bayes Optimal Binary Classifier

$$Y \in \{0, 1\}$$

- Suppose you knew $P(Y|X)$ exactly, how should you classify?

- Bayes-Optimal classifier:

$$f(x) = \arg \max_y \mathbb{P}(Y = y | X = x)$$

- Suppose we don't know $P(Y|X)$, but have n iid examples

$$\{(x_i, y_i)\}_{i=1}^n$$

- What is a natural estimator for $P(Y | X)$?

Bayes Optimal Binary Classifier

- Suppose we don't know $P(Y|X)$, but have n iid examples

$$\{(x_i, y_i)\}_{i=1}^n$$

$$Y \in \{0, 1\}$$

- What is a natural estimator for $P(Y | X)$?

Fix some $\tilde{x} \in X$

$$\tilde{D} = \{(\tilde{x}, \tilde{y}_1), \dots, (\tilde{x}, \tilde{y}_m)\}$$

Suppose $x_i = \tilde{x}$ for $m \leq n$ samples

What is a natural estimator for $\theta_* := \mathbb{P}(Y = 1 | X = \tilde{x})$?

If k of the m labels are equal to $Y = 1$ then $\frac{k}{m}$

$$\hat{p}(Y=1 | X=\tilde{x}) = \frac{k}{m}$$

unbiased

$$\mathbb{E}_D [\hat{p}(Y=1 | X=\tilde{x})] = p(Y=1 | X=\tilde{x})$$

$\hat{p} \rightarrow p, m \rightarrow \infty$

Bayes Optimal Binary Classifier

- Suppose we don't know $P(Y|X)$, but have n iid examples

$$\{(x_i, y_i)\}_{i=1}^n \quad Y \in \{0, 1\}$$

- What is a natural estimator for $\operatorname{argmax}_y P(Y = y | X)$?

If $X = \{0, 1\}^d$, or is generally discrete, continuous case

$$\hat{f}(x) = \operatorname{argmax}_{y \in \{0, 1\}} \frac{\sum_{i=1}^n \mathbf{1}[x_i = x, y_i = y]}{\sum_{i=1}^n \mathbf{1}[x_i = x]}$$

(1) we may not see all (x, y) pairs, $2^d - 2$ samples

(2)

(3) $\hat{p}(y|x) \approx p(y|x)$, require sample

(4)

X can be continuous

Issues?

for classification

Logistic Regression



Process

Collect a **dataset**

$$\text{Data } \{ (x_i, y_i) \}_{i=1}^n, \quad x_i \in \mathbb{R}^d, \quad y_i \in \{0, 1\}$$

Decide on a **model**

$$f: \mathbb{R}^d \rightarrow \{0, 1\}, \quad p \in \mathcal{F}: \text{function class}$$
$$f(x) = \underset{y}{\text{argmax}} p(y|x)$$

Find the function which fits the data best

Choose a loss function

$$l(f(x), y)$$

Pick the function which minimizes loss on data

$$\hat{p} = \underset{p \in \mathcal{F}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i)$$

Use function to make prediction on new examples

$$x_{\text{new}}, \quad \hat{f}(x_{\text{new}}) = \underset{y}{\text{argmax}} \hat{p}(y|x)$$

Decide on a model, Binary Classification

To make predictions for unseen inputs (x s),

need a **general** model for $\mathbb{P}(Y = 1|X = x)$

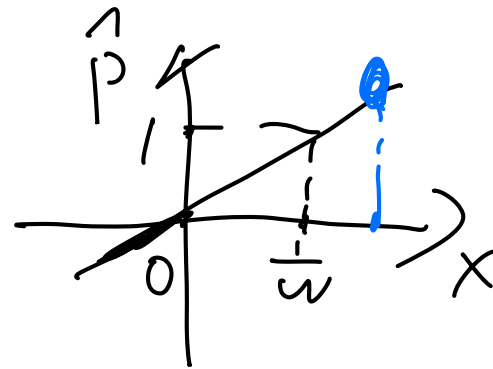
$$w^T x \in (-\infty, \infty)$$

- **What about standard linear regression model?**

$$\hat{p}(y/x) = w^T x$$

$$x \in \mathcal{R}$$

$$w > 0$$



- **Need to map real values to $[0,1]$** $\sigma(\cdot) : (-\infty, \infty) \rightarrow [0,1]$
- **We call such maps “link functions”** $\hat{p}(y/x) = \sigma(w^T x)$

Logistic Regression

$$z \rightarrow -\infty, \sigma(z) = 0$$
$$z \rightarrow \infty, \sigma(z) = 1$$

Actually classification, not regression :)

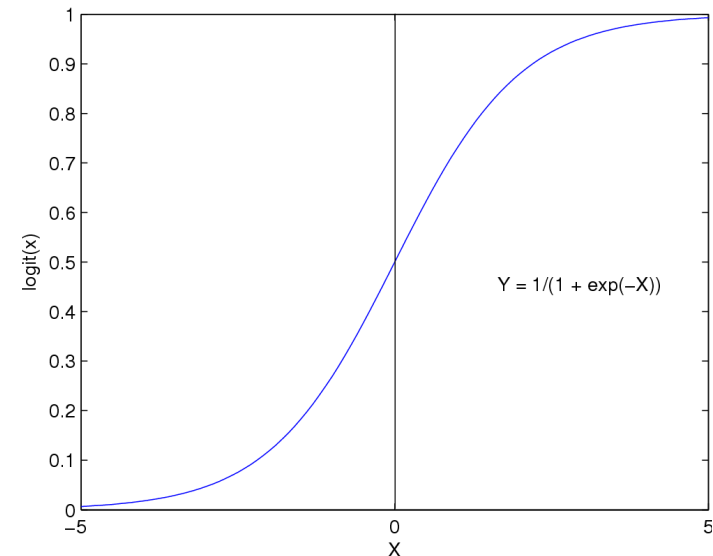
Learn $\mathbb{P}(Y = 1|X = x)$ using $\sigma(w^T x)$, for link function $\sigma(z) =$

Logistic function(or Sigmoid):

$$\frac{1}{1 + \exp(-z)}$$

$$\mathbb{P}[Y = 1|X = x, w] = \sigma(w^T x) = \frac{1}{1 + \exp(-w^T x)}$$

$$\mathbb{P}[Y = 0|X = x, w] = 1 - \sigma(w^T x) = \frac{\exp(-w^T x)}{1 + \exp(-w^T x)}$$
$$= \frac{1}{1 + \exp(w^T x)}$$



X

Features can be discrete or continuous!

k -class classification

w_1, \dots, w_k

$x \mapsto \exp(-w_i^T x)$

$$\hat{p}(y=i|x) = \frac{\exp(-w_i^T x)}{\sum_{j=1}^k \exp(-w_j^T x)}$$

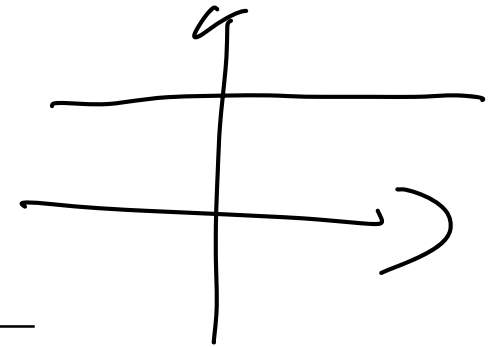
Understanding the sigmoid

if $w_0 = 0, w_1 \rightarrow 0$

$$p(Y=0 | X=x)$$

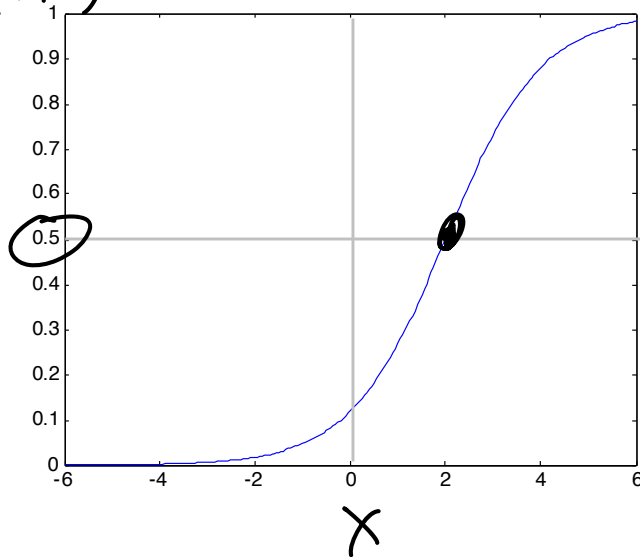
$$\sigma(w_0 + \sum_k w_k x_k) = \frac{1}{1 + e^{w_0 + \sum_k w_k x_k}}$$

offset

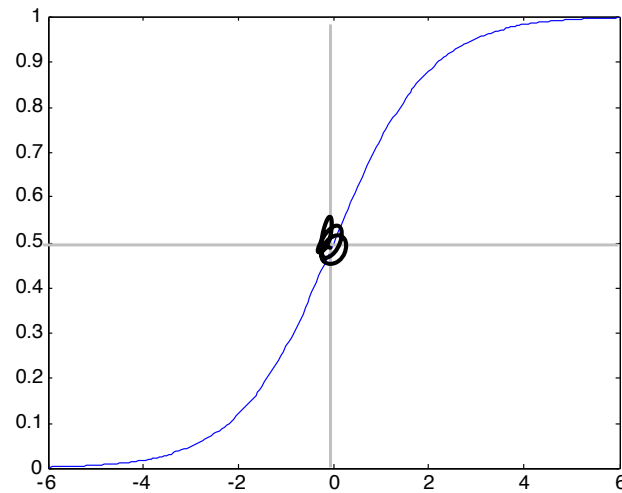


$x \in \mathcal{R}^1$

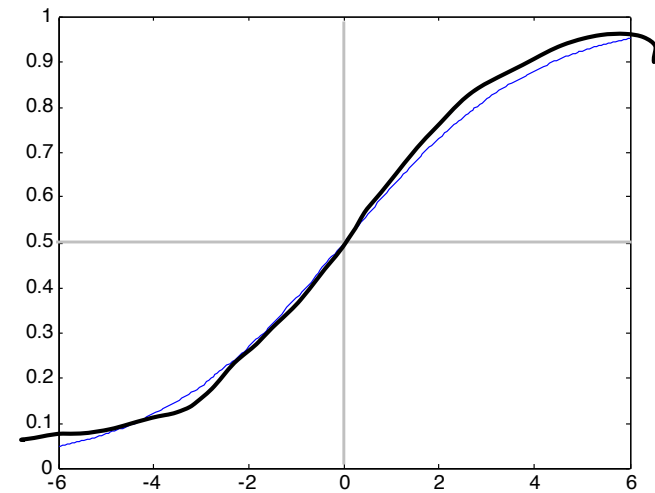
$w_0 = -2, w_1 = -1$



$w_0 = 0, w_1 = -1$



$w_0 = 0, w_1 = -0.5$



Logistic Regression

Actually classification, not regression :)

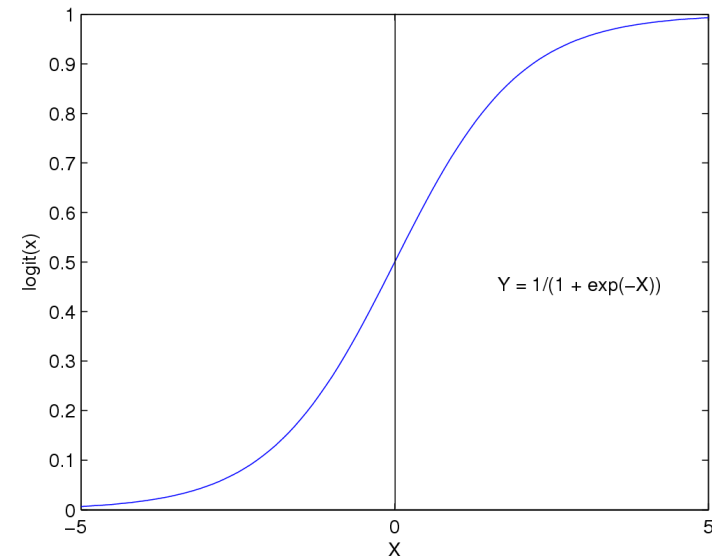
Learn $\mathbb{P}(Y = 1|X = x)$ using $\sigma(w^T x)$, for link function $\sigma =$

Logistic function(or Sigmoid):

$$\frac{1}{1 + \exp(-z)}$$

$$\mathbb{P}[Y = 1|X = x, w] = \sigma(w^T x) = \frac{1}{1 + \exp(-w^T x)}$$

$$\begin{aligned}\mathbb{P}[Y = 0|X = x, w] &= 1 - \sigma(w^T x) = \frac{\exp(-w^T x)}{1 + \exp(-w^T x)} \\ &= \frac{1}{1 + \exp(w^T x)}\end{aligned}$$



Features can be discrete or continuous!

Sigmoid for binary classes

$$\mathbb{P}(Y = 0|w, X) = \frac{1}{1 + \exp(w_0 + \sum_k w_k X_k)}$$

$$\mathbb{P}(Y = 1|w, X) = 1 - \mathbb{P}(Y = 0|w, X) = \frac{\exp(w_0 + \sum_k w_k X_k)}{1 + \exp(w_0 + \sum_k w_k X_k)}$$

wr want avg max $p(Y|X)$ $> \Rightarrow$ predict 1

$$\frac{\mathbb{P}(Y = 1|w, X)}{\mathbb{P}(Y = 0|w, X)} = \exp(w_0 + \sum_k w_k X_k) \Rightarrow < 1 \Rightarrow \text{predict 0}$$

$= (=) 0 \text{ or } 1 \text{ or } 0.5$

Sigmoid for binary classes

$$\mathbb{P}(Y = 0|w, X) = \frac{1}{1 + \exp(w_0 + \sum_k w_k X_k)}$$

$$\mathbb{P}(Y = 1|w, X) = 1 - \mathbb{P}(Y = 0|w, X) = \frac{\exp(w_0 + \sum_k w_k X_k)}{1 + \exp(w_0 + \sum_k w_k X_k)}$$

$$\frac{\mathbb{P}(Y = 1|w, X)}{\mathbb{P}(Y = 0|w, X)} = \exp(w_0 + \sum_k w_k X_k)$$

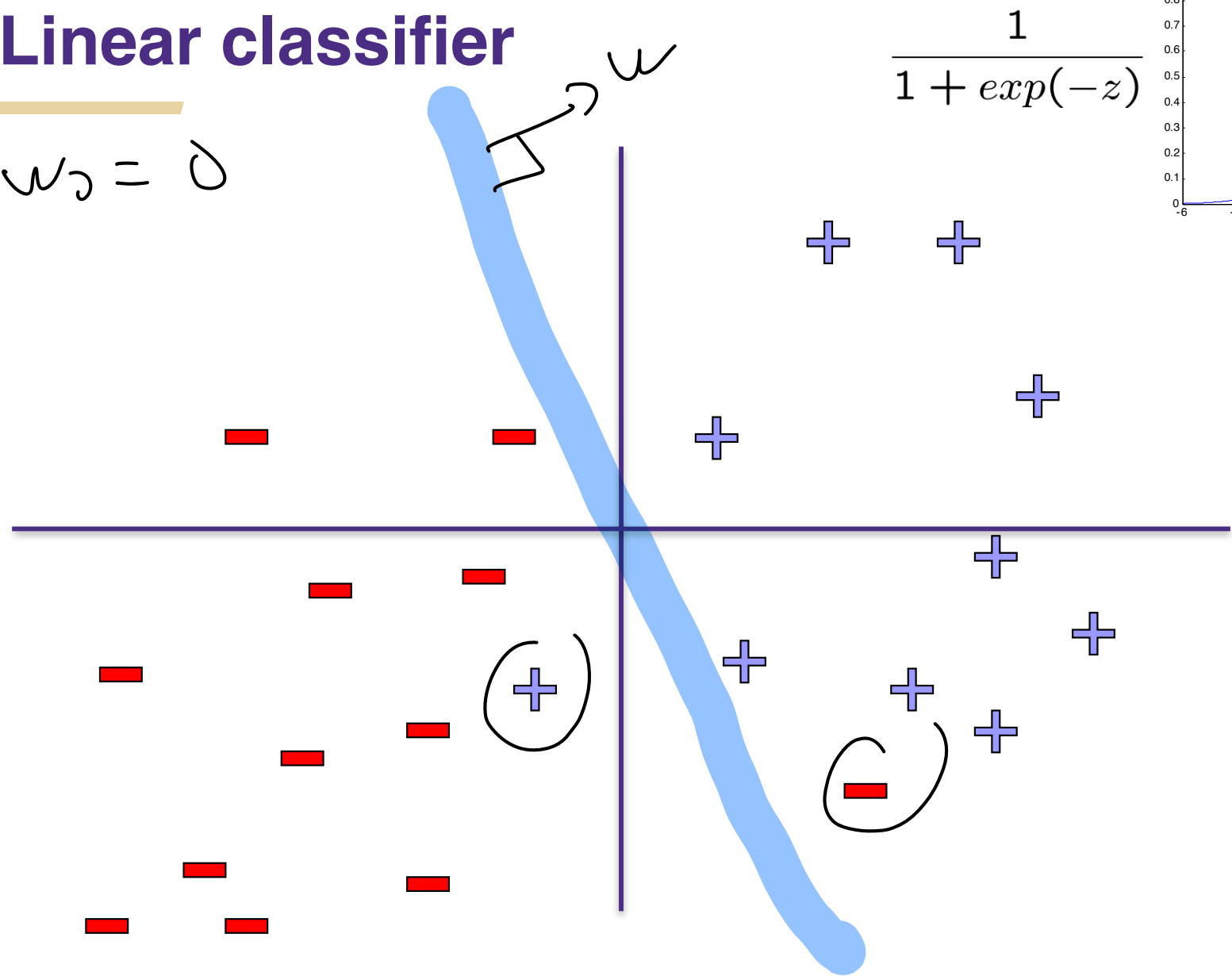
Linear Decision Rule!

$$\log \frac{\mathbb{P}(Y = 1|w, X)}{\mathbb{P}(Y = 0|w, X)} = w_0 + \sum_k w_k X_k \Rightarrow$$

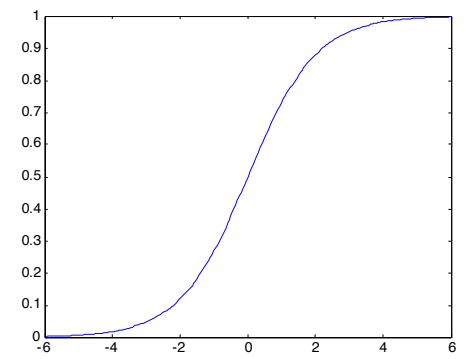
$> 0 \Rightarrow$ predict 1
 $< 0 \Rightarrow$ predict 0
 $= 0 \Rightarrow$ 0 or 1 is OK

Logistic Regression – a Linear classifier

$w_0 = 0$



$$\frac{1}{1 + \exp(-z)}$$



$$\log \frac{\mathbb{P}(Y = 1|w, X)}{\mathbb{P}(Y = 0|w, X)} = w_0 + \sum_k w_k X_k$$

Process

Decide on a **model** ✓

Find the function which fits the data best

Choose a loss function

Pick the function which minimizes loss on data

Use function to make prediction on new examples

what about 0/1 loss

$$\min_f \sum_{i=1}^n \mathbb{1}\{f(x_i) \neq y_i\}$$

- (1) not differentiable
- (2) convex

Loss function: Conditional Likelihood

- Have a bunch of iid data:

$$\{(x_i, y_i)\}_{i=1}^n \quad x_i \in \mathbb{R}^d, \quad y_i \in \{-1, 1\}$$

$$P(Y = -1|x, w) = \frac{1}{1 + \exp(w^T x)}$$

$$P(Y = 1|x, w) = \frac{\exp(w^T x)}{1 + \exp(w^T x)}$$

encoding is just
for simplicity
if $y_i \in \{0, 1\}$
 $\Rightarrow 2y_i - 1 \in \{-1, 1\}$

- This is equivalent to:

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y w^T x)}$$

- So we can compute the maximum likelihood estimator:

$$\hat{w}_{MLE} = \arg \max_w \prod_{i=1}^n P(y_i|x_i, w)$$

Loss function: Conditional Likelihood

- Have a bunch of iid data: $\{(x_i, y_i)\}_{i=1}^n$ $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y w^T x)}$$

$$\begin{aligned} \hat{w}_{MLE} &= \arg \max_w \prod_{i=1}^n P(y_i|x_i, w) \quad \text{take log, negate} \\ &= \arg \min_w \sum_{i=1}^n \log(1 + \exp(-y_i x_i^T w)) \end{aligned}$$

Logistic Loss: $\ell_i(w) = \log(1 + \exp(-y_i x_i^T w))$

Squared error Loss: $\ell_i(w) = (y_i - x_i^T w)^2$

(MLE for Gaussian noise)

Process

Decide on a **model** ✓

Find the function which fits the data best

Choose a loss function ✓

Pick the function which minimizes loss on data

$$\min_w \sum_{i=1}^n \ell(w, x_i, y_i)$$

Use function to make prediction on new examples

Loss function: Conditional Likelihood

- Have a bunch of iid data: $\{(x_i, y_i)\}_{i=1}^n$ $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$

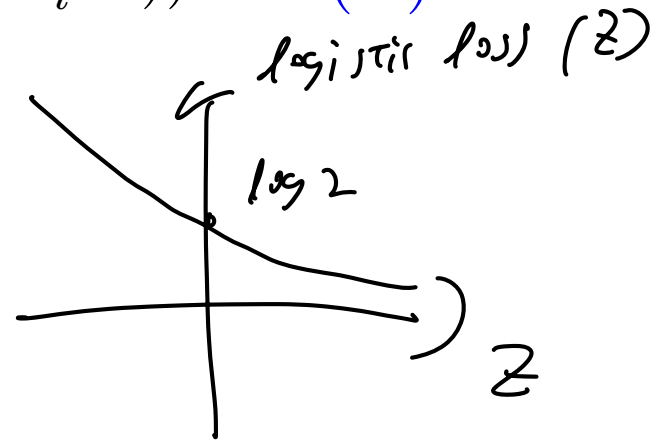
$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y w^T x)}$$

$$\hat{w}_{MLE} = \arg \max_w \prod_{i=1}^n P(y_i|x_i, w)$$

$$= \arg \min_w \sum_{i=1}^n \log(1 + \exp(-y_i x_i^T w)) = J(w)$$

What does $J(w)$ look like? Is it convex?

differentiable



Loss function: Conditional Likelihood

Loss function: Conditional Likelihood

- Have a bunch of iid data: $\{(x_i, y_i)\}_{i=1}^n$ $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$

$$P(Y = y|x, w) = \frac{1}{1 + \exp(-y w^T x)}$$

$$\begin{aligned}\hat{w}_{MLE} &= \arg \max_w \prod_{i=1}^n P(y_i|x_i, w) \\ &= \arg \min_w \sum_{i=1}^n \log(1 + \exp(-y_i x_i^T w)) = J(w)\end{aligned}$$

Good news: $J(\mathbf{w})$ is convex function of \mathbf{w} , no local optima problems

Bad news: no closed-form solution to maximize $J(\mathbf{w})$

Good news: convex functions easy to optimize

Overfitting and Linear Separability

$$y_i \in \{-1, 1\}$$

$$\arg \min_w \sum_{i=1}^n \log(1 + \exp(-y_i x_i^T w))$$

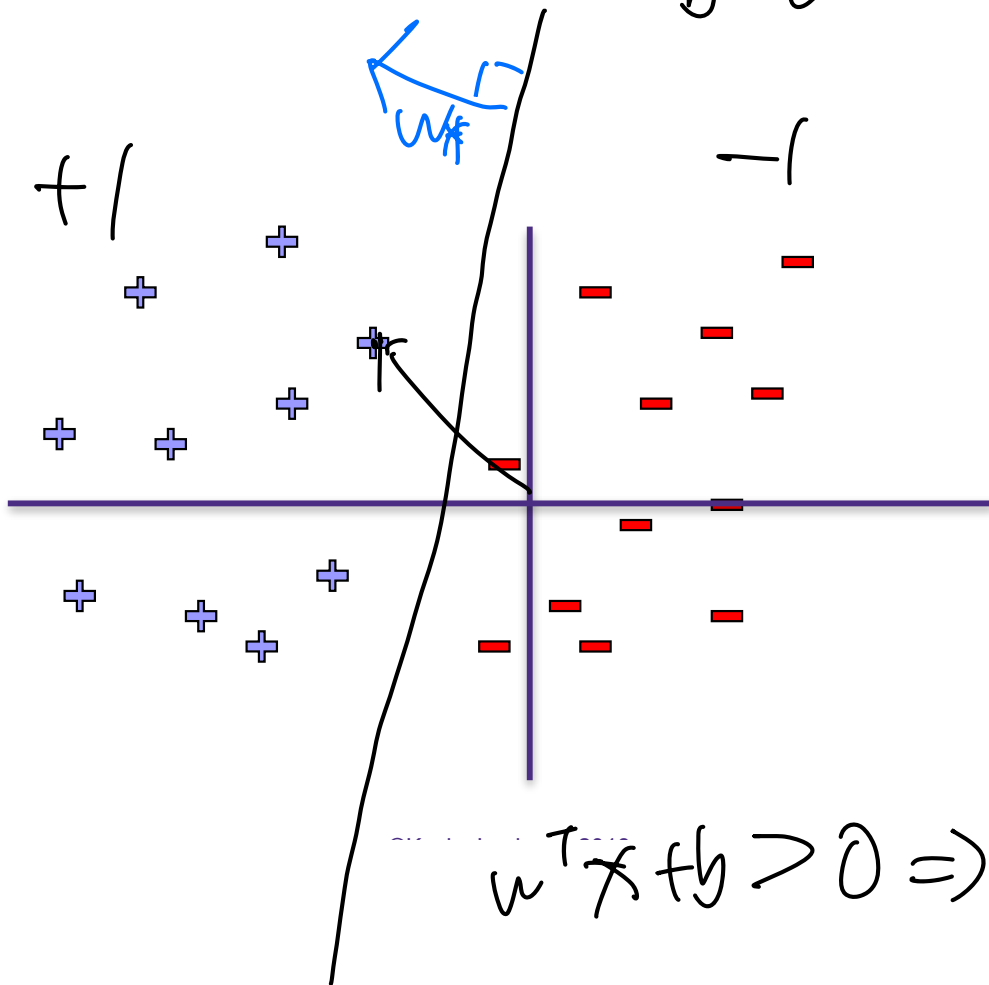
When is this loss small?

(1) $\text{Sign}(x_i^T w) = y_i$
 \cup

(2) $x_i^T w$ large

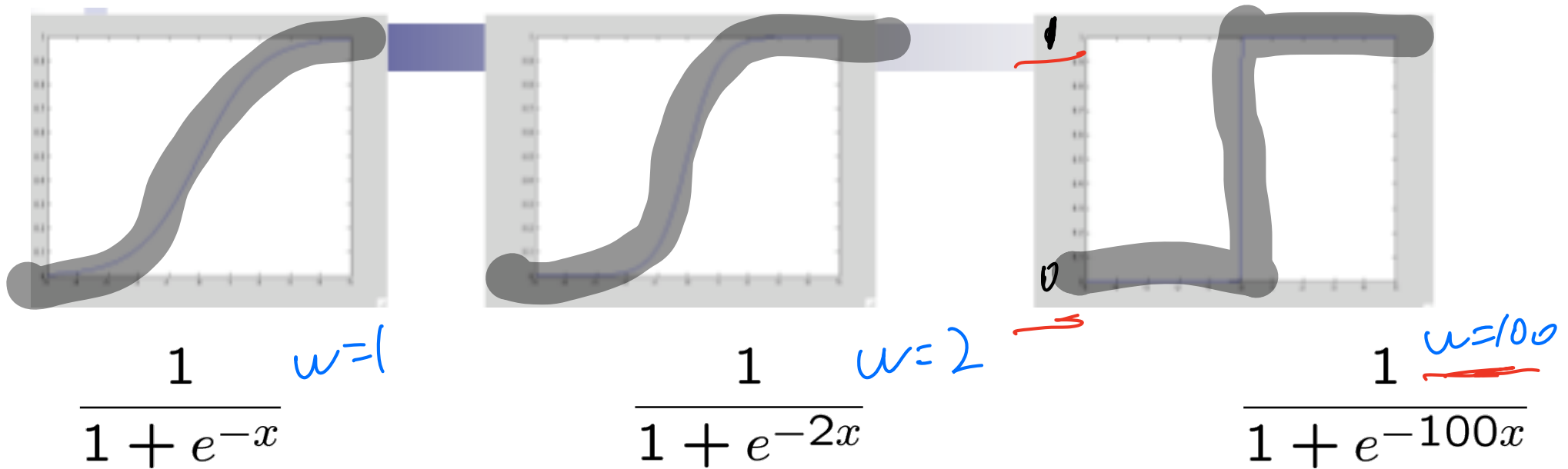
$2w^*$

$w^* \rightarrow \infty$
with same direction



Large parameters \rightarrow Overfitting

When data is linearly separable, weights $\Rightarrow \infty$



Overfitting

Penalize high weights to prevent overfitting?

Add a penalty to avoid high weights/overfitting?:

$$\arg \min_{w,b} \sum_{i=1}^n \log (1 + \exp(-y_i (x_i^T w + b))) + \lambda \|w\|_2^2$$

Be sure to not regularize the offset b !

