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

$f(X_1)$ = "Person"
$f(X_2)$ = "Animal"

Encoding:
let {"Person" = 1
"Animal" = 0

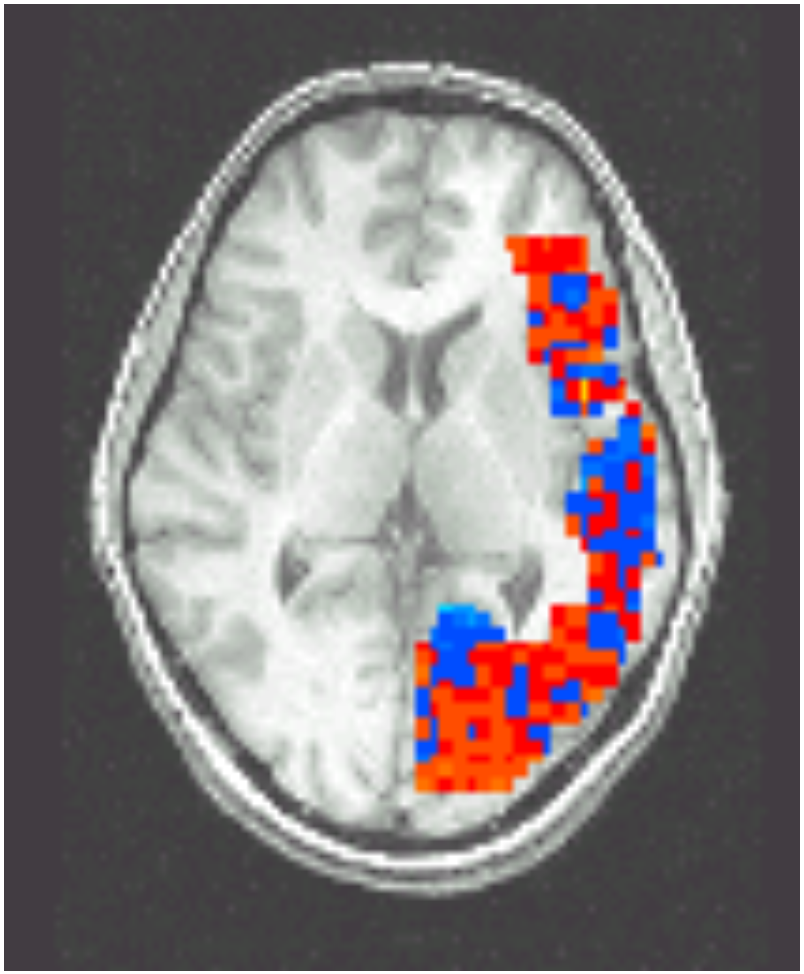
$f(X_1) = 1$
$f(X_2) = 0$

[Mitchell et al.]

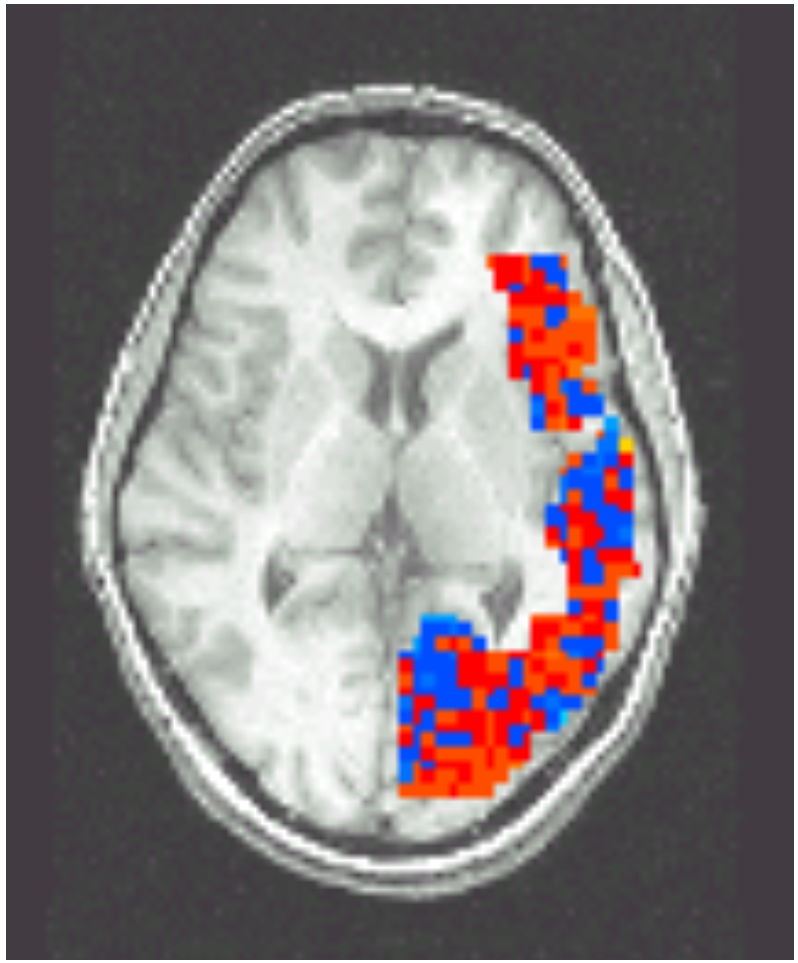
Pairwise classification accuracy: 85%

Person

Animal



X_1



X_2

Classification

- Learn f: X -> Y
  - X - features $x \in \mathbb{R}^d$
  - Y - target classes $\{0,1\}$ binary, $\{1,\dots,K\}$ multi-class
- Loss Function $\ell(f(x), y) = \mathbb{1}\{f(x) \neq y\}$ — 1 wrong, 0 correct
- Expected loss of f:

Performance measure

$$\mathbb{E}_{XY}\left[\ell(f(x), y)\right] = \mathbb{E}_{XY}\left[\mathbb{1}\{f(x) \neq y\}\right] = \mathbb{E}_X \mathbb{E}_{Y|X}\left[\mathbb{1}\{f(x) \neq y\} \mid X = x\right]$$

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

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

$$= 1 - P(Y = f(x) \mid X = x)$$

$\mathbb{E}_{XY}$ ∈ joint distribution over X, Y

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

Classification

- Learn f: X -> Y
  - X - features
  - Y - target classes

- Loss Function $\ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}$
- Expected loss of f:

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

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

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

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

Classification

- Learn f: X -> Y
  - X - features
  - Y - target classes

- Loss Function $\ell(f(x), y) = \mathbf{1}\{f(x) \neq y\}$
- Expected loss of f:

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

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

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

min f

- 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(YIX) 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(YIX), but have n iid examples

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

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

Bayes Optimal Binary Classifier

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

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

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

need $\forall x \in X$

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

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

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 $\frac{k}{m}$

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

unbiased $\quad \mathbb{E}_{\tilde{D}}[\hat{p}(Y=1 \mid X=\tilde{x})] = p(Y=1 \mid X=\tilde{x})$

$\hat{p} \to p, \quad m \to \infty$

Bayes Optimal Binary Classifier

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

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

- **What is a natural estimator for argmax_y P(Y = y | X)?**

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

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

Issues?

(1) we may not see all (X, Y) pairs, $2^d \cdot 2$ samples

(2) ~ ~ ~ ~ X, 2^d

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

(4) X can be continuous

for classification

Logistic Regression

Process

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

Collect a **dataset** Data $\{(x_i, y_i)\}_{i=1}^{n}$, $y_i \in \{0, 1\}$

Decide on a **model** $f : \mathbb{R}^d \to \{0, 1\}$, $f \in F$: function class

$f(x) = \arg\max_y p(y|x)$

Find the function which fits the data best

- **Choose a loss function** $l(f(x), y)$
- **Pick the function which minimizes loss on data** $\hat{f} = \arg\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} l(f(x), y)$

Use function to make prediction on new examples x_{new}, $\hat{f}(x_{new}) = \arg\max_y \hat{p}(y|x)$

Decide on a model, Binary Classification

To make predictions for unseen inputs (xs),

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

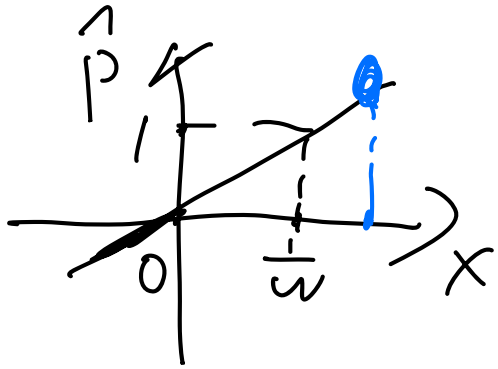
$w^T x \in [-\omega, \omega]$

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

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

$x \in \mathbb{R}$

$w > 0$



- **Need to map real values to [0,1]**
  - **We call such maps "link functions"**

$\sigma(\cdot): [-\omega, \omega] \to [0,1]$

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

Logistic Regression

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

$z = \infty, \ \sigma(z) = 1$

Actually classification, not regression :)

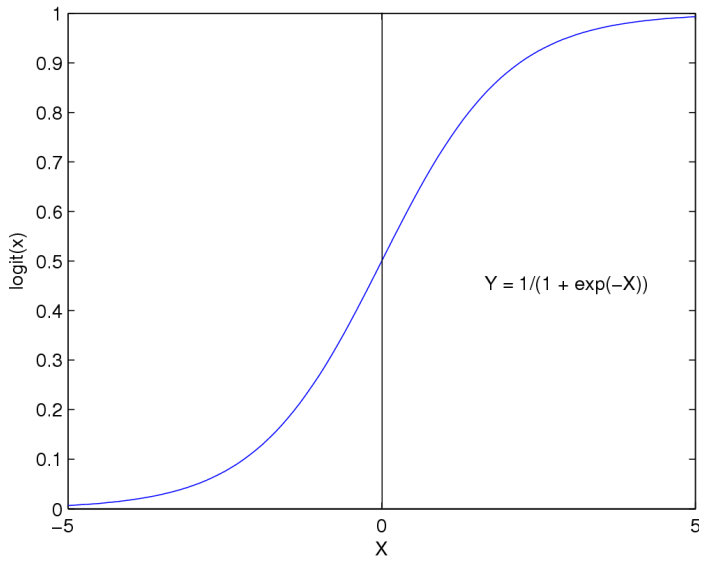
Learn $\mathbb{P}(Y = 1|X = x)$ using $\sigma(w^T x)$, for link function $\sigma(z) =$

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

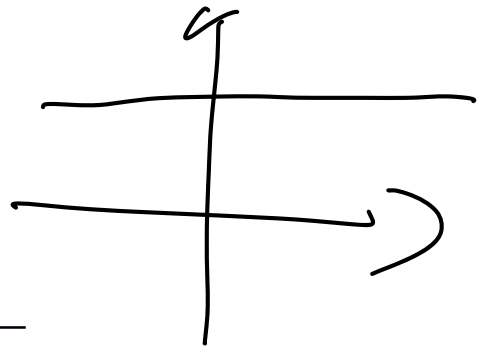
$w_1, \ldots, w_K$

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

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

Understanding the sigmoid

if $w_0 = 0$, $w_1 \to \infty$



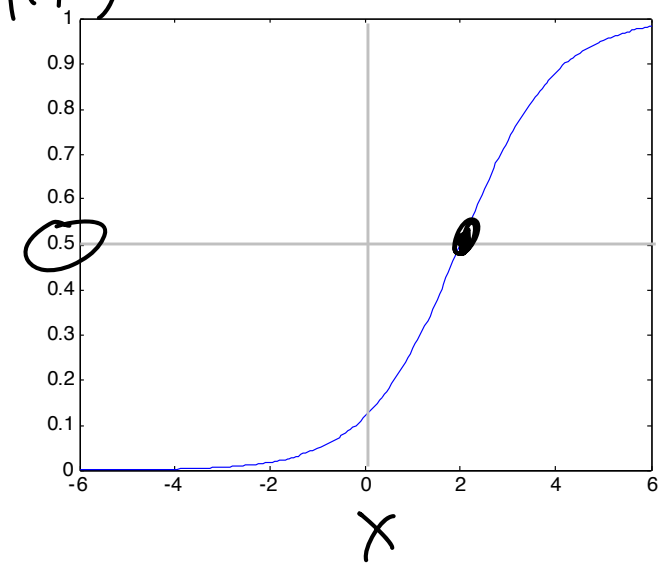
$p(Y=0 \mid X=x)$

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

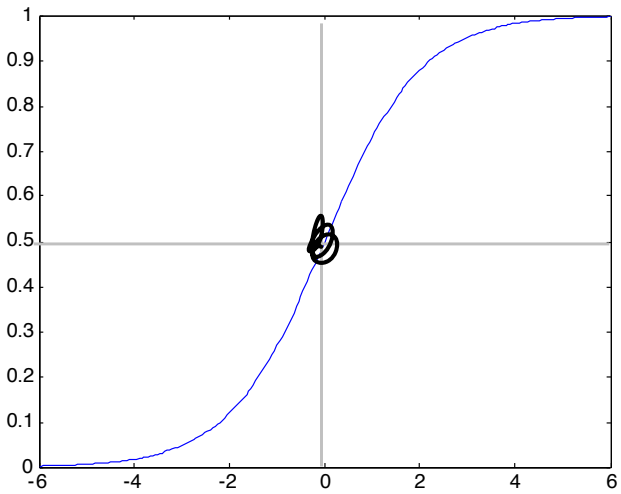
offset

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

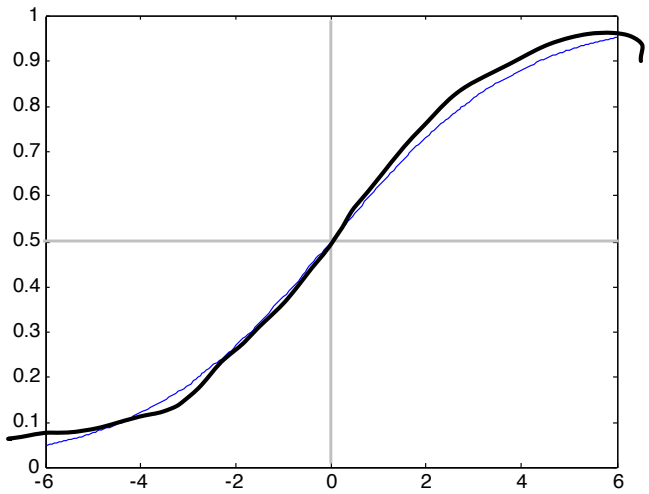
w_0=-2, w_1=-1



w_0=0, w_1=-1



w_0=0, w_1=-0.5



Logistic Regression

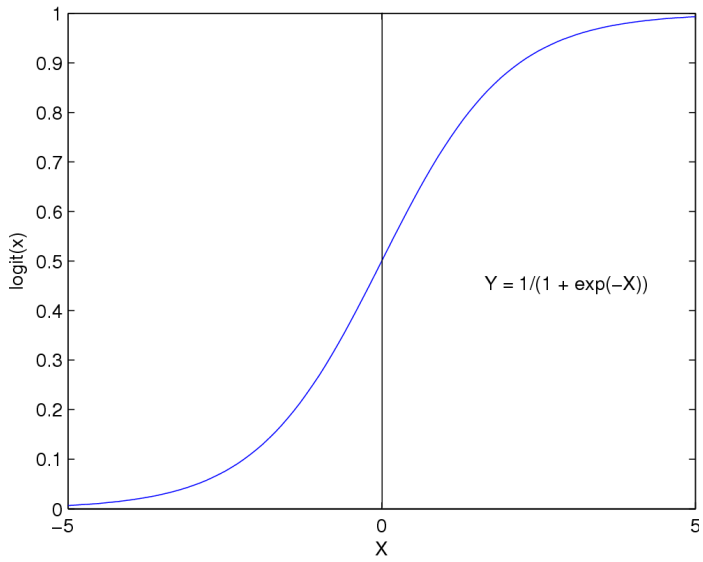
Actually classification, not regression :)

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

Logistic function(or Sigmoid): $\dfrac{1}{1 + exp(-z)}$

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

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



Features can be discrete or continuous!

Sigmoid for binary classes

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

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

we want $\arg\max_y p(Y|X)$

$$\frac{\mathbb{P}(Y=1|w,X)}{\mathbb{P}(Y=0|w,X)} = \exp\left(w_0+\sum_k w_k X_k\right) \Rightarrow \begin{cases} >1 \Rightarrow \text{predict } 1 \\ <1 \Rightarrow \text{predict } 0 \\ =1 \Rightarrow 0 \text{ or } 1 \text{ ok} \end{cases}$$

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!

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

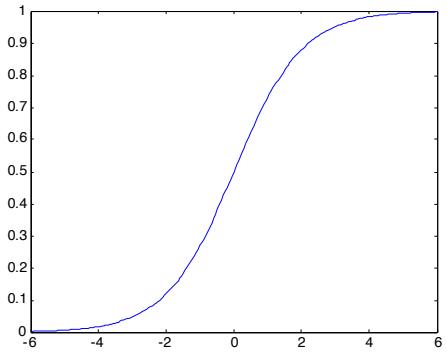
$> 0 \Rightarrow$ predict 1

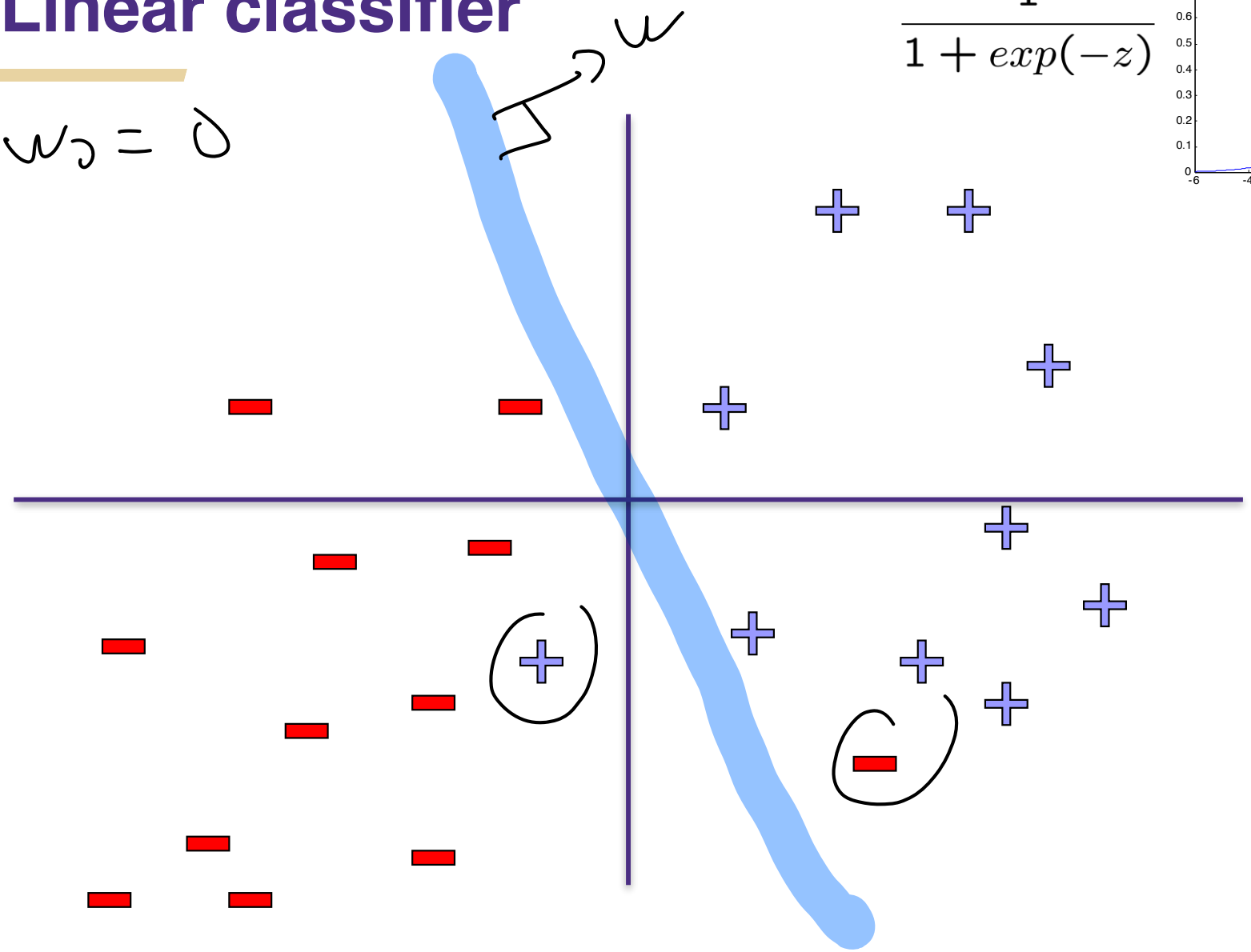
$< 0 \Rightarrow$ predict 0

$= 0 \Rightarrow$ 0 or 1 is OK

Logistic Regression – a Linear classifier

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





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

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

Loss function: Conditional Likelihood

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

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

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

- **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:**

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

Loss function: Conditional Likelihood

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

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

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

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

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**

Use function to make prediction on new examples

Loss function: Conditional Likelihood

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

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

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

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

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

Loss function: Conditional Likelihood

Loss function: Conditional Likelihood

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

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

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

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

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

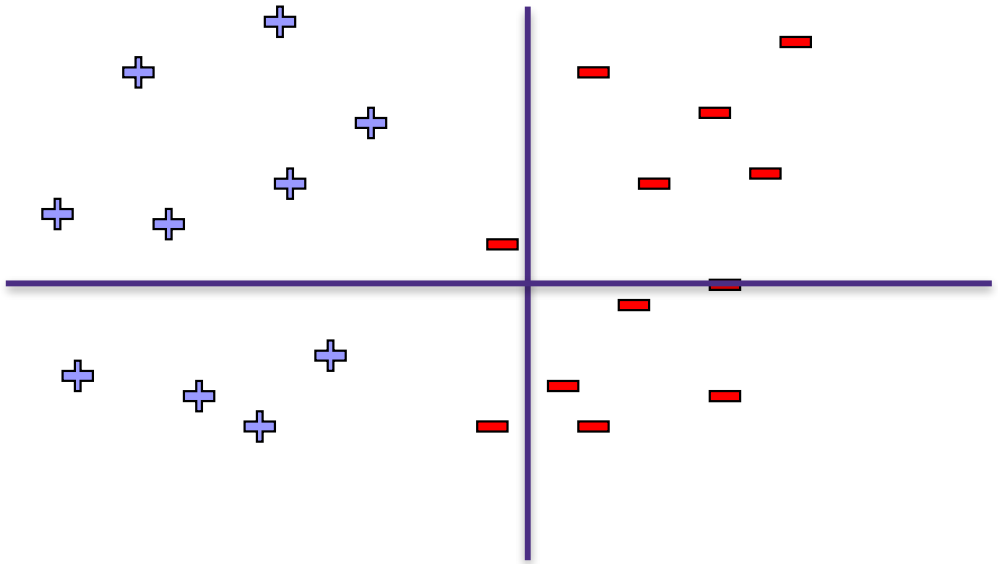
Bad news: no closed-form solution to maximize $J(\mathbf{w})$

Good news: convex functions easy to optimize

Overfitting and Linear Separability

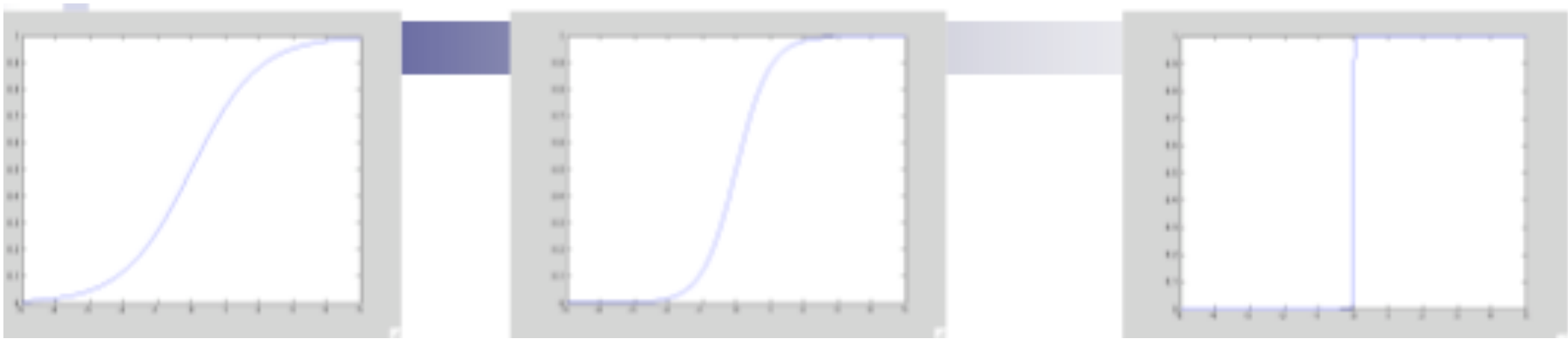
$$\arg\min_{w} \sum_{i=1}^{n} \log(1 + \exp(-y_i \, x_i^T w))$$

When is this loss small?



Large parameters → Overfitting

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



$$\frac{1}{1+e^{-x}} \qquad \frac{1}{1+e^{-2x}} \qquad \frac{1}{1+e^{-100x}}$$

Overfitting

Penalize high weights to prevent overfitting?

Add a penalty to avoid high weights/overfitting?:

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

Be sure to not regularize the offset b!