# Classification

UNIVERSITY *of* WASHINGTON

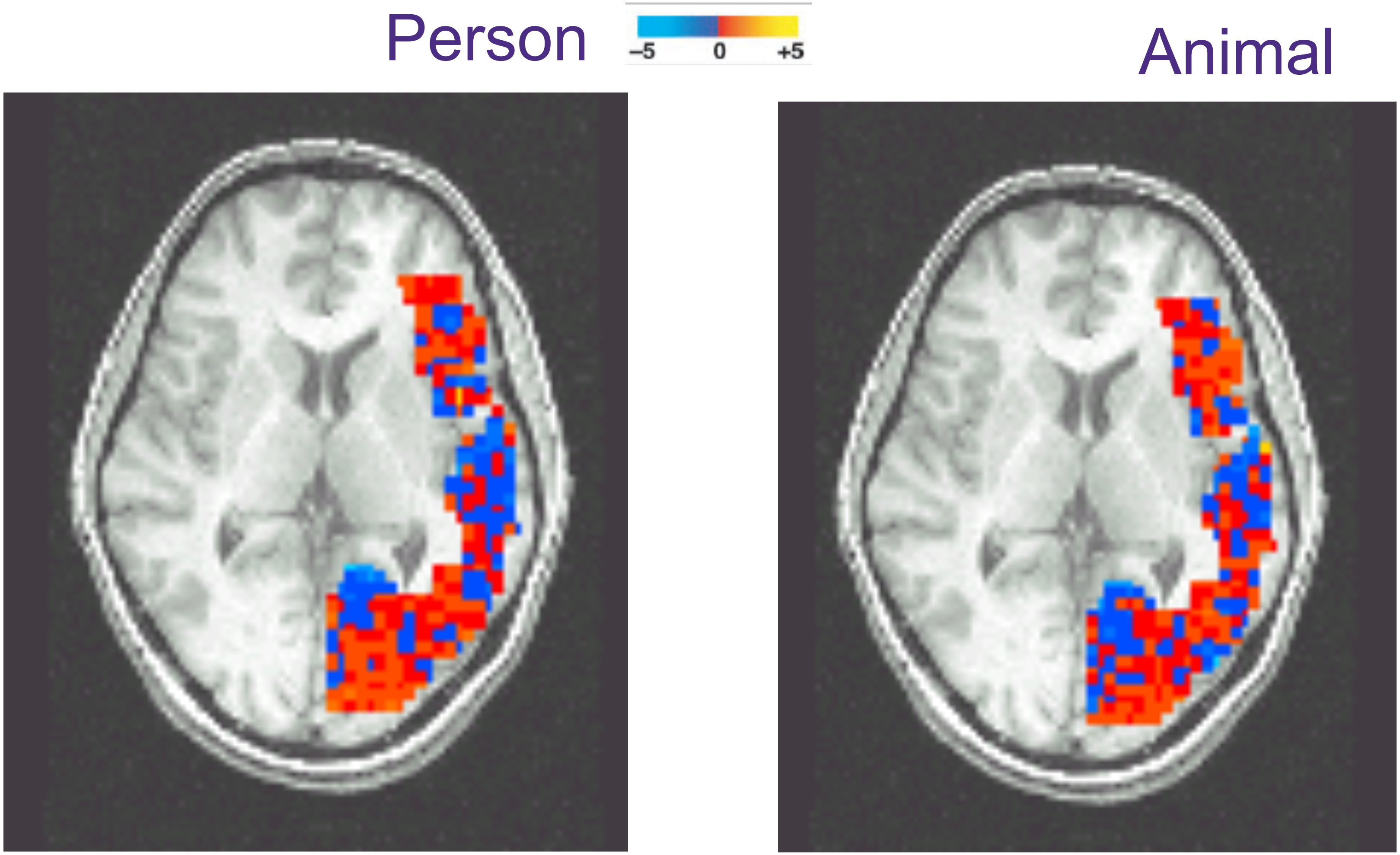
# Thus far, regression:

predict a continuous value given some inputs

# Reading Your Brain, Simple Example

[Mitchell et al.]

Pairwise classification accuracy: 85%



# Classification

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

- Loss Function
- Expected loss of f:

# 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:**

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

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

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

- **Suppose you knew P(YIX) 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**

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

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

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

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

# 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 argmax_y P(Y = y | X)?**

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

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

Issues?

# Logistic Regression

# Process

Collect a **dataset**

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

# Decide on a model, Binary Classification

To make predictions for unseen inputs ($x$s),

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

- What about standard linear regression model?

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

# Logistic Regression

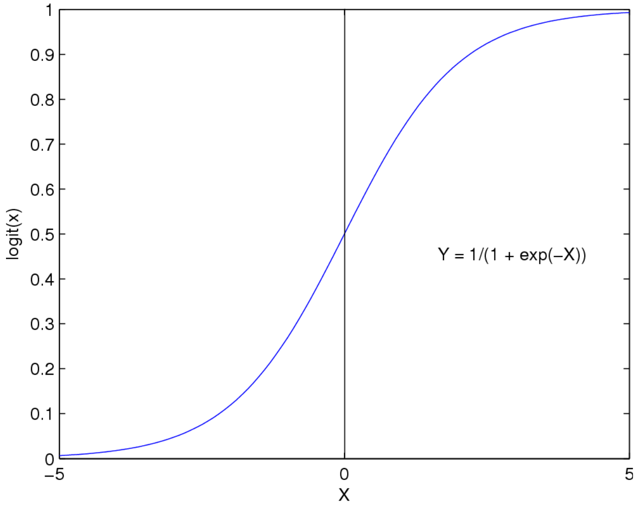
## Actually classification, not regression :)

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

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

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

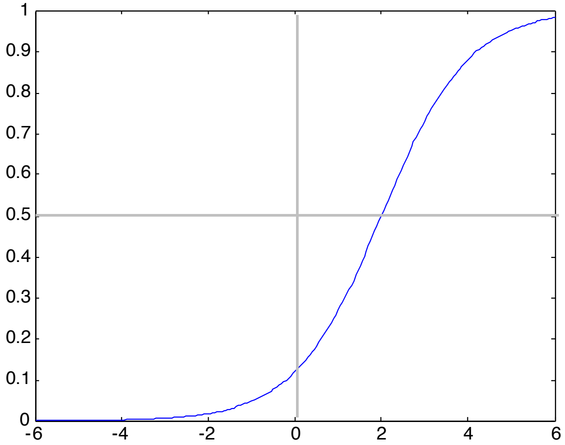
$$\mathbb{P}[Y = 0|X = x, w] = 1 - \sigma(w^T x) = \frac{\exp(-w^T x)}{1 + \exp(-w^T x)}$$
$$= \frac{1}{1 + \exp(w^T x)}$$

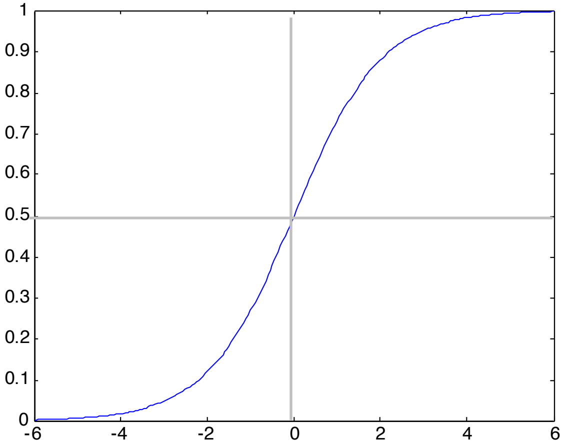


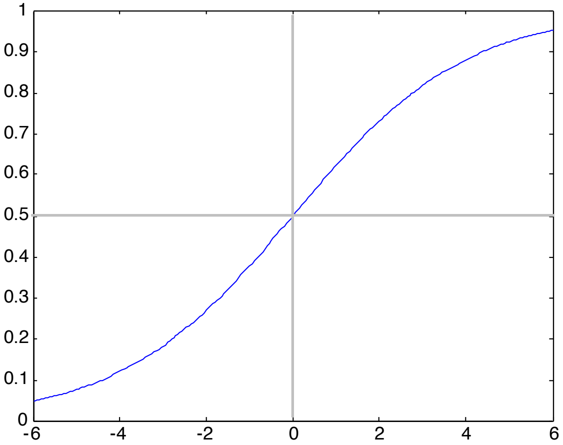
**Features can be discrete or continuous!**

# Understanding the sigmoid

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


$w_0$=-2, $w_1$=-1


$w_0$=0, $w_1$=-1


$w_0$=0, $w_1$=-0.5

# Logistic Regression

## Actually classification, not regression :)

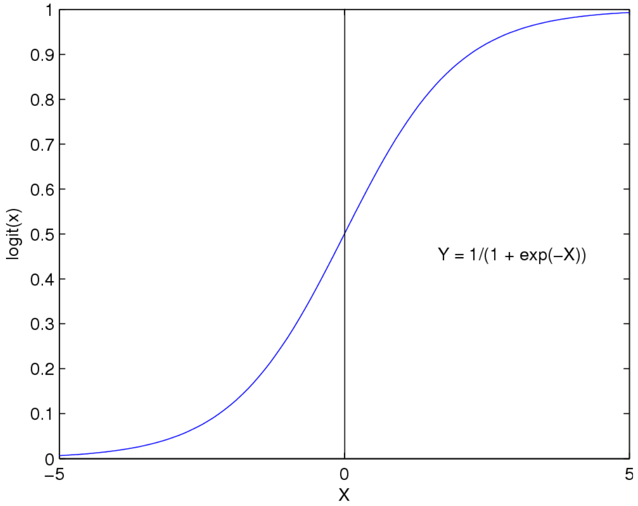
Learn $\mathbb{P}(Y=1|X=x)$ using $\sigma(w^T x)$, for link function $\sigma =$

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

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

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

$$= \frac{1}{1+\exp(w^T x)}$$



**Features can be discrete or continuous!**

# Sigmoid for binary classes

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

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

$$\frac{\mathbb{P}(Y = 1|w, X)}{\mathbb{P}(Y = 0|w, X)} =$$

# Sigmoid for binary classes

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

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

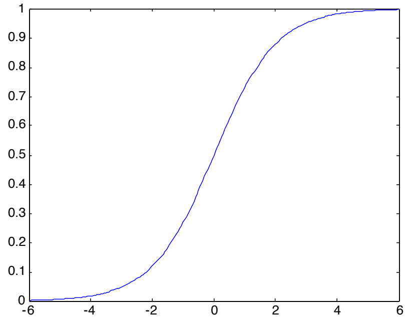
$$\frac{\mathbb{P}(Y=1|w,X)}{\mathbb{P}(Y=0|w,X)} = \exp(w_0+\sum_k w_k X_k)$$

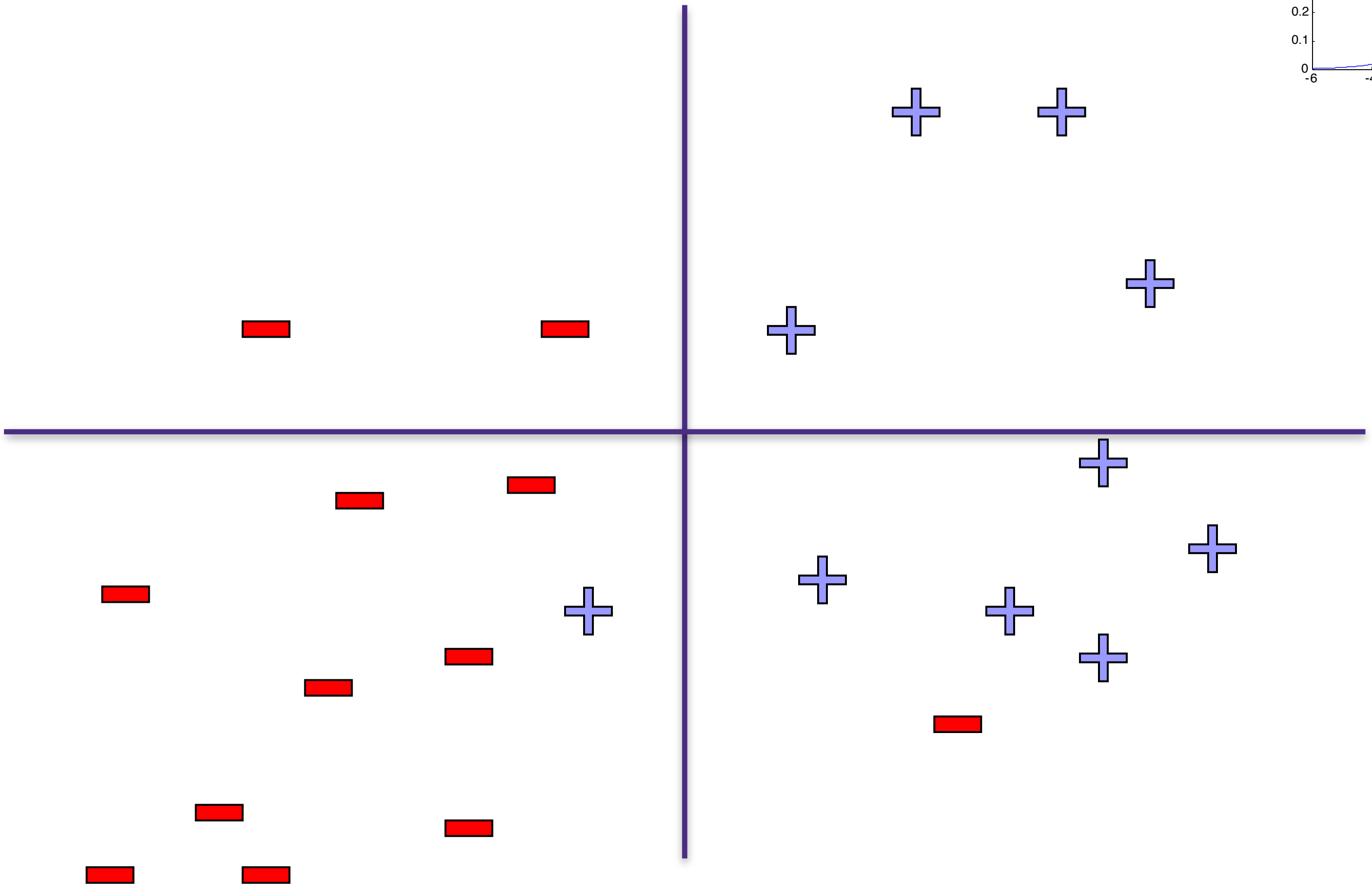
**Linear Decision Rule!**

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

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

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

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

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**

Use function to make prediction on new examples

# Loss function: Conditional Likelihood

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

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

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

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

# Loss function: Conditional Likelihood

# Loss function: Conditional Likelihood

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

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

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

Good news: *J*(**w**) is convex function of **w**, no local optima problems

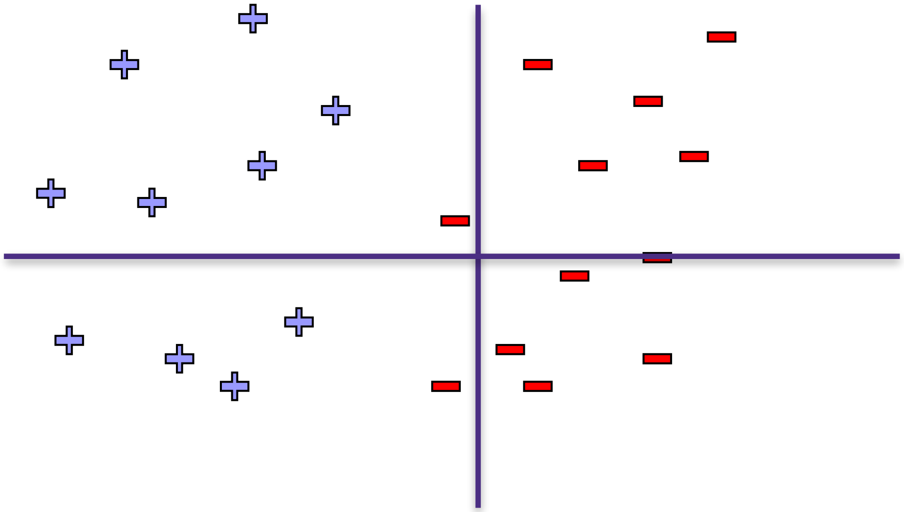
Bad news: no closed-form solution to maximize *J*(**w**)

Good news: convex functions easy to optimize

# Overfitting and Linear Separability

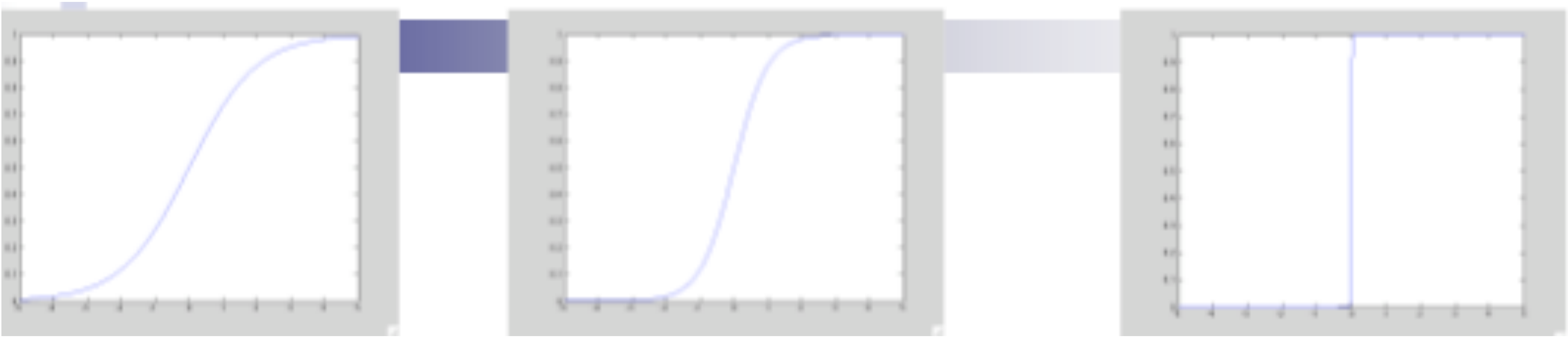
$$\arg\min_{w} \sum_{i=1}^{n} \log(1 + \exp(-y_i\, x_i^T\, w))$$

When is this loss small?



# Large parameters → Overfitting

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



$$\frac{1}{1+e^{-x}}$$

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

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

Overfitting

Penalize high weights to prevent overfitting?

## Add a penalty to avoid high weights/overfitting?:

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

Be sure to not regularize the offset $b$!