Announcement

1) HW 1 due today
2) HW 2 release today, due 5/5

Logistic Regression

Process

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

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

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

Find the function which fits the data best

Choose a loss function $\ell(f(x), y)$

Pick the function which minimizes loss on data $\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$

Use function to make prediction on new examples $x_{new} \in \mathbb{R}^d$, $\hat{f}(x_{new})$

Logistic Regression

$x \in \mathbb{R}^d$
$w \in \mathbb{R}^d$
$y \in \{0,1\}$

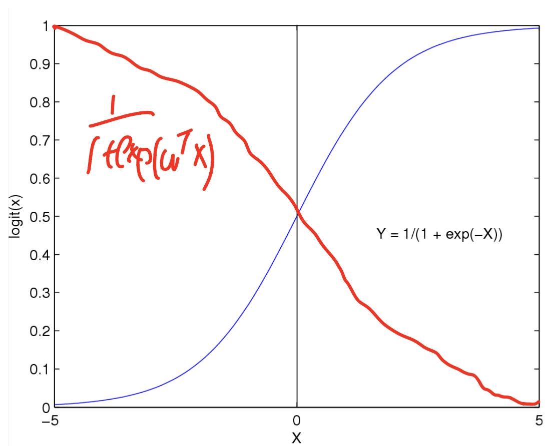
Actually classification, not regression :)

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

$w_0, w_1, \dots, w_d \in \mathbb{R}$

w_0: offset

$X_k \in \mathbb{R}$, $X \in \mathbb{R}^d$

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

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

$$\frac{\mathbb{P}(Y = 1|w, X)}{\mathbb{P}(Y = 0|w, X)} = \exp\left(w_0 + \sum_{k=1}^{d} w_k X_k\right)$$

exp in X, w

if magnitude is large (or small)

ratio is extremely large

($\to 0$)

Sigmoid for binary classes

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

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

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

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

$\Rightarrow$ if $>1 \Rightarrow$ predict 1
$<1 \Rightarrow$ predict 0
$=1 \Rightarrow$ both are OK

$\Leftrightarrow$

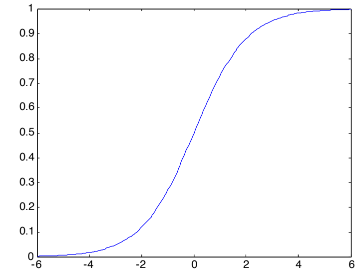
Linear Decision Rule!

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

$\Rightarrow$ $>0 \Rightarrow$ predict 1
$<0 \Rightarrow$ predict 0
$=0 \Rightarrow$ both are OK

Logistic Regression – a Linear classifier

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



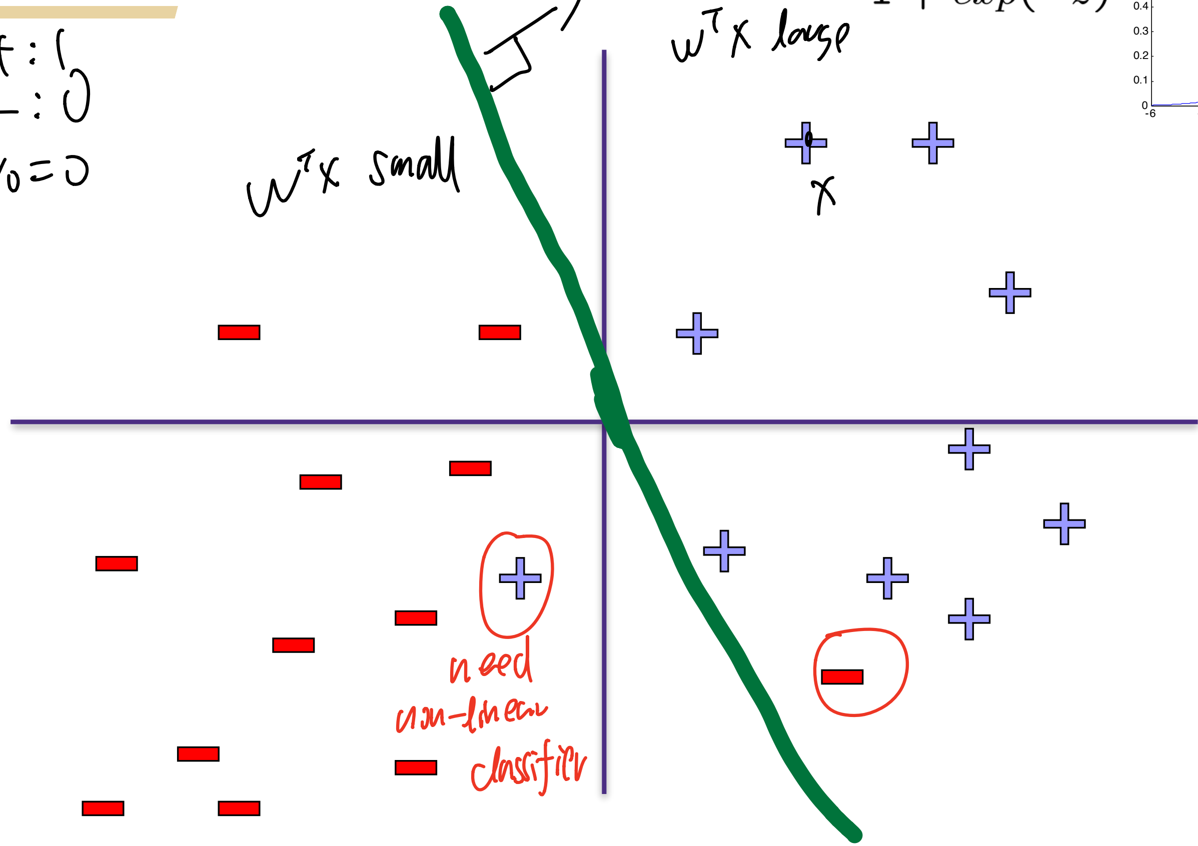
$+ : 1$
$- : 0$
$w_0 = 0$

w

$w^T x$ large

$w^T x$ small

$x \in \mathbb{R}^2$



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

Process

$$f(x) = \begin{cases} 1 & w^T x > 0 \\ 0 & \text{o.w.} \end{cases}$$

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

encoding 시
only simplicity
거 $y_i \{0,1\}$
$\Rightarrow 2y_i - 1 \in \{-1,1\}$

- **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:**

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

w: parameter
want to learn

Loss function: Conditional Likelihood

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

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

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

log is monotonic

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

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

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

(MLE for Gaussian noise)

Process

what we really care

- 0/1: $\mathbb{1}\{f(x) \neq y\}$
- $\log(1+\exp(-y w^T x))$

for training

- MLE principle
- 0/1 is hard to optimize

Decide on a **model**

Find the function which fits the data best

Choose a loss function

Pick the function which minimizes loss on data $\arg\min_{w} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$

Use function to make prediction on new examples

Loss function: Conditional Likelihood

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

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

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

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

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

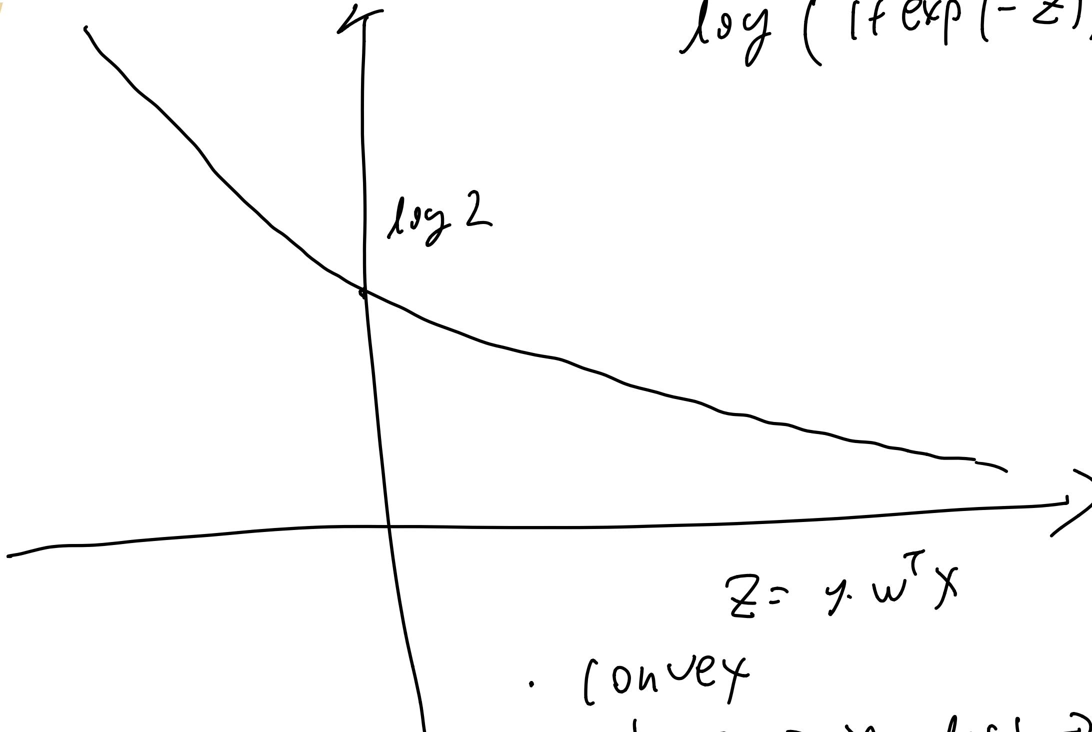
What does $J(w)$ look like? Is it convex?

$J''(w) \geq 0$

$w \in \mathbb{R}$

Loss function: Conditional Likelihood

$\log(1+\exp(-z))$



$z = y \cdot w^T x$

- convex
- if $z \to \infty$, loss $\to 0$
  $z \to -\infty$, loss $\to \infty$

intuition: if y & $w^T x$ have same sign $\to$ loss small
if y & $w^T x$ have different sign $\to$ loss big

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

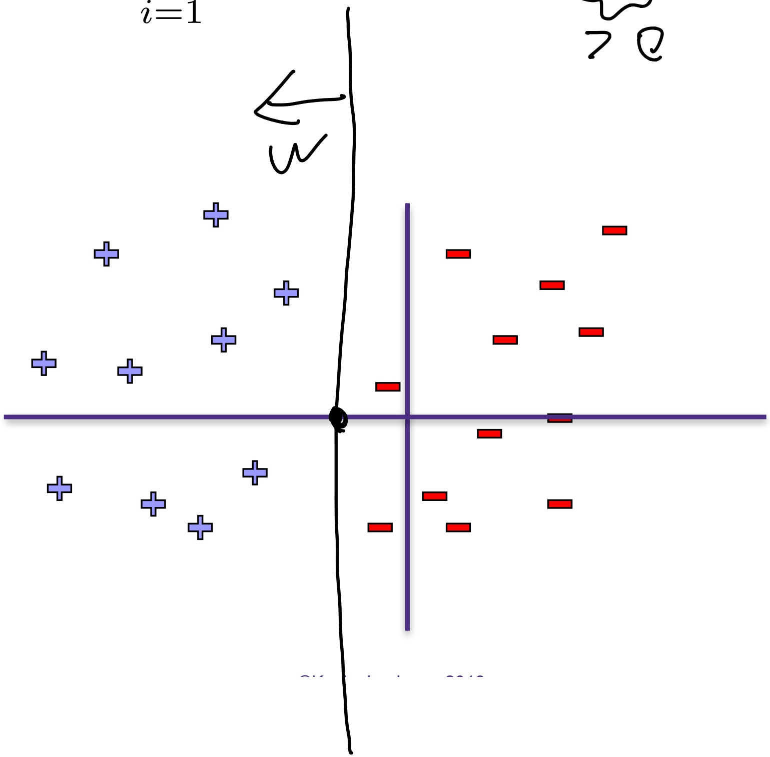
Bad news: no closed-form solution to maximize $J(\mathbf{w})$

Good news: convex functions easy to optimize

(gradient descent)

Overfitting and Linear Separability

$$\arg\min_w \sum_{i=1}^{n} \log(1 + \exp(-\underbrace{y_i\, x_i^T w}_{>0}))$$



When is this loss small?

(1) direction, offset

$\mathrm{sign}(x_i^T w) = \mathrm{sign}(y_i)$

(2) magnitude of w

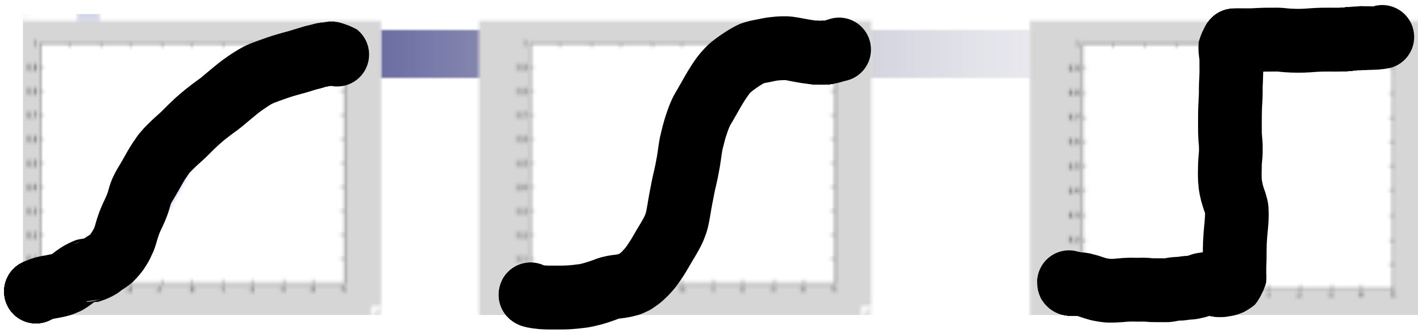
$w \to 2w \to 4w \cdots$

$\|w\|_2 \to \infty$

$J(w) \to 0$

Large parameters → Overfitting

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



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

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

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

Overfitting

large weight $P(y|x) \to 0$ or 1

often not accurate → want $P(y|x) \in (0+c, 1-c)$

c is constant

Penalize high weights to prevent overfitting?

Regularized Conditional Log Likelihood

$w, (x, y) \to y > 0 \quad b < 0 \qquad x^T w + b \ \& \ y$

$x^T w > 0 \to$ if $\|w\|$ small, $x^T w + b < 0$, if $\|w\|$ large, $w^T x + b > 0$

Add a penalty to avoid high weights/overfitting?:

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

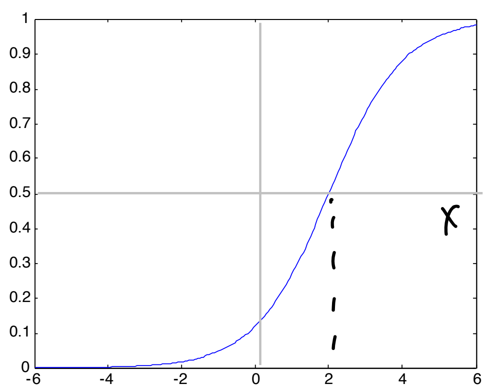
can also use ℓ_1

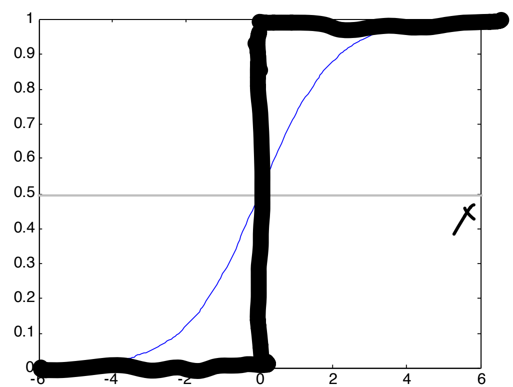
Be sure to not regularize the offset b!

$y = \sigma(w_0 + w_1 \cdot x)$

$w_1 = 1$

$w_0 = -2$





$w_1 = 1$

$w_0 = 0$