Logistic regression

- Logistic model for distributions over $\mathbb{R}^d \times \{+1, -1\}$ (or $\mathbb{R}^d \times \{0, 1\}$).

Definition: We say that a distribution D over $\mathbb{R}^d \times \{+1, -1\}$ satisfies the logistic model if there exists $w \in \mathbb{R}^d$ such that if $(X, Y) \sim D$ then

$$\Pr[Y = y \mid X = x] = \frac{1}{1 + e^{-y w^T x}}$$

for any $(x, y) \in \mathbb{R}^d \times \{+1, -1\}$.

Note: Logistic model does not place any assumptions on the

place ... marginal distribution of X:

$$\Pr[X=x, Y=y] = \Pr[X=x] \cdot \Pr[Y=y \mid X=x]$$

this can be anything

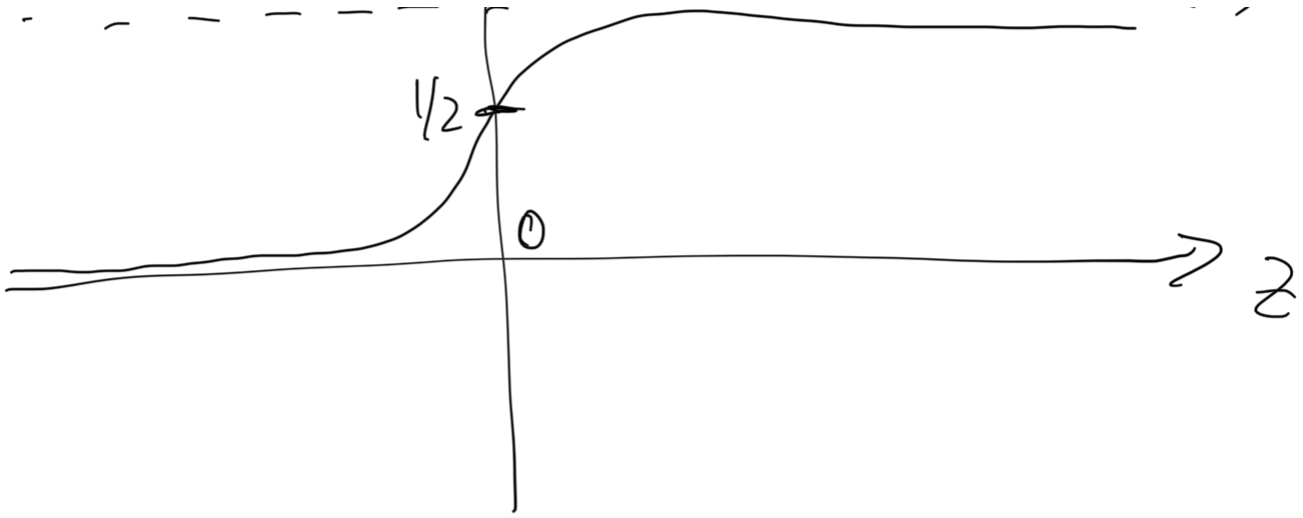
this has to satisfy logistic model assumption

Why logistic regression?

1) Define $\sigma: \mathbb{R} \to (0,1)$

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

$\sigma(z)$



$\sigma(z)$ is called

a) the sigmoid function

because of its S-shape

b) logistic function

(Etymology of the word "logistic" is bit fuzzy.)

$$Pr[Y=y \mid X=x] = \sigma(y w^T x)$$

or

$$Pr[Y=1 \mid X=x] = \sigma(w^T x)$$

↑ Generalized linear model

2) "Odds" of probability event A are

"$Pr[A] : Pr[A^c]$"

- Mathematically, odds are $\frac{Pr[A]}{Pr[A^c]}$

- The conditional odds for

logistic regression are

$$\frac{Pr[Y=1 \mid X=x]}{Pr[Y=-1 \mid X=x]}$$

$$= \frac{\sigma(w^T x)}{\sigma(-w^T x)}$$

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

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

$$= \frac{e^{w^T x} + \left(e^{w^T x}\right)^2}{e^{w^T x} + 1}$$

$$= e^{w^T x}$$

- Log odds (logarithm of odds) are a linear function of x:

$$\ln\left(\frac{\Pr[Y=1 \mid X=x]}{\Pr[Y=-1 \mid X=x]}\right) = w^T x$$

- Logistic function $\sigma : \mathbb{R} \to (0,1)$ maps log-odds to probabilities:

$$\sigma(w^T x) = \sigma\left(\ln\left(\frac{\Pr[Y=1 \mid X=x]}{\Pr[Y=-1 \mid X=x]}\right)\right)$$

$$= \Pr[Y=1 \mid X=x]$$

log odds

probability

Training logistic regression model

- Let $S = ((x_1, y_1), \ldots, (x_m, y_m))$ be a labeled data set where $x_1, x_2, \ldots, x_m \in \mathbb{R}$ and $y_1, y_2, \ldots, y_m \in \{+1, -1\}$.

- We want find model

- parameters $w \in \mathbb{R}^d$ that "best fit the data."

- This is (usually) done with maximum likelihood principle.

- Likelihood function $L: \mathbb{R}^d \to \mathbb{R}$

$$L(w) = \Pr_w[y_1, y_2, \ldots, y_m \mid x_1 \ldots x_m]$$

where subscript w indicates logistic model with parameter w.

- Likelihood $\neq$ Probability

1) Likelihood is a function of w

2) Probability is a measure i.e. a function of events.

- Suppose $(X_1, Y_1) \dots (X_m, Y_m)$ is an i.i.d. sample. Then,

$$L(w) = \prod_{i=1}^{m} Pr_w[y_i \mid x_i]$$

$$= \prod_{i=1}^{m} \sigma(-y_i w^T x_i)$$

by independence

- We want to find

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmax}} \; L(w)$$

- This is equivalent to

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; \underbrace{-\ln\left(L(w)\right)}$$

Negative log-likelihood

$$-\ln\left(L(w)\right) = -\ln \prod_{i=1}^{m} \sigma(y_i w^T x_i)$$

$$= -\sum_{i=1}^{m} \ln \sigma(y_i w^T x_i)$$

$$= -\sum_{i=1}^{m} \ln \frac{1}{1 + e^{-y_i w^T x_i}}$$

$$\sum_{i=1} \qquad 1 + e^{\quad} \quad x_i$$

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

- We define <u>log-loss</u> or <u>logistic loss</u>

$$\ell(w, x, y) = \ln\left(1 + e^{-y\, w^T x}\right)$$

- There is no "closed form" expression for

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{m} \ln\left(1 + e^{-y_i w^T x_i}\right)$$

- The minimum needs to be computed by "numerical" methods.

- One can show that the function

$$f: \mathbb{R}^d \to \mathbb{R},$$

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

is convex.

- Convexity makes finding minimum of $f(w)$ easy to find.