1 Binary Logistic Regression

Model

1. Binary classification: $y \in \{0, 1\}$
2. Want to predict probability of being in a particular class: $P(y = 1|\mathbf{x}; \mathbf{w})$
3. Could fit a linear model: $f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T \mathbf{x}$
4. But this could give predictions outside $[0, 1]$ for some test inputs (invalid probabilities)
5. Use the sigmoid function to force the output to lie in the $[0, 1]$ range:

$$f(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$$

6. Interpret $f(\mathbf{x}; \mathbf{w}) = P(y = 1|\mathbf{x}; \mathbf{w})$, implying $P(y = 0|\mathbf{x}; \mathbf{w}) = 1 - f(\mathbf{x}; \mathbf{w})$

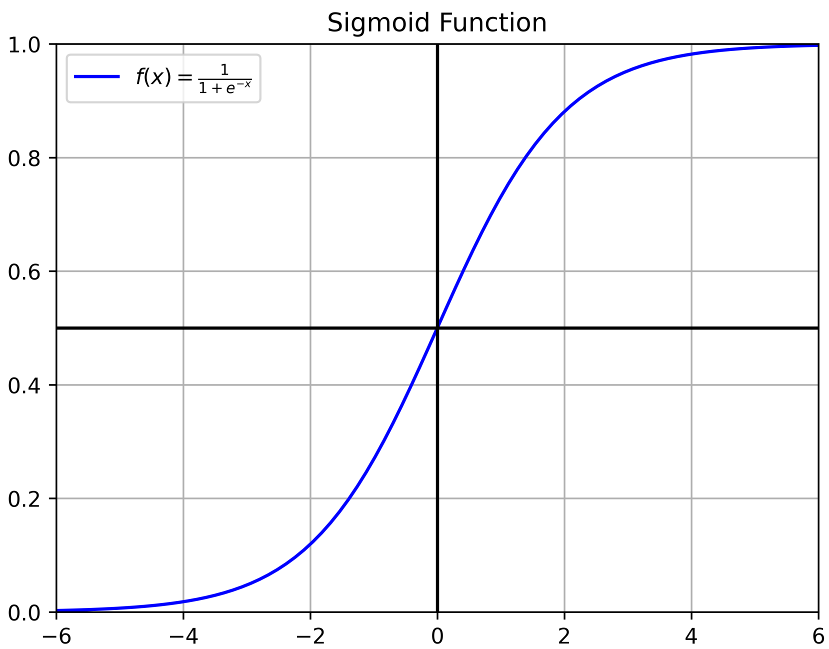

Figura 1: Function used to force the output to lie in the $[0, 1]$ range

Loss Funciton

We observe data $\{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$, with $y \in \{0,1\}$, Using maximum likehood:

$$\begin{aligned} L(\mathbf{w}) &= P(y^{(1)}|\mathbf{x}^{(1)};\mathbf{w}) \cdot P(y^{(2)}|\mathbf{x}^{(2)};\mathbf{w}) \cdots P(y^{(n)}|\mathbf{x}^{(n)};\mathbf{w}) \\ &= \prod_{n=1}^{N} P(y^{(n)}|\mathbf{x}^{(n)};\mathbf{w}) \end{aligned}$$

minimising the negative log likehood

$$\begin{aligned} J(\mathbf{w}) = -\log L(\mathbf{w}) = -\log \prod_{n=1}^{N} P(y^{(n)}|\mathbf{x}^{(n)};\mathbf{w}) &= -\sum_{n=1}^{N} \log P(y^{(n)}|\mathbf{x}^{(n)};\mathbf{w}) \\ (*)\ P(y|\mathbf{x};\mathbf{w}) = \begin{cases} f(\mathbf{x};\mathbf{w}) & if \quad y=1 \\ 1-f(\mathbf{x};\mathbf{y}) & if \quad y=0 \end{cases} &= \begin{cases} \sigma(\mathbf{w}^T;\mathbf{x}) & if \quad y=1 \\ 1-\sigma(\mathbf{w}^T;\mathbf{x}) & if \quad y=0 \end{cases} \\ \implies P(y|\mathbf{x};\mathbf{w}) &= \sigma(\mathbf{w}^T;\mathbf{x})^y(1-\sigma(\mathbf{w}^T;\mathbf{x}))^{1-y} \\ &= -\sum_{n=1}^{N} \log\left[\sigma(\mathbf{w}^T;\mathbf{x}^{(n)})^y(1-\sigma(\mathbf{w}^T;\mathbf{x}^{(n)})^{1-y^{(n)}})\right] \\ &= -\sum_{n=1}^{N} [\log \sigma(\mathbf{w}^T;\mathbf{x}^{(n)})^y + (1-y^{(n)}) \cdot \log(1-\sigma(\mathbf{w}^T;\mathbf{x}^{(n)}))] \end{aligned}$$