Introduction to Machine Learning

Notes taken by Runqiu Ye
Carnegie Mellon University

Spring 2025

Contents

1 Supervised Learning

1.1 Logistic Regression

Logistic regression is used for classfication problems. Logistic regression takes in input feature $x \in \mathbb{R}^n$, and output a prediction $y \in \{0, 1\}$. The hypotheses function $h_\theta(x)$ is chosen as

$$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$

where

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

is the sigmoid function.

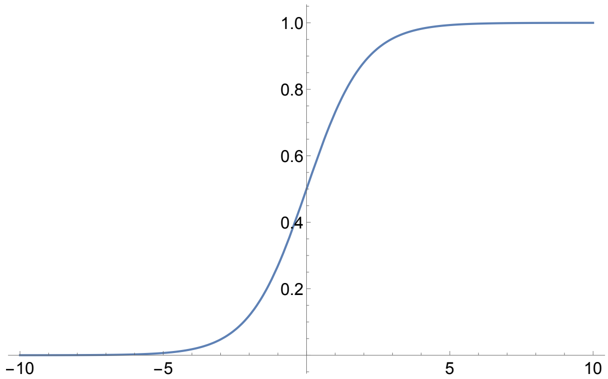

Figure 1: A plot of the sigmoid function $\sigma(z)$.

A plot of the sigmoid function is shown in Figure 1. The range of the sigmoid function is bounded in $[0, 1]$. In particular, $\sigma(z) \to 1$ when $z \to \infty$ and $\sigma(z) \to 0$ as $z \to -\infty$. A useful property about the sigmoid function is its derivative. It is easy to verify that

$$\sigma'(z) = \frac{e^{-z}}{(1 + e^{-z})^2} = \sigma(z)(1 - \sigma(z)).$$

To fit the parameter θ to dataset, assume that

$$\begin{aligned} p(y = 1 \mid x; \theta) &= h_\theta(x), \\ p(y = 0 \mid x; \theta) &= 1 - h_\theta(x). \end{aligned}$$

Note that

$$p(y \mid x; \theta) = h_\theta(x)^y (1 - h_\theta(x))^{1-y}.$$

Assuming n independent training examples, the likelihood function

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^{n} h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}. \end{aligned}$$

It is easier to maximize the log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} y^{(i)} h_\theta(x^{(i)}) + (1 - y^{(i)})(1 - h_\theta(x^{(i)})).$$

This is called the logisitic loss or the binary cross-entropy.