Binary Classification

- Given a dataset $\mathcal{D}$, we like to construct a function

$$f : \mathcal{D} \to \{0, 1\},$$

where 0 = (category 0) and 1 = (category 1).

- Examples
  1. $\mathcal{D} = \{\text{SAT scores}\}$, 0 = rejection, 1 = admission
  2. $\mathcal{D} = \{\text{emails}\}$, 0 = non-spam, 1 = spam

- More precisely, we will construct a function

$$f : \mathcal{D} \to [0, 1]$$

so that $f(d)$, $d \in \mathcal{D}$, represents the **probability** that d belongs to (category 1).

- In the SAT scores example,

$$f(1350) = 0.732$$

would mean ”A student with SAT score 1350 is accepted with probability 0.732”.

Q: How can we construct such a function?

- In linear regression, we use a linear model and minimize the mean square error (MSE).
- In logistic regression, our strategy will be similar to that of linear regression, and the method is called **maximum likelihood estimation** (MLE).

A *training set* $\mathcal{T}$ is given with known classification.

- Step 1: Using a probabilistic model, write a function out of $\mathcal{T}$ with unknown *parameters* or *weights*.
- Step 2: Determine the parameters so that the known classification may have the maximum likelihood.

It is customary to take the negative log of the likelihood function, and the resulting function is called the cross-entropy. Then we need to *minimize* the cross-entropy.

The cross-entropy function will look like

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\},$$

where

$$y_n = \sigma(w_1 x_{n1} + w_2 x_{n2} + \cdots + w_k x_{nk} + w_{k+1})$$

and

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

Goal: Determine $\boldsymbol{w} = (w_1, w_2, \ldots, w_{k+1})$ which minimizes $E(\boldsymbol{w})$.

- The cross-entropy $E(\boldsymbol{w})$ is not linear, and it is not possible to calculate a closed-form formula for $\boldsymbol{w}_*$ which minimizes $E(\boldsymbol{w})$.
- On the other hand, it can be shown that $E(\boldsymbol{w})$ is convex, and a global minimum exists.
- We will find an approximate value for $\boldsymbol{w}_*$ using gradient descent and Newton's method.
- The method of gradient descent is widely used in many other parts of machine learning.

Things to Do for Binary Classification

1. Gradient Descent and Newton's Method
2. Probability Theory
3. Logistic Regression