Maximum Likelihood Estimation (MLE)

Goal:

Find parameters that maximize likelihood of observed data

1

Likelihood Function:

$$L(w,b) = \prod P(y_i|x_i)$$

2

Log-Likelihood (easier to optimize):

$$\ell = \sum log\ P(y_i|x_i)$$

3

For Binary Classification:

$$\ell = \sum[y_i\ log(p) + (1-y_i)log(1-p)]$$

Maximizing log-likelihood = Minimizing negative log-likelihood

?

Why MLE?

Principled probabilistic approach

Connection:

Links to cross-entropy loss function