Homework 5 for Machine learning: Theory and Algorithms

Gabriele Genovese

18 October 2024

1 Exercise 1

The relationship between empirical risk minimization and maximum likelihood estimation is expressed in equation (4). $P_w(S)$ represents the total probability of the dataset S given the model parameters w. This probability is equal to the mean of the probabilities of all items in the dataset. The probability of each item in the dataset is computed using the sigmoid function, which returns a probability value between 0 and 1. Therefore, the right loss function to use in those kind of problems is $\log(1 + \exp(-y_t w^T x_t))$.

2 Exercise 2

We need to show that the loss function $\log(1 + \exp(-y_t w^\top x_t))$ is convex in w. The logistic loss function is the composition of two convex functions: $\exp(-z)$ is convex, and $\log(1 + u)$ is convex for $u > 0$. Since convexity is preserved under composition, the logistic loss function is convex in w. The empirical risk is the average of convex functions (logistic losses over all samples), and the average of convex functions is also convex.

3 Exercise 3

In the re-labeled case, where $y_t \in \{0, 1\}$, the likelihood function $P_w(S)$ becomes:

$$P_w(S) = \prod_{t=1}^{m} h_w(x_t) y_t \cdot \left(1 - h_{w(x_t)}\right)^{1-y_t}$$

Taking the logarithm of both sides and the mean it becomes:

$$\frac{1}{m}\log(P_w(S)) = \frac{1}{m}\sum_{t=1}^{m} (y_t \log(h_w(x_t)) + (1 - y_t)\log(1 - h_w(x_t)))$$

So, the loss function on a single data point is:

$$\mathrm{Loss}(h_w(x_t), y_t) = y_t \log(h_w(x_t)) + (1 - y_t)\log(1 - h_w(x_t))$$

4 Exercise 4

The likelihood can be viewed as a measure of how well the model $h_w(x_t)$ estimates the probabilities of the true values y_t. Since y_t is binary (0 or 1), the contribution to the loss function changes depending on the value of y_t:

- If $y_t = 1$, the loss is $\log\left(\frac{1}{h_w}(x_t)\right)$, which penalizes the model's predictions if the probability of 1 is low.
- If $y_t = 0$, the loss is $\log\left(\frac{1}{1-h_w(x_t)}\right)$, which penalizes the predictions if the probability of 0 is low.

In this way, the negative likelihood becomes the sum of penalties for all incorrect predictions made by the model, and the loss function reflects the discrepancy between the model's predictions and the true values.