Homework 1

February 2022

1 Linear regression

Solve a simple linear regression problem using Mean Squared Error (MSE) loss. You need to find a linear function f between Forced Exhalation Volume (FEV) and age, *i.e.* $y = f(x)$. y denotes FEV and x denotes the age. The dataset (index.txt) consists of 654 children between 3 and 19 years old. The MSE loss is given as

$$Loss_{MSE} = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2, \tag{1}$$

where x_i and y_i denote FEV and age of the i-th training sample, respectively. N denotes the number of training samples.

Specifically, you can use 90% samples to be the training set, and 10% samples to be the test set. Please provide the predication error on the test set, *i.e.* $Error_{test} = \sum_{i=1}^{m}\left(y_i - f(x_i)\right)^2$ and m denotes the number of the testing samples. Note that the splited training set and the testing set should also be provided.

2 Logistic regression

Solve a binary classification problem using logistic loss. Consider a tabular classification task, the lower back pain symptoms dataset (Dataset_spine.csv) has 13 columns where the first 12 are the features and the last column is the target column ("Abnormal" or "Normal"). That is, you need to find a non-linear function f to infer from the 12 features whether the result is normal or not, *i.e.* $y = f(\mathbf{x})$, where $\mathbf{x}$ is a 12-dimensional vector, and $y \in \{0, 1\}$ is the boolean label.

The dataset has a total of 300 samples. You can use 80% samples to be the training set, and 20% samples to be the test set. The order of the samples can be shuffled to ensure that the training and test sets evenly contain "Abnormal" and "Normal" samples. And samples from the test set cannot appear in the training set. Please provide the curves of train loss, test loss, train accuracy and test accuracy. The logistic loss you need to implement is on page 4 of lecture 1.