Machine Learning **Universität Hamburg**
Spring 2024 Sören Laue

Exercise Sheet 4

Exercise 1

Consider measuring the length of a certain object. Since it is difficult to measure it exactly, there have been many attempts in measuring it. Assume that the correct length is l^* and the measurements $l^{(1)}, l^{(2)}, ..., l^{(n)}$ that have been obtained are the true length plus a noise term, i.e.,

$$l^{(i)} = l^* + \varepsilon$$

1. Assume that the noise term ε follows a Gaussian distribution with mean 0 and variance 1, i.e., $\varepsilon \sim N(0,1)$. Determine the maximum likelihood estimator (MLE) for this case. Provide the derivation of it and also provide a closed form solution for it.

2. Assume now that the noise term ε follows a Laplacian distribution $Laplace(\mu, b)$ with location parameter $\mu = 0$ and scale parameter $b = 1$. The probability density function of the Laplacian distribution $Laplace(\mu, b)$ is defined as

   $$\frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

   Again, compute the maximum likelihood estimator (MLE) for this case. Provide the derivation of it and also provide a closed for solution for it.

Which of the two estimators is more robust against outliers?

Exercise 2

Consider the following regression problem: You are given 40 data points each having 200 features and a real-valued response for each data point. Find a good linear regressor that explains the data well and at the same time is fairly sparse. You encounter such problems usually when dealing with gene expression data where you have few patients but many genes that might cause an illness.
Come up with a good solution for this problem and write some code to solve it. You are now allowed to use scikit-learn. You will find a training and a test data set `dataset_sparse_train.npy` and `dataset_sparse_test.npy`. Report your best regressor. What does best mean here for you?

Please turn in your solutions by Thursday, May 2nd.