Adv Big Data

Professor Gu

Xueyan Bai

Assignment 6: Summary of Distance Function

With the study in class, we learnt about distance functions such as L1 Regularization, Lasso and KL divergence. In this paper, the author mainly discussed about Kullback-Leibler Divergence.

KL divergence is a method of comparing two probability distribution. While we use simpler and similar distribution to replace the complex observations or distribution, KL divergence could help us to figure the loss of information after the replacement. For instance, if we have a dataset with unknown distribution, we need to reduce this data to a simple model with one or two parameters due to expense and difficulty of data transmission. There might have several distributions work. To find the best model, we need KL divergence to exam which model preserves the most information from data.

The process to quantify how much information is in data calls information theory. The most important metric in information theory is called Entropy (H). The definition of Entropy:

The author also mentioned a special notation , which represents as “the minimum of bits it would take us to encode our information”. To measure information lost using KL divergence, we have data probability distribution p and approximating distribution q:

From above, we could see that KL divergence is the expectation of the log difference between the original distribution and approximating distribution:

Since the larger indicates the greater loss of information from approximating distribution, we should choose the approximating distribution with smaller . Given the asymmetry, it is important to remember that KL divergence cannot be used to measure the distance between two distributions:

Also, while p(x) = q(x), . Otherwise, will be greater than 0 since is the strictly convex function. Since p(x) represents the probability of real data, will be a constant.

Thus, to minimize the cross entropy is to minimize the KL divergence, which is equivalent to find the maximum likelihood estimation.