Master's Thesis

Estimation of mixing ratios for a mixture distribution using the Wasserstein distance

Guidance

Professor Nobuo Yamashita
Program-Specific Associate Professor Hiroyuki Sato

Kohdai NAGASHIO

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University

Feburary 2024

Abstract

A mixture distribution is a weighted sum of distributions, and it is useful for analyzing data with different labels. For example, the distribution of overall human height can be considered a mixture distribution of human height at each age. In some applications of data science and machine learning, prior knowledge of the mixing ratio of data with different labels enables us to get better results. However, in practice, the true mixing ratio is generally not known in advance. There exist methods of to estimate the mixing ratio from distributions of explanatory variables for the data of interest and a distribution of overall data. The key of the estimation is how to measure a distance between these distributions. The estimation method based on a distance called the Pearson divergence has been proposed. It can estimate the ratio in a short time, but its estimation is not good enough.

In this thesis, we propose to use the Wasserstein distance for the estimation. When sample data of distributions are given, the problem of estimation with the Wasserstein distance is formulated as a linear optimization (LO) problem. Since the size of the LO is proportional to the number of sample data, it is difficult to solve it with the standard LO solvers for big data. Therefore, we develop an alternating direction method of multipliers (ADMM) which is suitable for the LO. The computational time of each iteration of the ADMM is proportional to the square of the number of data.

Finally, numerical experiments shows that our proposed method using the Wasserstein distance can estimate the mixing ratio more accurately than the method using the Pearson divergence. The experimental results show that the solution obtained by the ADMM method are consistent with those obtained by the LO method.