Abstract

Machine learning is a data processing technology to make an intellectual inference (e.g., regression analysis or classification) by automatically extracting important patterns from data. It has maintained reciprocal relation with empirical science because both machine learning and empirical science share a common interest in methodology to make a reasonable inference from limited information.

A salient example of the shared methodology is *matrix completion* (MC) that exploits the low-rankness often present in real-world matrix-shaped data. It allows one to retrieve a matrix from its partial information that is deficient due to various factors, e.g., missing values, noise, or discretization.

One crucial type of measurement limitations widely found in empirical science is the *ceiling effect*. It is a type of measurement limitations in which values are clipped at a predefined threshold before observation. The ceiling effect is also conceivable in the context of machine learning, e.g., in recommendation systems with a five-star rating. After rating an item with a five-star, a user may find another item much better later. In this case, the true rating for the latter item should be above five, but the recorded value is still a five-star.

In this thesis, I consider the recovery of a matrix-formed data from its observations affected by ceiling effects. However, the current theoretical guarantees for low-rank matrix completion do not apply to clipped matrices because whether clipping occurs depends on the underlying values. Therefore, the feasibility of *clipped matrix completion* (CMC) is not trivial. This thesis first provides a theoretical guarantee for the exact recovery of CMC by using a trace-norm minimization algorithm. Furthermore, practical CMC algorithms are proposed by extending ordinary MC methods. The extension is to use the squared *hinge* loss in place of the squared loss for reducing the penalty of over-estimation on clipped entries. Also, a novel regularization term tailored for CMC is proposed. It is a combination of two trace-norm terms, and the recovery error under the regularization is theoretically bounded. While the problem setting is motivated by potential applications in scientific research, the developed methodology is also beneficial to recommender systems, a machine learning application, as indicated by experiments.

This thesis is a step towards extending the frontier of matrix completion, and the developed methodology may pioneer potential applications. For example, a biological scientist who has to use measurement equipment prone to ceiling effect may measure multiple cells under multiple environmental conditions, organize the data into a matrix, and apply CMC to obtain a reasonable estimate of the true values. The present thesis demonstrates the possibility of the low-rank completion principle in constructing remedies for information deficits which may depend on the underlying values. Regarding the generality of the low-rank completion principle, it is also promising to consider matrix completion under other more complex measurement limitations. Further extensions to address different measurement limitations are left for future work.