

High-impact variable selection for clustering in multinomial logistic regression

Hai-Tang Chiou¹, Chun-Wei Dong², Chun-Ying Lin², and Shih-Feng Huang²

¹ Department of Mathematics, National Chung Cheng University

² Graduate Institute of Statistics, National Central University

Abstract

This study investigates high-impact variable selection for identifying clusters in high-dimensional multinomial logistic regression. For datasets with known class labels, variables are initially selected using an aggressive variant of the Chebyshev greedy algorithm, and then refined through a high-dimensional information criterion. For unlabeled data, pseudo-labels are generated via clustering, and the variable selection is applied iteratively, updating both labels and selected variables at each step. Simulation studies demonstrate that the proposed methods effectively identify variables relevant for distinguishing underlying classes.

Keyword: Chebyshev greedy algorithm, High-dimensional information criterion, High-dimensional variable selection