

Vom Umgang mit kategorialen Prädiktoren – Selektion und Clustering

Gerhard Tutz

Institut für Statistik
Seminar für angewandte Stochastik
Ludwig-Maximilians-Universität München

In den letzten Jahren wurden diverse Regularisierungsmethoden für metrische Prädiktoren entwickelt und untersucht. Kategoriale Prädiktoren wurden hingegen vergleichsweise selten betrachtet. Die Herausforderung besteht insbesondere in der Dimension, da kategoriale Prädiktoren schnell zu hoher Parameterzahl führen.

Es werden zuerst Regularisierungstechniken betrachtet, die Variablenselektion bewirken. Darüberhinaus stellt sich für kategoriale Prädiktoren das Problem, ob sich Kategorien im Hinblick auf die abhängige Variable unterscheiden, oder ob sie nicht besser zu Clustern zusammengefasst werden sollten. Es empfiehlt sich dabei zwischen ordinal-kategorialen und nominal-kategorialen Prädiktoren zu unterscheiden. Entsprechende Verfahren, die Clustering von Kategorien und Variablenselektion erreichen werden dargestellt und illustriert.