Probability machines for categorical outcome using machine learning methods

Andreas Ziegler[1,2], Jochen Kruppa[1],
Theresa Holste1[1]

1. Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany
2. Zentrum für Klinische Studien, Universität zu Lübeck, Germany

Abstract

Probability estimation for dichotomous and multicategory outcome using logistic and multinomial logistic regression has a long-standing tradition in biostatistics. Biases may occur if the model is misspecified. An alternative approach is the estimation of probabilities by machine learning methods, such as k nearest neighbors (k-NN), bagged nearest neighbors (b-NN), support vector machines (SVM) or random forests (RF). In this talk, we first describe the fundamental idea for the consistent estimation of probabilities using machine learning methods. We specifically show that probability estimation can be embedded into the class of nonparametric regression. Next, we introduce random RF, k-NN, b-NN and SVM and summarize their theoretical properties for dichotomous outcome. The algorithms naturally extend to multicategory outcome. In simulation studies we demonstrate the validity of the methods. The approaches are illustrated using real data sets.