Discriminative and Generative models for clinical risk estimation: An empirical comparison

John Stamford[1] and Chandra Kambhampati[2]

1 **Department of Computer Science, The University of Hull, Kingston upon Hull, HU6 7RX**
j.stamford@2014.hull.ac.uk

2 **Department of Computer Science, The University of Hull, Kingston upon Hull, HU6 7RX**
c.kambhampati@hull.ac.uk

Abstract

Linear discriminative models, in the form of Logistic Regression, are a popular choice within the clinical domain in the development of risk models. Logistic regression is commonly used as it offers explanatory information in addition to its predictive capabilities. In some examples the coefficients from these models have been used to determine overly simplified clinical risk scores. Such models are constrained to modeling linear relationships between the variables and the class despite it known that this relationship is not always linear. This paper compares the conditions under which linear discriminative and linear generative models perform best. This is done through comparing logistic regression and naïve Bayes on real clinical data. The work shows that generative models perform best when the internal representation of the data is closer to the true distribution of the data and when there is a very small difference between the means of the classes. When looking at variables such as sodium it is shown that logistic regression can not model the observed risk as it is non-linear in its nature, whereas naïve Bayes gives a better estimation of risk. The work concludes that the risk estimations derived from discriminative models such as logistic regression need to be considered in the wider context of the true risk observed within the dataset.

1998 ACM Subject Classification D.4.8 Performance

Keywords and phrases Discriminative, Generative, Naïve Bayes, Logistic Regression, Clinical Risk

Digital Object Identifier 10.4230/OASIcs.ICCSW.2017.5