Bayesian variable selection for mixture of nominal and ordinal responses via an indicator matrix

Eleftheria Kotti*, Ioanna Manolopoulou, Tom Fearn
Department of Statistical Science, University College London, United Kingdom
*e.kotti.12@ucl.ac.uk

September 25, 2015

Abstract

Selection of significant features/variables from Big (in this case high-dimensional) Data is an important problem in Statistics and Machine Learning, since including unnecessary variables adds noise to estimation and omitting them can improve prediction and classification performance. Motivated by a $p > n$ problem of spectral measurements of tissue from different stages in the progression of Barrett's oesophagus, where the stages of disease are a combination of nominal and ordinal categories, we develop a Bayesian variable selection approach. The aim is to identify the best individual variables and the best model in order to allow for accurate prediction in the general case of a mixture of ordinal and nominal responses.

We construct a probit model with latent variables that, in a single model, takes into account both types of responses, using one latent variable with a vector of thresholds for the ordinal responses and an additional latent vector for each nominal response. We assume that the covariance matrix of the distribution of the latent variables is known and it has a common variance across ordinal responses, a different variance across different nominal responses and zero covariances in both cases.

In order to perform variable selection for the multi-class multivariate case with both ordinal and nominal categories, different indicator vectors (indicating presence of the covariate in the linear combinaton making up the latent variable) are used across different latent variables. The advantage using different indicator vectors is that different variables may be important for identifying different stages of disease. We generate the indicator matrix using all those vectors.

From the joint posterior distribution we calculate conditional distributions and apply Gibbs sampling to allow for efficient inference. The conditional distribution on the indicator matrix does not have closed-form solution. At this step we use Metropolis-within-Gibbs.

We apply our methodology on simulated data, where the number of variables is bigger than the number of samples. We compare the classification accuracy of our method with existing ones from classical Statistics and Machine Learning scientific area. The proposed method can be applied, for example, in bioinformatics.

Keywords: Bayesian variable selection, Markov chain Monte Carlo, latent variables, nominal and ordinal responses, different indicator vector across different responses.

Acknowledgments: We want to thank the 'Foundation for Education and European Culture' for the financial support provided to Eleftheria Kotti.