Sparse Bayesian kernel projections for classification of near-infrared spectroscopy data

Katarina Domijan[1*]

1. Department of Mathematics and Statistics, NUI Maynooth, Maynooth, Ireland

*Contact author: Katarina.Domijan@maths.nuim.ie

Keywords: Bayesian inference, reproducing kernel Hilbert spaces, Bayesian decision theory, NIR spectroscopy

A Bayesian classification method is applied to near infrared spectroscopic data obtained from from several food authenticity studies. The datasets record spectra over the visible and near infra-red wavelength range for relatively small numbers of training samples. Typically the dimension of the variables exceeds ten to twenty - fold the number of samples, however, the reflectance measurements are sequentially highly correlated. The classifier is based on the reproducing kernel Hilbert spaces (RKHS) theory which allows for nonlinear generalization of linear classifiers by implicitly mapping the classification problem into a high dimensional feature space where the data are thought to be linearly separable. The proposed algorithm performs the classification of the projections of the data to the principal axes of the feature space. The projections are uncorrelated and sparse, leading to large computational savings and improved classification performance. The likelihood is modeled through the multinomial logistic regression model and the relatively standard hierarchical prior structure for Bayesian generalized linear models is assumed. The degree of sparsity is regulated in a novel framework based on Bayesian decision theory. The Gibbs sampler is implemented to find the posterior distributions of the parameters, thus probability distributions of prediction can be obtained for new data points, which gives a more complete picture of classification. The classifier obtains good classification rates for the datasets considered and does not require pre-processing steps to reduce their dimensionality. The results from several classification algorithms available in *R* programming language are included for comparison. A collection of *R* functions for classification and data visualization is available.