A Hierarchical Bayesian Approach to Regression and its Application to Predicting Survival Times in Cancer

Lars Kaderali

Abstract

A hierarchical Bayesian approach to regression is developed in this thesis, enabling regression in high dimensional spaces when only a low number of data points is available, and in addition, it can be assumed that most of the input measurements are irrelevant. The model is developed for a simple linear regression problem first, and then adapted to a nonlinear problem: The prediction of survival times of cancer patients from gene expression data.

Two methods of computation and prediction with the model are developed: A maximum-a-posteriori procedure based on gradient descent, and a Markov chain Monte Carlo approach. The latter not only enables the prediction of regression estimates and the assessment of the relevance of individual input measurements, but can also be used to compute confidence intervals and visualize distributions over both regression estimates and input relevance values.

The algorithms developed are then tested on both simulated and real world data. The methods are used to predict survival times of cancer patients from gene expression data, on publicly available datasets from two clinical studies, on diffuse large B-cell lymphoma and on adenocarcinomas of the lung. It is shown in comparison to other existing methods and to clinical staging that the methods developed yield improved predictions. A separate preprocessing step to reduce dataset dimensionality can be abandoned almost completely, and in contrast to many existing methods, true survival times in years and months are predicted for the individual patient, and not just two or more distinct risk classes. The probabilistic approach in addition enables the computation of confidence intervals on the predictions made.