Master thesis in Machine learning with applications in genetics
Random Forest, Gradient Boosting Trees and Bayesian Additive Regression Trees to detect complex genetics interactions

The purpose of this project is to compare three ensemble regression tree methods of their capabilities to detect complex epistatic interactions in large scale genomic data. The methods that will be compared are the Random Forest (RF; Breiman 2001), Gradient Boosting Trees (GBT; Friedman 2001) and Bayesian Additive Regression Trees (BART; Chipman et al. 2010).

Ensemble of regression trees are non-parametric statistical learning methods that have the important feature that they can handle all forms of genetic effects in a very sparse way. However, since ensemble regression tree methods are black-box approaches it is rather difficult to perform variable selection and evaluate the genetic mode of a certain marker. Permutation and bootstrap procedures have been suggested for variable selection, and these seems to work fine for the RF and GBT.

One of the tasks of this project will be to implement the permutation procedure for BART using high performance parallel based computing (Pratola et al. 2013). Partial dependence plots provides a useful approximation to visualize epistatic interactions between important variables in ensemble regression tree methods.

The other task will be to develop algorithms for application of partial dependence plots to all three methods. The performance of the methods will be tested on simulated data and real genomic data from animals. Implementation of algorithms will require some knowledge of R and C.

Contact Information:
Behrang Mahjani (Behrang.mahjani@it.uu.se)
Patrik Waldmann (patrik.waldmann@slu.se)
Salman Toor (salman.toor@it.uu.se)

References: