AUSHANG

FREIE UNIVERSITÄT BERLIN

Fachbereich Mathematik und Informatik

Promotionsbüro, Arnimallee 14, 14195 Berlin

DISPUTATION

Donnerstag, 6. Juni 2019, 14:00 Uhr

Ort: Seminarraum 1
(Max-Planck-Institut f. molek.Genetik, Ihnestraße 63, 14195 Berlin)

Disputation über die Doktorarbeit von

Frau Anna Ramisch

Thema der Dissertation:

Enhancer Prediction Based on Epigenomic Data

Thema der Disputation:

Bagging, Boosting and Decision Trees

Die Arbeit wurde unter der Betreuung von **Prof. Dr. M. Vingron** durchgeführt.

Abstract: Decision trees are prominent “off-the-shelf” methods to quickly approach a multitude of machine learning problems, traditionally classification tasks. They can handle different types of input data, e.g. categorical or numerical data, as well as missing values, outliers and irrelevant input features. The main idea is to entangle complicated feature interactions by creating a branch-like partition of the input feature space into smaller manageable subsets. Hence, decision trees are also easy to interpret, as long as the trees are not grown too deep. However, due to their hierarchical structures, classification errors at the top easily propagate down to all the following decisions. Thus, small changes in the data can lead to very different trees and unstable classification predictions. Furthermore, decision trees are prone to overfitting when grown too deeply. In my presentation, I will explain two strategies, bagging and boosting, of how to tackle the shortcomings of decision trees while maintaining most of their advantages. Bagging and boosting are so called ensemble methods which combine multiple weak learners to produce more reliable predictions. Bagging is a parallel ensemble approach meaning that it learns multiple classifiers in parallel and finally combines their prediction to a final classification vote. It is also called ‘Bootstrap Aggregation’ since it is based on multiple randomly subsampled training sets which build the basis of a set of decision trees. Applied on a new input sample, each bootstrap tree has a vote, and the final bagging class estimator is simply their majority vote. Boosting, on the other hand, is a sequential ensemble approach where a weak base learner, e.g. a decision tree, is optimised in an iterative procedure depending on its performance on the training set. In each iteration step, the composition of the training data is altered putting higher weights on wrongly classified samples which leads to a set of classifiers specialised on different parts of the training set. The final boosting estimator is a weighted sum of all the classifiers’ votes. In the second part of the presentation I will give a summary of the tree-based enhancer prediction method explained in my thesis entitled "Enhancer Prediction based on Epigenomic Data".

Die Disputation besteht aus dem o. g. Vortrag, danach der Vorstellung der Dissertation einschließlich jeweils anschließenden Aussprachen.

Interessierte werden hiermit herzlich eingeladen

Der Vorsitzende der Promotionskommission
Prof. Dr. M. Vingron