Master oder Bachelor Arbeit (Statistik, Data Science oder Informatik)
Thema: Regression und Klassifikation für sehr große Datenmengen

In einigen Bereichen der Wissenschaft oder der Wirtschaft haben wir es mit derartig großen Datensätzen zu tun, dass klassische Verfahren zur Regression oder Klassifikation an die Grenzen ihrer Effizienz stoßen. Daher müssen neue Verfahren entwickelt werden, welche die Analyse massiver Datensätze ermöglichen.

Um in einer solchen Situation weiterhin die theoretischen und praktischen Eigenschaften der klassischen Algorithmen zu bewahren, wollen wir uns in dieser Abschlussarbeit dem sogenannten "sketch-and-solve" Paradigma widmen.

Hierbei wird im ersten Schritt eine Datenreduktion erzielt (z.B. durch Subsampling von Datenpunkten) und anschließend ein klassischer Algorithmus zur Lösung des Problems auf dem reduzierten Datensatz ausgeführt. Die Herausforderung hierbei besteht darin, eine geeignete Reduktionsmethode zu entwickeln, sodass die gefundenen Lösung beweisbar nah an der optimalen Lösung liegt, welche wir durch Analyse des vollen Datensatzes erhalten hätten.

Mittels geschicktem Subsampling von Datenpunkten möchten wir eine sogenannte Kernmenge für das Probit-Modell entwickeln. Diese garantiert, dass die Likelihood sogar für jeden möglichen Parametervektor approximativ erhalten bleibt und nicht nur für den "optimalen" Maximum-Likelihood Schätzer. Eine Konsequenz hiervon ist, dass wir in der Bayesianischen Betrachtung auch die Verteilung der Parametervektoren punktweise approximieren können. Die Arbeit wird abgerundet durch eine empirische Evaluierung der Methode im frequentistischen sowie im Bayesianischen Fall.

Mathematische Statistik und biometrische Anwendungen / Dortmund Data Science Center
Ansprechperson: Dr. Alexander Munteanu

Master's or Bachelor's thesis (Statistics, Data Science or Computer Science)
Topic: Regression and classification of very large data

In several areas of research and in the industry we are faced with massively large data sets, such that classical methods for regression or classification reach their limits of efficiency. Therefore, new methods need to be developed for handling massive data sets and their statistical analysis.

In order to preserve the theoretical and practical properties of classical algorithms in such a situation, we will focus on the so-called "sketch-and-solve" paradigm in this thesis.

First, a data reduction is applied (e.g. by subsampling data points) before a classical algorithm is executed to solve the problem on the reduced data set. The challenge here is to develop a suitable reduction method so that the solution found is provably close to the optimal solution that we would have obtained by analyzing the full data set.

Our aim is to develop a so-called coreset for the probit model by sampling a small number of data points from a carefully designed importance sampling distribution. This will guarantee that the likelihood is approximated for every possible parameter vector, not only for the "optimal" maximum likelihood estimator. As a consequence this allows us to approximate the distribution of parameter vectors in the Bayesian setting. Finally, the thesis is rounded up by an empirical evaluation of the method in the frequentist as well as in the Bayesian case.

Chair for Mathematical Statistics with Applications in Biometrics / Dortmund Data Science Center
Contact: Dr. Alexander Munteanu