LMU München, Institut für Statistik, Seminar für angewandte Stochastik

Bachelor/Master-Seminar im Sommersemester 2014

Regularisierungstechniken in der Regression

Prof. Dr. G. Tutz;
M. Berger, M.Sc.; Dipl.-Stat. M. Oelker; Dipl.-Stat. W. Pössnecker;
G. Schauberger, M.Sc.; Dipl.-Stat. M. Schneider;

Bei Regressionsproblemen mit einer großen Zahl an erklärenden Größen und/oder hoher Korrelation zwischen diesen sind Standardverfahren der Statistik, wie etwa einfache Maximum-Likelihood-Schätzung, oft instabil oder gar nicht durchführbar. Abhilfe schaffen Regularisierungstechniken, die zum einen die Schätzung interessierender Parameter stabilisieren, und zum anderen oft die Interpretierbarkeit der Ergebnisse erhöhen. Zwei klassische Verfahren sind hier die sog. Ridge Regression (Hoerl & Kennard, 1970) sowie Lasso (Tibshirani, 1996). Während die Ridge Regression den Schwerpunkt auf die Stabilisierung der Schätzung (und folglich Verringerung des MSE) legt, ermöglicht Lasso zusätzlich Variablen-Selektion und trägt somit auch zur Interpretierbarkeit des sich ergebenden Modells entscheidend bei. Im Seminar wird ausführlich in die Thematik eingeführt, es werden sowohl klassische als auch neuere Ansätze der Regularisierung behandelt.

Das Seminar richtet sich an Studierende im Bachelor- und Masterstudiengang Statistik. Als Hintergrund-Literatur, in dem viele der Verfahren kurz skizziert sind, dient das im Netz verfügbare Buch:

> Hastie, T., Tibshirani, R. & Friedman, J. (2009): The Elements of Statistical Learning – Data Mining, Inference and Prediction, 2nd Edition, New York: Springer.

Weitere Literatur zu den einzelnen Themen wird im Seminar bekannt gegeben.

Das Seminar findet dienstags von 16 – 18 Uhr im Seminarraum des Instituts für Statistik, Ludwigstr. 33, statt. Die Anmeldung zum Seminar erfolgt über das LSF. Erster Seminar-Termin ist Dienstag, der 8.4.2014 (Einführung, Themenvergabe etc.). Weitere Informationen finden Sie auf der Seminar-Homepage:

`http://www.statistik.lmu.de/institut/lehrstuhl/semsto/Lehre/SeminarSS2014/indexSem.html`

Themenvorschläge und Literatur:

- Ridge Regression
  - Hoerl & Kennard (1970): Ridge regression: biased estimation for non-orthogonal problems, *Technometrics 12.*
  - Nyquist (1991): Restricted estimation of generalized linear models, *Journal of Applied Statistics 40.*
- Lasso
  - Tibshirani (1996): Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society B 58.*
  - Park & Hastie (2007): L1-regularization path algorithm for generalized linear models, *Journal of the Royal Statistical Society B 69.*
- Grouped Lasso
  - Yuan & Lin (2006): Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society B 68.*
  - Meier et al. (2008): The group lasso for logistic regression, *Journal of the Royal Statistical Society B 70.*
- Grouped Lasso für multinomialen Response
  - Tutz, Pößnecker & Uhlmann (2012): Variable Selection in General Multinomial Logit Models. *Department of Statistics: Technical Reports, No.126.*
- Fused Lasso
  - Tibshirani et al. (2005): Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society B 67.*
  - Gertheiss & Tutz (2010): Sparse modeling of categorial explanatory variables, *The Annals of Applied Statistics 4(4).*
- Elastic Net
  - Zou & Hastie (2005): Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society B 67.*
- Algorithmus-basierte Variablenselektion in Regression und Klassifikation
  - Friedman, J. H. (2012): Fast sparse regression and classification. *International Journal of Forecadting 28(3)*, 722-738.

- Trees and Random Forests
  - Breiman et al. (1984): *Classification and regression tress*, Wadsworth and Brooks, Monterey CA.
  - Breiman (2001): Random forests, *Machine Learning 45*.

- Sparse Boosting
  - Bühlmann & Yu (2006): Sparse boosting, *Journal of Machine Learning Research 7*.

- Likelihood-based Boosting
  - Tutz & Binder (2006): Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics, 62, 961-971*.

- Multivariates Boosting
  - Lutz & Bühlmann (2006): Boosting for high-multivariate responses in high-dimensional linear regression, *Statistica Sinica 16*.

- Lasso in Rasch Models
  - Tutz & Schauberger (2014): A Penalty Approach to Differential Item Functioning in Rasch Models. *Psychometrika*, to appear.

- Least Angle Regression
  - Efron et al. (2004): Least angle regression, *The Annals of Statistics 32*.

- SCAD
  - Fan & Li (2001): Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association 96*.

- Dantzig Selector
  - Candes & Tao (2007): The Dantzig selector: statistical estimation when p is much larger than n, *The Annals of Statistics 35*
  - James & Radchenko (2009): A generalized dantzig selector with shrinkage tuning, *Biometrika 96*.