Bachelorseminar im Sommersemester 2017

Regularisierte Schätzverfahren in Regressionsmodellen

Prof. Dr. M. Spindler;
Philipp Bach, M. Sc.; Sven Klaaßen, M. Sc; Jannis Kück, M. Sc.

In Regressionsproblemen mit einer großen Anzahl an Einflussgrößen ist es notwendig die relevanten Variablen bzw. Wirkungsstrukturen zu identifizieren, da die übliche Maximum-Likelihood-Schätzung oft nicht durchführbar oder zumindest nicht optimal ist. Das ist von besonderer Bedeutung bei sogenannten hochdimensionalen Datensätzen und bei der Analyse von "Big Data", die von immer größerer Bedeutung im Bereich Wirtschaftswissenschaften und betrieblicher Praxis werden.
Es werden unterschiedliche Ansätze betrachtet, wie sich die Prädiktorenstruktur identifizieren lässt und wie Regularisierungstechniken einsetzbar sind, um Parameter schätzbar zu machen und relevante Strukturen zu identifizieren. Ein wesentlicher Schwerpunkt ist die Variablenselektion. Viele der betrachteten Verfahren zielen auf eine Kombination von Variablenselektion und Verbesserung der Modellperformance ab.

Das Seminar richtet sich an Studierende im Bachelorstudiengang Betriebswirtschaftslehre.
Als Hintergrund-Literatur, in der 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 [1] .

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

Das Seminar findet als Blockveranstaltung vom 30.06-02.07.2017 statt. Die Anmeldung für das Seminar ist vom 02.01. bis zum 11.01.2017 zentral über STiNE möglich. Die Vorbesprechung und Vorbereitung der Themenvergabe ist für den 04.04.2017 von 16-18 Uhr im Raum 1068 angesetzt.[2] Voraussetzung für das Bestehen des Seminars ist die Teilnahme an der Blockveranstaltung. Der Leistungsnachweis wird erbracht durch eine Hausarbeit (ca. 13 Seiten) und einen Seminarvortrag von ca. 45 Minuten mit anschließender Diskussion und Beurteilung des Vortrags.

[1] `http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf`

[2] Zur Themenvergabe kann eine Liste mit den drei meist präferierten Themen mitgebracht bzw. abgegeben werden. So weit möglich, wird versucht, diese zu berücksichtigen.

Themenvorschläge und Literatur

- Rigde Regression
  - Hoerl & Kennard (1970): Ridge regression: biased estimation for nonorthogonal 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*.
- Einführung in GLMs / Lasso in GLMs
  - 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*.
- Fused Lasso
  - Tibshirani et al. (2005): Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society B 67*.
- Elastic Net
  - Zou & Hastie (2005): Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society B 67*.
- Lava
  - Chernozhukov, Hansen, and Liao (2016+): A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, forthcoming.
- 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*.
- Trees and Random Forests
  - Breiman et al. (1984): Classification and regression tress, Wadsworth and Brooks, Monterey CA.
  - Breiman (2001): Random forests *Machine Learning 45*.
- Bagging
  - Bühlmann & Yu (2003): Analyzing bagging, *The Annals of Statistics 30*, Number 4 (2002), 927-961.
  - Zhou (2012): Ensemble Methods: Foundations and Algorithms (Chapman Hall/CRC Data Mining and Knowledge Discovery Serie).

- Boosting
  - Friedman, J. (2001): Greedy function approximation: A gradient boosting machine. *Annals of Statistics 29*, 11891232.
  - Friedman, J.,Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *Annals of Statistics 28*, 337407.
- L_2-Boosting
  - Bühlmann & Yu (2003): Boosting with the L 2 loss: regression and classification, *Journal of the American Statistical Association 98 (462)*, 324-339.
  - Bühlmann & Hothorn (2007): Boosting Algorithms: Regularization, Prediction and Model Fitting, *Statistical Science*, 477-505.
- Stability Selection
  - Meinshausen & Bühlmann (2010): Stability Selection, *Journal of the Royal Statistical Society: Series B* 72(4), 417-473.
  - Bühlmann, Kalisch, Meier (2014): High-Dimensional Statistics with a View Toward Applications in Biology, *Annual Review of Statistics and its Applications 1.*
- Sure Independence Screening
  - Fan & Lv (2008): Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B* 70(5), 849-911.