Comprehensible models from high-dimensional or complex data

Niklas Lavesson, BTH

Introduction

- Transparency in predictive modeling is important in many domains
- Some domain experts are more likely to prefer models with
  - Clear convincing reasons for predictions
  - A small number of relevant features
  - A way to make decisions that emulate human decision making

Terminology

- Common terms that refer to similar topics
  - understandability, comprehensibility, interpretability, transparency
- Additional terms that may refer to such topics
  - interestingness, complexity, model size
- Can we identify semantic differences and, if so, what is the impact of using some semantically different terms interchangeably?

Motivation

- Comprehensible models may provide
  - machine learning researchers with a new source for knowledge about the studied algorithms
  - applied researchers with new knowledge about the domain in question
  - professionals with an increased sense of trust
- In comparison to other topics, and in relation to its potential impact, comprehensibility is an understudied topic in machine learning

Known Challenges

- Comprehensibility is, in most instances, a subjective and qualitative model property
- It may be difficult to establish universal or even widely accepted definitions
- Many comprehensibility measures are model-specific and, thus, cannot be used to compare different models of different type
- Experimental evaluation of comprehensibility is an open question (cf Freitas, 2014)

Potential Future Challenges

- Size and complexity increase
  - Problems, Data, Algorithms, Implementations
- As machine learning grow in popularity, the amount of non-expert users increase
  - The demand for more easily understandable models may increase as well
- What should be done in data science to tackle existing and future challenges?

Initial Discussion Points

- Could something be learned from other areas with similar challenges?
- Are some foci more promising than others?
  - Generic or domain-specific measures
  - Direct or indirect comprehensibility assessment
  - Measure to understand or to evaluate
- Suitable research study designs
- Factors that may impact comprehensibility

Factors that may impact model comprehensibility

- Data dimensionality and sparsity
- Integration of data from heterogeneous sources
- Noise and overlapping class boundaries
- Disagreement with existing knowledge

Selected Literature

- Chorowski, J. (2012). **Learning Understandable Classifier Models**. Doctoral thesis. University of Louisville, Kentucky.
- Freitas, A. (2013). **Comprehensible Classification Models: A Position Paper**. ACM SIGKDD Explorations Newsletter.
- Gleicher, M. (2014). **Towards Comprehensible Predictive Modeling**. IEEE VIS 2014 Workshop Visualization for Predictive Analytics.
- Piltaver, R., Lustrek, M., Gams, M., Martincic-Ipsic, S. (2014). **Comprehensibility of Classification Trees - Survey Design**. Proc. 6th International Conference on Information Technologies and Information Society.
- Rudin, C. (2014). **Algorithms for Interpretable Machine Learning**. Invited Talk. Data Mining for the Social Good. Proc. 20 ACM SIGKDD Conference on Knowledge Discovery & Data Mining.
- Tan, C. (2013). **More than Accuracy: Interpretability**. Talk at the MLDG group. Cornell University.
- Zhou, J., Wang, Y., Li, Z., Chen, F. (2013). **Transparent Machine Learning - Revealing Internal States of Machine Learning**. 18th International Conference on Intelligent User Interfaces.