T2: Structured Sparsity in Natural Language Processing: Models, Algorithms and Applications

André F. T. Martins, Mário A. T. Figueiredo, and Noah A. Smith

ABSTRACT

This tutorial will cover recent advances in sparse modeling with diverse applications in natural language processing (NLP). A sparse model is one that uses a relatively small number of features to map an input to an output, such as a label sequence or parse tree. The advantages of sparsity are, among others, compactness and interpretability; in fact, sparsity is currently a major theme in statistics, machine learning, and signal processing. The goal of sparsity can be seen in terms of earlier goals of feature selection and therefore model selection (Della Pietra et al., 1997; Guyon and Elisseeff, 2003; McCallum, 2003).

This tutorial will focus on methods which embed sparse model selection into the parameter estimation problem. In such methods, learning is carried out by minimizing a regularized empirical risk functional composed of two terms: a "loss term," which controls the goodness of fit to the data (e.g., log loss or hinge loss), and a "regularizer term," which is designed to promote sparsity. The simplest example is L1-norm regularization (Tibshirani, 2006), which penalizes weight components individually, and has been explored in various NLP applications (Kazama and Tsujii, 2003; Goodman, 2004; Gao, 2007). More sophisticated regularizers, those that use mixed norms and groups of weights, are able to promote "structured" sparsity: i.e., they promote sparsity patterns that are compatible with a priori knowledge about the structure of the feature space. These kind of regularizers have been proposed in the statistical and signal processing literature (Yuan and Lin, 2006; Zhao et al., 2009; Kim et al., 2010; Bach et al., 2011) and are a recent topic of research in NLP (Eisenstein et al., 2011; Martins et al, 2011, Das and Smith, 2012). Sparsity-inducing regularizers require the use of specialized optimization routines for learning (Wright et al., 2009; Xiao, 2009; Langford et al., 2009).

The tutorial will consist of three parts: (1) how to formulate the problem, i.e., how to choose the right regularizer for the kind of sparsity pattern intended; (2) how to solve the optimization problem efficiently; and (3) examples of the use of sparsity within natural language processing problems.

OUTLINE

1. Introduction

(30 minutes)

- What is sparsity?
- Why sparsity is often desirable in NLP
- Feature selection: wrappers, filters, and embedded methods
- What has been done in other areas: the Lasso and group-Lasso, compressive sensing, and recovery guarantees
- Theoretical and practical limitations of previous methods to typical NLP problems
- Beyond cardinality: structured sparsity

2. Group-Lasso and Mixed-Norm Regularizers

(45 minutes)

- Selecting columns in a grid-shaped feature space
- Examples: multiple classes, multi-task learning, multiple kernel learning
- Mixed L2/L1 and L_{inf}/L1 norms: the group Lasso
- Non-overlapping groups
- Example: feature template selection
- Tree-structured groups
- The general case: a DAG
- Coarse-to-fine regularization

3. Coffee Break

(15 minutes)

4. Optimization Algorithms

(45 minutes)

- Non-smooth optimization: limitations of subgradient algorithms
- Quasi-Newton methods: OWL-QN
- Proximal gradient algorithms: iterative soft-thresholding, forward-backward and other splittings
- Computing proximal steps
- Other algorithms: FISTA, Sparsa, ADMM, Bregman iterations
- Convergence rates

- Online algorithms: limitations of stochastic subgradient descent
- Online proximal gradient algorithms
- Managing general overlapping groups
- Memory footprint, time/space complexity, etc.
- The "Sparseptron" algorithm and debiasing

5. **Applications**

   (30 minutes):

   - Sociolinguistic association discovery
   - Sequence problems: named entity recognition, chunking
   - Multilingual dependency parsing
   - Lexicon expansion

6. **Closing Remarks and Discussion**

   (15 minutes)

BIOS

André F. T. Martins
Instituto de Telecomunicações, Instituto Superior Técnico
Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal,
and Priberam Informática
Al. Afonso Henriques, 41 - 2., 1000-123 Lisboa, Portugal
afm--AT--cs.cmu.edu

A. Martins is a final year Ph.D. student in Carnegie Mellon's School of Computer Science and the Instituto de Telecomunicações at Instituto Superior Técnico, where he is working on a degree in Language Technologies. Martins' research interests include natural language processing, machine learning, convex optimization, and sparse modeling. His dissertation focuses on new models and algorithms for structured prediction with non-local features. His paper "Concise Integer Linear Programming Formulations for Dependency Parsing" received a best paper award at ACL 2009.

Mário A. T. Figueiredo
Instituto de Telecomunicações, Instituto Superior Técnico
Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
mario.figueiredo--AT--lx.it.pt

M. Figueiredo is a professor of electrical and computer engineering at Instituto Superior Técnico (the engineering school of the Technical University of Lisbon) and his main

research interests include machine learning, statistical signal processing, and optimization. He recently guest co-edited a special issue of the IEEE Journal on Special Topics in Signal Processing devoted to compressive sensing (one of the central areas of research on sparsity) and gave several invited talks (and a tutorial at ICASSP 2012) on optimization algorithms for problems involving sparsity.

Noah A. Smith
School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
nasmith--AT--cs.cmu.edu

N. Smith is Finmeccanica associate professor in language technologies and machine learning at CMU. His research interests include statistical natural language processing, especially unsupervised methods, machine learning for structured data, and applications of natural language processing, including machine translation and statistical modeling of text and quantitative social data. He recently published a book, Linguistic Structure Prediction, about many of these topics; in 2009 he gave a tutorial at ICML about structured prediction in NLP.

ACKNOWLEDGMENTS

This tutorial was enabled by support from the following organizations:

- National Science Foundation (USA), CAREER grant IIS-1054319.
- Fundação para a Ciência e Tecnologia (Portugal), grant PEst-OE/EEI/LA0008/2011.
- Fundação para a Ciência e Tecnologia and Information and Communication Technologies Institute (Portugal/USA), through the CMU-Portugal Program.
- QREN/POR Lisboa (Portugal), EU/FEDER programme, Discooperio project, contract 2011/18501.