Bayesian Learning of Probabilistic Language Models

by

Andreas Stolcke

Abstract

The general topic of this thesis is the probabilistic modeling of language, in particular natural language. In probabilistic language modeling, one characterizes the strings of phonemes, words, etc. of a certain domain in terms of a probability distribution over all possible strings within the domain. Probabilistic language modeling has been applied to a wide range of problems in recent years, from the traditional uses in speech recognition to more recent applications in biological sequence modeling.

The main contribution of this thesis is a particular approach to the learning problem for probabilistic language models, known as *Bayesian model merging*. This approach can be characterize as follows.

- Models are built either in batch mode or incrementally from samples, by *incorporating* individual samples into a working model
- A uniform, small number of simple operators works to gradually transform an instance-based model to a generalized model that abstracts from the data.
- Instance-based parts of a model can coexist with generalized ones, depending on the degree of similarity among the observed samples, allowing the model to adapt to non-uniform coverage of the sample space.
- The generalization process is driven and controlled by a uniform, probabilistic metric: the Bayesian posterior probability of a model, integrating both criteria of goodness-of-fit with respect to the data and a notion of model simplicity ('Occam's Razor').

The Bayesian model merging framework is instantiated for three different classes of probabilistic models: Hidden Markov Models (HMMs), stochastic context-free grammars (SCFGs), and simple probabilistic attribute grammars (PAGs). Along with the theoretical background, various applications and case studies are presented, including the induction of multiple-pronunciation word models for speech recognition (with HMMs), data-driven learning of syntactic structures (with SCFGs), and the learning of simple sentence-meaning associations from examples (with PAGs).

Apart from language learning issues, a number of related computational problems involving probabilistic context-free grammars are discussed. A version of Earley's parser is presented that solves the standard problems associated with SCFGs efficiently, including the computation of sentence probabilities and sentence prefix probabilities, finding most likely parses, and the estimation of grammar parameters.

Finally, we describe an algorithm that computes n-gram statistics from a given SCFG, based on solving linear systems derived from the grammar. This method can be an effective tool to transform part of

the probabilistic knowledge from a structured language model into an unstructured low-level form for use in applications such as speech decoding. We show how this problem is just an instance of a larger class of related ones (such as average sentence length or derivation entropy) that are all solvable with the same computational technique.

An introductory chapter tries to present a unified view of the various model types and algorithms found in the literature, as well as issues of model learning and estimation.

Prof. Jerome A. Feldman, Dissertation Chair

Contents