Proceedings of HLT-NAACL
Tutorials , pg. 8
Edmonton, May-June 2003

Optimization, Maxent Models, and Conditional Estimation without Magic

Christopher Manning and Dan Klein
Stanford University

Abstract

This tutorial aims to cover the basic ideas and algorithms behind techniques such as maximum entropy modeling, conditional estimation of generative probabilistic models, and issues regarding the use of models more complex than simple Naive Bayes and Hidden Markov Models. In recent years, these sophisticated probabilistic methods have been used with considerable success on most of the core tasks of natural language processing, for speech language models, and for IR tasks such as text filtering and categorization, but the methods and their relationships are often not well understood by practitioners. Our focus is on insight and understanding, using graphical illustrations rather than detailed derivations whenever possible. The goal of the tutorial is that the inner workings of these modeling and estimation techniques be transparent and intuitive, rather than black boxes labeled "magic here".

The tutorial decomposes these methods into optimization problems on the one side, and optimization methods on the other. The first hour of the tutorial presents the basics of non-linear optimization, assuming only knowledge of basic calculus. We begin with a discussion of convexity and unconstrained optimization, focusing on gradient methods. We discuss in detail both simple gradient descent and the much more practical conjugate gradient descent. The key ideas are presented, including a comparison/contrast with alternative methods. Next, the case of constrained optimization is presented, highlighting the method of Lagrange multipliers and presenting several ways of translating the abstract ideas into a concrete optimization method. The principal goal, again, is to make Lagrange methods appear as intuitively natural, rather than as mathematical sleight-of-hand.

The second part of the tutorial begins with a presentation of maximum entropy models from first principles, showing their equivalence to exponential models (also known as loglinear models, and particular versions of which give logistic regression, and conditional random fields). We present many simple examples to build intuition for what maxent models can and cannot do. Finally, we discuss how to find parameters for maximum entropy models using the previously presented optimization methods. By this point in the tutorial, audience members should have a clear understanding of how to build a system for estimating maxent models. We conclude with a discussion of issues specific to the language technology domain, including conditional estimation of generative models, and the issues involved in choosing model structure (such as independence, label and observation biases, and so on). We also discuss methods of smoothing, focusing on how smoothing works differently for maxent models than for standard relative-frequency-based distributions.

The tutorial will run 3 hours, with a break in the middle. Participants will be assumed to know basic calculus and probability theory, and to have some exposure to models such as Naive Bayes and HMMs, but need only have a basic awareness of language technology problems.