Maximum Entropy Modeling

This page dedicates to a general-purpose machine learning technique called **Maximum Entropy Modeling** (MaxEnt for short). On this page you will find:

What is Maximum Entropy Modeling

In his famous 1957 paper, Ed. T. Jaynes wrote:
Information theory provides a constructive criterion for setting up probability distributions on the basis of partial knowledge, and leads to a type of statistical inference which is called the maximum entropy estimate. It is least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information.
That is to say, when characterizing some unknown events with a statistical model, we should always choose the one that has Maximum Entropy.

Maximum Entropy Modeling has been successfully applied to Computer Vision, Spatial Physics, Natural Language Processing and many other fields. This page will focus on applying Maxent to Natural Language Processing (NLP).

The concept of Maximum Entropy can be traced back along multiple threads to Biblical times. However, not until the late of 21st century has computer become powerful enough to handle complex problems with statistical modeling technique like Maxent.

Maximum Entropy was first introduced to NLP area by Berger, et al (1996) and Della Pietra, et al. 1997. Since then, Maximum Entropy technique (and the more general framework Random Fields) has enjoyed intensive research in NLP community.

Tutorials for Maximum Entropy Modeling

Here is an (incomplete) list of tutorials & introduction for Maximum Entropy Modeling.

- A Brief Maxent Tutorial
  Good online tutorial by Adam Berger
- A Simple Introduction to Maximum Entropy Models for Natural Language Processing
  This is an introductory paper by Adwait Ratnaparkhi. ftp download
- Maxent Models, Conditional Estimation, and Optimization, without the Magic
  A (not short) tutorial by Dan Klein and Chris Manning. This is really a good tutorial for Maxent modeling in NLP. However, I think it will be more readable if it was 50% shorter.
- Introduction to Natural Language Processing: Maximum Entropy
- The Maximum Entropy Method of Data Analysis

Maxent related software

Here is an incomplete list of software found on the net that are related to Maximum Entropy Modeling.

- maxent.sf.net Great java maxent implementation with GIS training algorithm. Part of OpenNlp project.
- Amis -- A maximum entropy estimator for feature forests. A maximum entropy estimator with GIS, IIS and L-BFGS algorithms.
- maxent Another Maximum Entropy Modeling Package with Ruby binding, GIS, Gaussian Prior smoothing and XML data format.
- Predictive Modeling Toolkit
- Robert Malouf's Maximum Entropy Parameter Estimation software, now available as Toolkit for Advanced Discriminative Modeling on sourceforge.net, has GIS, IIS, L-BFGS and Gradient Descent training methods and parallel computation ability through PETSc. You may want to read his paper first.
- YASMET -- Yet Another Simple Maximum Entropy Toolkit with Feature Selection
- YASMET(2) -- Yet Another Small MaxEnt Toolkit. Believe it or not, this implementation is written in only 132 lines of C++ code and still has feature selection and gaussian smoothing. You need GCC 2.9x to compile the source. link2
- MEGA Model Optimization Package. A recently appeared ME implementation by Hal Daumé III. The software features CG and LM-BFGS Optimization and is written in OCaml. Although I no longer use OCaml, I'd say that's a great language, and is worth learning.
- Text Modeller A python implementation of a **joint** Maximum Entropy model (aka. Whole Sentence Language Model) with sampling based training. Now seems to be part of scipy.
- Stanford Classifer is another open source implementation of Maximum Entropy Model in java, suitable for NLP tagging and parsing tasks.
- NLTK includes a maxent classifier written entirely in Python. IIS and GIS training methods available. Suitable for text categorization and related NLP tasks.
- Here is another small maxent package in C++ with a BSD-like license, written by Dekang Lin.
- SharpEntropy, a C# port of the java maxent package (http://maxent.sf.net) mentioned above.
- Maxent software for species habitat modeling by Robert E. Schapire et al. Registration needed for downloading.
- My (Yet another...) Maxent implementation in C++ with Python binding, GIS, L-BFGS and Gaussian Prior Smoothing

Annotated papers on Maximum Entropy Modeling in NLP

Here is a list of recommended papers on Maximum Entropy Modeling with brief annotation.

- A Maximum Entropy Approach to Natural Language Processing (Berger, et al. 1996)

  A must read paper on applying maxent technique to Natural Language Processing. This paper describes maxent in detail and presents an Increment Feature Selection algorithm for increasingly construct a maxent model as well as several examples in statistical Machine Translation.

- Inducing Features of Random Fields (Della Pietra, et al. 1997)

  Another must read paper on maxent. It deals with a more general frame work: *Random Fields* and proposes an *Improved Iterative Scaling* algorithm for estimating parameters of Random Fields. This paper gives theoretical background to Random Fields (and hence Maxent model). A greedy *Field Induction* method was presented to automatically construct a detail random fields from a set of atomic features. An word morphology application for English was developed. longer version.

- Adaptive Statistical Language Modeling: A Maximum Entropy Approach (Rosenfeld, 1994)

  This paper applies ME technique to statistical language modeling task. More specifically, it builds a conditional Maximum Entropy model that incorporates traditional N-gram, distant N-gram and trigger pair features. Significantly perplexity reduction over baseline trigram model was reported. Later, Rosenfeld and his group proposed a *Whole Sentence Exponential Model* that overcome the computation bottleneck of conditional ME model. You can find more on my SLM page.

- Maximum Entropy Models For Natural Language Ambiguity Resolution (Ratnaparkhi, 1998)

  This dissertation discusses the application of maxent model to various Natural Language Dis-ambiguity tasks in detail. Several problems were attacked within the ME framework: sentence boundary detection, part-of-speech tagging, shallow parsing and text categorization. Comparison with other machine learning technique (Naive Bayes, Transform Based Learning, Decision Tree etc.) was given. Ratnaparkhi also had a short introduction paper on ME.

- The Improved Iterative Scaling Algorithm: A Gentle Introduction

  This paper describes IIS algorithm in detail. The description is easier to understand than (Della Pietra, et al. 1997), which involves more mathematical notations.

- Stochastic Attribute-Value Grammars (Abney, 1997)

  Abney applies Improved Iterative Scaling algorithm to parameters estimation of Attribute-Value grammars, which can not be corrected calculated by ERF method (though it works on PCFG). Random Fields is the model of choice here with a general Metropolis-Hasting Sampling on calculating feature expectation under newly constructed model.

- A comparison of algorithms for maximum entropy parameter estimation (Malouf, 2003)

  Four iterative parameter estimation algorithms are compared on several NLP tasks. L-BFGS is observed to be the most effective parameter estimation method for Maximum Entropy model, much better than IIS and GIS. (Wallach 02) reported similar results on parameter estimation of Conditional Random Fields. Here is Malouf's Maximum Entropy Parameter Estimation software.

- A Mathematical Theory of Communication

  Claude Elwood Shannon's influential 1948 paper that laid the foundation of information theory and changed the whole world since then. I see no reason who has read the above papers does not want to read this one.

- Information Theory and Statistical Mechanics (Jaynes, E. T., 1957)
  Having read all the above papers? Well, it's time to have a look at this one. Edwin Thompson Jaynes presented some insightful results of maximum entropy principle in this 1957 paper published in *Physics Reviews*. This is also his first paper in information theory. Interestingly, this influential work was published over the objection of a reviewer.

Other recommended papers:

- Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data
- A Gaussian Prior for Smoothing Maximum Entropy Models (Chen and Rosenfeld, 1999)
- An Introduction to Conditional Random Fields for Relational Learning (Charles Sutton and Andrew McCallum, 2006)

Other MaxEnt related resources on the web

- Ed. T. Jaynes
  A collection of web pages devoted to the life of the great mathematician Ed. T. Jaynes, the pioneer of Maximum Entropy Modeling. Here is a photo of E.T. Jaynes.
- There is a Wiki Entry for Maximum Entropy Principle. Worth looking on.
- Looking for a package for **Conditional Random Fields**?
  Here's a CRF implementation in java, by Prof. Sunita Sarawagi. And the MALLET toolkit by Andrew McCallum also contains a class for training CRF. Also check out CRF++: Yet Another CRF toolkit written

by Taku Kudo in C++. Kevin Murphy has written many graphical related software in matlab, including a CRF toolbox.
