Spectral methods for estimating probabilistic language models

Lyle Ungar, Paramaveer Dhillon, Jordan Rodu, Michael Collins, and Dean Foster
University of Pennsylvania and Columbia University

NLP researchers have long computed the singular vectors of matrices of words and the documents or contexts they occur in, and used these singular vectors as low dimension representations of the words [1, 2]. When the immediate context of words–the words before and after each target word–are used, the projections of words onto their corresponding singular vectors forms the basis of a spectral method for estimating HMMs [3, 4].

More generally, the hidden states in a variety of directed dynamic Bayes nets, including HMMs and probabilistic grammars, can be estimated using spectral methods. These methods are attractive as they are very fast and scalable and provide globally consistent estimates of the model parameters. Spectral methods–unlike EM or Gibbs samples methods which, respectively, can get stuck in local minima or take arbitrarily long to converge–provide strong guarantees on computational and model estimation accuracy.

We have extended this class of spectral methods to learn probabilities of dependency trees and constituent grammars. We propose a simple yet powerful latent variable generative model for dependency parsing, and a spectral learning method to efficiently estimate it. As a pilot experimental evaluation, we use the marginal probabilities estimated by our model as features in a discriminative dependency parser, showing substantial benefit.

In this talk, we briefly present examples of reduced dimension spectral modeling for HMMs and dependency grammars. In both cases, we use spectral methods to learn low dimensional context-oblivious, and context-specific word representations from unlabeled data. These representation features can then be used with any supervised learner. For HMM models, we show the benefit of including these features in discriminative models for POS tagging, named entity recognition (NER) and chunking problems.

Example: HMM

To give the flavor of the power these models, we will now briefly describe how we estimate HMMs, and sketch the type of theorems that have been proved. We have developed similar models and proofs for dependency and constituent grammars.

All of these methods rely on using a matrix U which is, for example, the dominant singular vectors form the SVD between pairs of adjacent words in a large corpus (we often use the Google n-gram collection). We then project δ_{x_t}, the unit vector corresponding to observing word x_t as the tth item in a word sequence, down to a low dimension vector $y_t = U^\top \delta_{x_t}$.

The U matrix contains a vector for each word in the vocabulary, such that vectors for distributionally similar words are close to each other. To illustrate this, a 30-dimensional vector for each word was estimated by calculating the singular values of the (sparse) matrix containing the second (middle) word of each trigram in the Google n-gram collection versus the corresponding first and third words in the trigrams.

Since visualizing 30 dimensions is hard, we select small sets of words, and plot them in their first two principle components of the 30 dimensions. This fully unsupervised process gives some insight into what information is captured. Show below are two examples: 1) a random set of nouns and verbs, which are cleanly separated, and 2) a set of names, which can be seen to separate both in terms of male/female an in terms of formality.

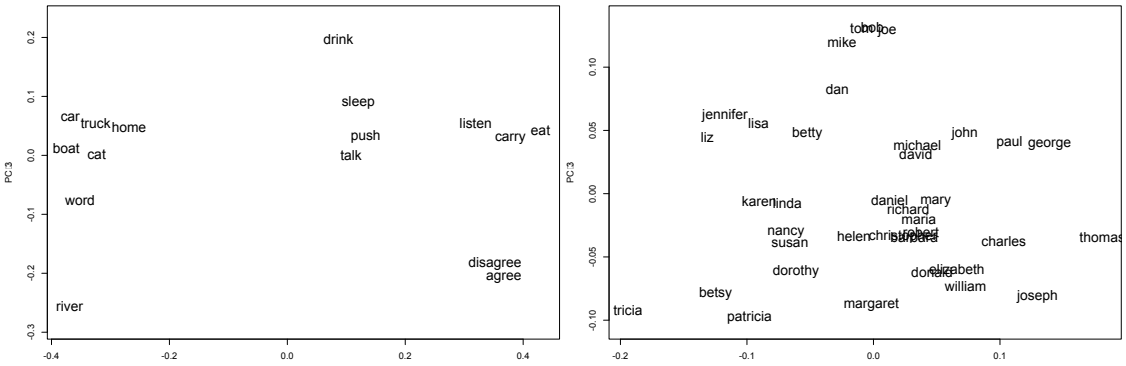

Figure 1: Projections onto two dimension of selected words in different categories using SVD on trigrams from Google n-grams.

We show that the probability of observing a word sequence

$$Pr(x_1, x_2, \ldots, x_t) = c_\infty^\top C_{y_t} C_{y_{t-1}} \cdots C_{y_1} c_1$$

holds where

$$\begin{aligned} c_1 &= \mu \\ c_\infty^\top &= \mu^\top \Sigma^{-1} \\ C_y \equiv C(y) &= K(y)\Sigma^{-1} \end{aligned}$$

and $\mu = E(y_1)$, $\Sigma = E(y_2 y_1^\top)$, and the tensor $K(a) = E(y_3 y_1^\top y_2^\top)a$ are easy to estimate using the method of moments.

Those familiar with the work of Hsu et al [ref] will note that our formulation uses the $k \times k \times k$ tensor C_y, rather than their $k \times v \times x$ tensor B_x, giving us a much more compact model when the vocabulary v is much bigger than the hidden state size k. Our lower dimensional model gives better complexity bounds, both in theory and in practice.

We prove theorems showing consistency of the method, and the rates of convergence of the distribution p defined by the method to the true distribution p. For example, the following theorem gives the finite sample bound in terms of a sample complexity:

Theorem 0.1. *Let X_t be generated by an $m \geq 2$ state HMM. Suppose we are given a U which has the property that range$(O) \subset$ range(U) and $|U_{ij}| \leq 1$ and suppose we use the above equation to estimate the probability based on N independent triples. Then*

$$N \geq \frac{128m^2(2t+3)^2}{\epsilon^2\ \Lambda^2 \sigma_m^4} \log\left(\frac{2m}{\delta}\right) \cdot \overbrace{\frac{\epsilon^2/(2t+3)^2}{(\sqrt[2t+3]{1+\epsilon}-1)^2}}^{\approx 1}$$

implies that

$$1 - \epsilon \leq \left| \frac{\widehat{\Pr}(x_1, \ldots, x_t)}{\Pr(x_1, \ldots, x_t)} \right| \leq 1 + \epsilon$$

holds with probability at least $1 - \delta$. where σ_m is the smallest singular value of Σ, and where Λ depends on properties of the model being estimated, and can be estimated from data.

References

[1] Landauer, T., Foltz, P., Laham, D.: An introduction to latent semantic analysis. In: Discourse processes. (2008)

[2] Turney, P., Pantel, P.: From frequency to meaning: vector space models of semantics. Journal of Artificial Intelligence Research **37** (2010) 141–188

[3] Hsu, D., Kakade, S.M., Zhang, T.: A spectral algorithm for learning hidden markov models. In: COLT. (2009)

[4] Dhillon, P., Foster, D., Ungar, L.: Multi-view learning of word embeddings via cca. In: NIPS. (2011)