A Weighted Mixture Approach to Text Prediction

STA 531 Final Project

Lei Qian
lei.qian@duke.edu

Wei (Emily) Shao
wei.shao@duke.edu

Leonardo Shu
leonardo.shu@duke.edu

Victor Yifan Ye
yifan.ye@duke.edu

May 6, 2016

Content

Introduction
Motivation

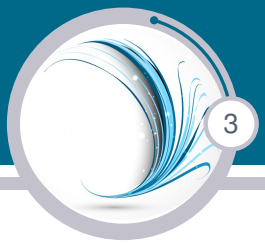
- Hidden Markov models (HMMs) have been used broadly in the predictive modeling of speech and text. However, achieving consistent predictions remains computationally costly.

Introduction
Motivation

- Hidden Markov models (HMMs) have been used broadly in the predictive modeling of speech and text. However, achieving consistent predictions remains computationally costly.
- Using the standard Baum-Welch (BW) algorithm As an example, a moderate-length input sequence of 20,000 words with 50 hidden states requires 10^8 calculations per iteration without taking M-step costs into consideration.

Introduction
Motivation

- Hidden Markov models (HMMs) have been used broadly in the predictive modeling of speech and text. However, achieving consistent predictions remains computationally costly.
- Using the standard Baum-Welch (BW) algorithm As an example, a moderate-length input sequence of 20,000 words with 50 hidden states requires 10^8 calculations per iteration without taking M-step costs into consideration.
- Subsetting the input sequence or utilizing a smaller number of hidden states will both significantly reduce the computational cost of Baum-Welch, yet at the cost of predictive performance.

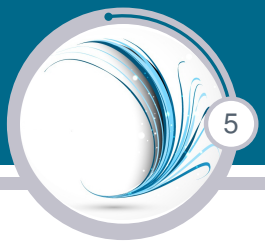


- The basic model for the new-word prediction problem is, for a new text chunk of size n, to draw the jth word x_j using a transition matrix $\mathbb{W}$ of size W x W where W equals the number of unique words in the text. The transition matrix has its probabilities calculated empirically by matching how often each words is next to all others and then standardizing throughout.
- Given the preceding word x^*_{j-1} and $\mathbb{W}$ with probability columns, the vector of probabilities given by the simple Monte Carlo model is the column of the unique word x_{i-1}, $V^{MC}_{x^*_{j-1}}$. We note that the length of $V^{MC}_{x^*_{j-1}}$ is the length of the total number of unique words in the text, with words that are never preceded by x^*_{j-1} assigned a probability of zero.

Goal: Find MLE of the parameters of the Hidden Markov Model.
Step 1: Expectation

- $$Q(\theta, \theta_k) = E_{\theta_k}(\log(P_\theta(x, z)|X = x))$$
$$= \sum_{i=1}^{m} P_{\theta_k}(Z_1 = i|x)\log(\pi_i)$$
$$+ \sum_{t=2}^{N}\sum_{i=1}^{m}\sum_{j=1}^{m} P_{\theta_k}(z_{t-1} = i, z_t = j|x)\log(T_{ij})$$
$$+ \sum_{t=1}^{N}\sum_{i=1}^{m} P_{\theta_k}(Z_t = i|x)\log(f_{\phi_i}(x_t))$$
$$\gamma_{ti} = P_{\theta_k}(Z_t = i|x)$$
$$\beta_{tij} = P_{\theta_k}(Z_{t-1} = i, Z_t = j|x)$$

$Q(\theta, \theta_k) = \sum_i^m \gamma_{1i}\log(\pi_i) + \sum_{t=2}^N \sum_{i=1}^m \sum_{j=1}^m \beta_{t_{ij}}\log(T_{ij}) +$
$\sum_{t=1}^N \sum_{j=1}^m \sum_{i=1}^m \gamma_{ti}\log(f_{\phi_i}(x_t))$

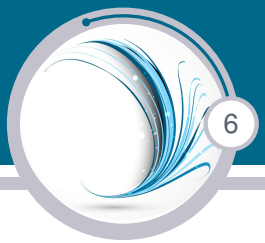


Step 2: Maximization

- $$\pi_i = \frac{\gamma_{1i}}{\sum_{j=1}^{m} \gamma_{1j}}$$

$$T_{ij} = \frac{\sum_{t=2}^{N} \beta_{tij}}{\sum_{t=2}^{N} \sum_{j=1}^{m} \beta_{tij}} = \frac{\sum_{t=2}^{n} \beta_{tij}}{\sum_{t=1}^{N-1} \gamma_{ti}}$$

$$\phi_{iw} = \frac{\sum_{x:x_t=w} \gamma_{ti}}{\sum_{w} \sum_{x:x_t=w} \gamma_{ti}}$$

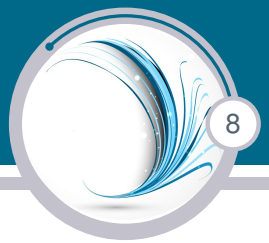


- 2^{nd} order Markov Chain where the probability of any given word (except the starting two) is dependent only on the previous pair.
- x_1^* and x_2^* fixed from the B-W algorithm output.
- Draw a vector of probabilities, $V^{MvS}_{x_{j-1},x_{j-2}}$ for each x_j^* where each element p_k of $V^{MvS}_{x^*_{j-1},x^*_{j-2}}$ is the probability of x_j^* being the unique word w_k given x^*_{j-1} and x^*_{j-2}
- Store the entire 3-dimension matrix $\mathbb{M}$ of size $W \times W \times W$ that consists of transition probabilities $(p_k, w_{x^*_{j-1}}, w_{x^*_{j-2}})$ for all unique 3-word sets

Toy Example

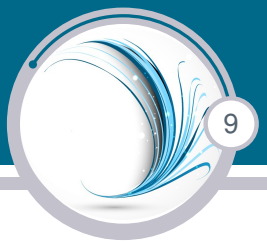
- $l = (l_1, l_2, l_3); \sum l = 1$
- Weighted Probability: $V^{mix}_{x^*_j} = l_1 * V^{MC}_{x^*_{j-1}} + l_2 * V^{HMM}_{z_j} + l_3 * V^{MvS}_{x^*_{j-1}, x^*_{j-2}}$
- Sequence of words $(A, B, C, E, D, A, B, B, E)$
- The unique three-word groups are:

$$\left\{ \begin{array}{l} \{A, B, C\} \\ \{B, C, E\} \\ \{C, E, D\} \\ \{E, D, A\} \\ \{D, A, B\} \\ \{A, B, B\} \\ \{B, B, E\} \end{array} \right\}$$



and unique two-word groups are:

- $$\left\{ \begin{array}{c} \{A, B\} \\ \{B, C\} \\ \{C, E\} \\ \{E, D\} \\ \{D, A\} \\ \{B, B\} \\ \{B, E\} \end{array} \right\}$$
- Suppose that the first two words have been selected as (A, B) via Baum Welch. Also suppose that for the hidden states z_3
- The vector of emission probabilities is given by $(p_A, p_B, p_C, p_D, p_E)'$. By Mark V. Shaney, the only two words that can follow from (A, B) are B and C. By simple Monte Carlo, the only words that follow B are B, C and E



- Emission probabilities given by B-W $(p_A, p_B, p_C, p_D, p_E)'$
- $V_{AB}^{MvS} = (0, 1/2, 1/2, 0, 0)'$
- $V_{B}^{MC} = (0, 1/3, 1/3, 0, 1/3)'$

$$V^{mix} = l_1 \cdot \begin{pmatrix} 0 \\ \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ 0 \end{pmatrix} + l_2 \cdot \begin{pmatrix} p_A \\ p_B \\ p_C \\ p_D \\ p_E \end{pmatrix} + l_3 \cdot \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} l_2 p_A \\ \frac{1}{3} l_1 + l_2 p_B + \frac{1}{2} l_3 \\ \frac{1}{3} l_1 + l_2 p_C + \frac{1}{2} l_3 \\ l_2 p_D \\ \frac{1}{3} l_1 + l_2 p_E \end{pmatrix}$$

	Jason Mraz	Adele
1	93 Million Miles	Chasing Pavements
2	Geek in the Pink	Hello
3	I Won't Give Up	Make You Feel My Love
4	I'm Yours	Rolling In The Deep
5	Life is Wonderful	Rumor has it
6	Make It Mine	Set Fire To The Rain
7	The Beauty in Ugly	Skyfall
8	The Remedy	Someone Like You
9	The Woman I Love	Turning Tables
10	Wordplay	When We Were Young

Table: Songs Names Used in the Analysis

Model
Prediction Output

The following is an example of our generated song lyrics for Adele.

"youd play . you played it , with a beating . throw your soul through every open door woah . youre gonna wish you never had met me . the scars of your despair . tears are gonna fall , rolling in the deep . we could have had it all . youre gonna wish you never had met me . it all . but my knees were far too tingle pavements when need your both build with but million shy blue you at im rolling reminded thinking made it you something all it . let the sky fall . when it crumbles . we will stand tall . i could stay there , close my eyes , feel you here forever . you had my heart inside of your love , they leave me breathless . i wish nothing but the best for you i ive made up my mind , dont need to do , if im wrong i am right , dont need to think it over , i must have called a thousand times . to make you feel my love . theres a fire starting in my heart drops , and my back begins to tingle ."

Results

- coherent sentences with high readability
- certain phrases are common
- certain words may have a limited amount of neighboring words
- ex: 'set' may be seen in phrase 'set fire to the rain'
- this is reflected in model

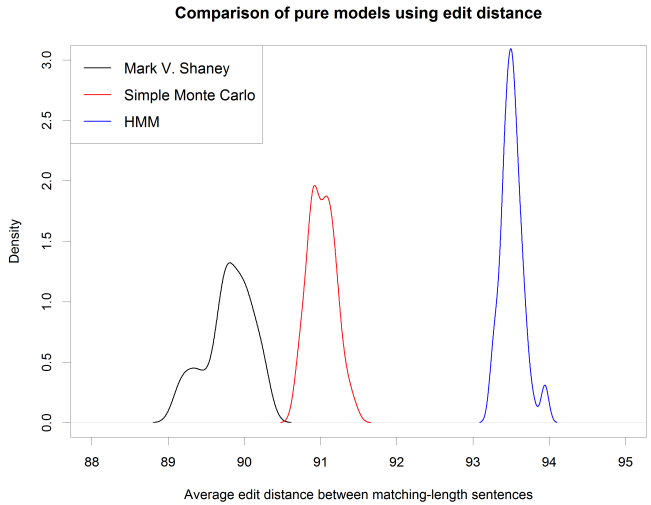
Advantages

- Computational complexity
- Versatility
  - third-order and fourth-order transition matrices
  - truncated HMM emission vectors
  - weights by word type
- (Desirable) Subjectiveness

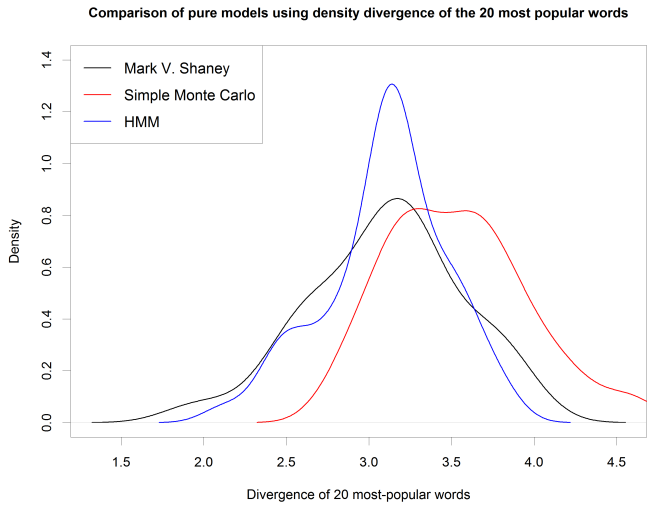
Training

- Indicators for readability
  - Zipf's Law - comparing distributions of most-frequent words
  - Levenshtein Distance
- Training algorithm: Metropolis

Training



(a) Edit Distance



(b) Word Frequency Divergence

Figure: Comparison of Models

Training

- Potential problem: we understand readability much better than the computer
- Potential solution: introducing a notion of grammar?

Applications

- Parody: DeepDrumpf
- Replication of text mining studies - avoiding copyright violations!
- Affordable procedural text generation: computer games?

Thank you for watching!