EXPECTATION MAXIMIZATION FOR ELECTION PREDICTION IN A HMM-LIKE PROBABILISTIC GRAPHICAL MODEL.

PATRICK VARIN AND ARJUN ALETTY

This paper is dedicated to our adviser, Paul Ruvolo.

Abstract. In this paper we present a probabilistic graphical model (PGM) similar to the Hidden Markov Model in which state transitions are also affected by previous observations. We also develop the Expectation Maximization algorithm specific to this PGM, an analogue to the Baum-Welch algorithm for the HMM. Finally we present this new PGM in the context of of election prediction and we show the performance of the EM algorithm in a series of experiments.

1. Introduction

The Markov condition is often a good descriptor of a variety of dynamical models. More likely than not, however, we cannot gather complete information from the system in question, in these cases a Hidden Markov Model is often used to analyze the system. Often times these systems are parametrized by a series of unknowns. These unknown parameters can be inferred, however, using the Expectation Maximization algorithm, which chooses the set of parameters that maximizes the likelihood of the observed dataset a famous application of this is in speech recognition, in which a speaker makes a series of utterances, the observed dataset, with the intent of constructing a word, the unknown parameter. The specific implementation of Expectation Maximization for HMMs is called the Baum-Welch algorithm.

There is a set of problems, however, in which the observations can influence the transition to the next state. In an election, for example, popular opinion can be influenced by the outcome of a poll. If an election cycle is considered a series of events in discrete time, the true public opinion can be considered the latent variable that influences the outcome of polls, the observed variables. There are a variety of parameters involved in this model, for example certain polls may have more influence on public opinion and certain polls may have some inherent bias towards or against different candidates.

In this paper we will discuss the computational methods necessary to analyze the graphical model discussed above. We will also revisit this example of using Expectation Maximization to predict the outcome of an election based on poll data.

Date: May 8, 2014.

2000 *Mathematics Subject Classification.* Primary 54C40, 14E20; Secondary 46E25, 20C20.

Key words and phrases. Probabalistic Graphical Model, Election Prediction, Expectation-Maximization, HMM, Baum-Welch.

2. The Model

The hidden Markov Model is given by a series of hidden states $\mathbf{Z} = \{Z_t\}$, each of which emits an observable variable $\mathbf{X} = \{X_t\}$. According to the Markov condition state transitions rely only on the previous state, and observations rely only on the immediate state. As a result the model can be described by three probability distributions, the initial state probability $P(Z_1)$, the transition probability $P(Z_t|Z_{t-1})$ and the emission probability $P(X_t|Z_t)$.

For problems in which there is no prior knowledge of the form of these probability distributions it is common to tabulate each of the probabilities in three matrices: $\pi = \{\pi_i\}$, $A = \{\alpha_{ij}\}$, $B = \{\beta_{ij}\}$. The initial state probability is given by $P(Z_1 = i) = \pi_i$, the transition probability is given by $P(Z_t = j|Z_{t-1} = i) = \alpha_{ij}$ and the emission probability is given by $P(X_t = j|Z_t = i) = \beta_{ij}$ The Expectation-Maximization algorithm, then attempts to compute each element in these tables. This specific implementation of Expectation-Maximization is called the Baum-Welch algorithm.

The model that we introduce here introduces a new dependency in the state transitions. In this new model these transitions rely not only on the previous state, but also the previous observation. This changes the form of the transition probability to $P(Z_t|Z_{t-1}, t_{x-1})$.

3. Expectation-Maximization

In some applications, in which the form of these probability distributions is unknown, it may make sense to parametrize the probability distributions in a similar manner to the HMM described above. In general, however, the structure of Expectation-Maximization is independent of the choice of parametrization and we will use the variable θ to describe a general set of parameters, e.g., the parametrization used in the Baum-Welch algorithm is $\theta = (\pi, A, B)$.

The Expectation-Maximization procedure is generally separated into two steps, the E-step and the M-step

The E-Step. The goal of the E-Step is to formulate a fuction, $Q(\theta, \theta')$, the computes the expected value of the log-likelihood of parameters with respect to the probabilities of the latent variables given the last best guess of the parameters.

$$Q(\theta, \theta') = \mathbf{E}_{\mathbf{Z}|\mathbf{X},\theta'}\left[\log P(\mathbf{X}, \mathbf{Z}|\theta)\right]$$

Computing the joint probability $P(\mathbf{X}, \mathbf{Z}|\theta)$ directly is often not computationally feasible. Using the causal structure of the model, however, we can simplify the joint probability as follows:

$$\begin{aligned}
P(\mathbf{X}, \mathbf{Z}|\theta) &= P(X_{1:T-1}, Z_{1:T-1}|\theta)P(X_T, Z_T|X_{1:T-1}, Z_{1:T-1}, \theta) \\
&= P(X_1, Z_1|\theta)\prod_{t=2}^{T} P(X_t, Z_t|X_{1:t-1}, Z_{1:t-1}, \theta) \\
&= P(X_1, Z_1|\theta)\prod_{t=2}^{T} P(X_t, Z_t|X_{t-1}, Z_{t-1}, \theta) \\
&= P(Z_1|\theta)P(X_1|Z_1, \theta)\prod_{t=2}^{T} P(Z_t|X_{t-1}, Z_{t-1}, \theta)P(X_t|Z_t, \theta)
\end{aligned}$$

or, in grouping like terms (initial, transition, and emission probabilities) it can be expressed more clearly as

$$P(Z_1|\theta)\prod_{t=2}^{T}P(Z_t|X_{t-1},Z_{t-1},\theta)\prod_{t=1}^{T}P(X_t|Z_t,\theta)$$

The log-likelihood can therefore be expressed as

$$\log P(\mathbf{X},\mathbf{Z}|\theta)=\log P(Z_1|\theta)+\sum_{t=2}^{T}\log P(Z_t|X_{t-1},Z_{t-1},\theta)+\sum_{t=1}^{T}\log P(X_t|Z_t,\theta)$$

which yields the expression for the expected value of the log-likelihood

$$\begin{aligned}\mathbf{E}_{\mathbf{Z}|\mathbf{X},\theta'}\left[\log P(\mathbf{X},\mathbf{Z}|\theta)\right]=&\sum_{\mathbf{Z}}P(\mathbf{Z}|\mathbf{X},\theta')P(Z_1|\theta)+\\ &\sum_{\mathbf{Z}}P(\mathbf{Z}|\mathbf{X},\theta')\sum_{t=2}^{T}\log P(Z_t|X_{t-1},Z_{t-1},\theta)+\\ &\sum_{\mathbf{Z}}P(\mathbf{Z}|\mathbf{X},\theta')\sum_{t=1}^{T}\log P(X_t|Z_t,\theta)\end{aligned}$$

The sum over the set of all possible state sequences $\mathbf{Z}$ is exponentially large. However, each of these terms can be simplified to reduce the computational complexity. For example, the first term can be simplified as follows:

$$\begin{aligned}\sum_{\mathbf{Z}}P(\mathbf{Z}|\mathbf{X},\theta')P(Z_1|\theta)&=\sum_{Z_{1:T}}P(Z_1|\mathbf{X},\theta')P(Z_{2:T}|Z_1,\mathbf{X},\theta')P(Z_1|\theta)\\ &=\sum_{Z_1}P(Z_1|\mathbf{X},\theta')P(Z_1|\theta)\sum_{Z_{2:T}}P(Z_{2:T}|Z_1,\mathbf{X},\theta')\\ &=\sum_{Z_1}P(Z_1|\mathbf{X},\theta')P(Z_1|\theta)\end{aligned}$$

Similarly, the second term simplifies as:

$$\sum_{\mathbf{Z}}P(\mathbf{Z}|\mathbf{X},\theta')\sum_{t=2}^{T}\log P(Z_t|X_{t-1},Z_{t-1},\theta)=\sum_{Z_{t-1:t}}\sum_{t=2}^{T}P(Z_{t-1:t}|\mathbf{X},\theta')\log P(Z_t|X_{t-1},Z_{t-1},\theta)$$

and the third term simplifies as:

$$\sum_{\mathbf{Z}}P(\mathbf{Z}|\mathbf{X},\theta')\sum_{t=1}^{T}\log P(X_t|Z_t,\theta)=\sum_{Z_t}\sum_{t=1}^{T}P(Z_t|\mathbf{X},\theta')\log P(X_t|Z_t,\theta)$$

These simplifications leave the problem computationally tractable, however it introduces two terms that we do not yet know how to compute: $P(Z_t|\mathbf{X},\theta')$ and $P(Z_{t-1:t}|\mathbf{X},\theta')$. The general approach here is to use the forwards-backwards algorithm.

To begin, we use Bayes rule and note the conditional independence of $X_{t+1:T}$ and $X_{1:t-1}$ given Z_t and X_t

$$P(Z_t|\mathbf{X},\theta')=\frac{P(Z_t|X_{1:t},\theta')P(X_{t+1:T}|Z_t,X_t)}{P(X_{t+1:T}|\theta')}$$

Letting α_t and β_t represent the forwards and backwards components, respectively we get $\alpha_t(i) = P(Z_t = z_i|X_{1:t}, \theta')$ and $\beta_t(i) = P(X_{t+1:T}|Z_t = z_i, X_t)$. In order to compute α_t we expand as follows

$$\begin{aligned}
\alpha_t(i) &= P(Z_t = z_i|X_{1:t}, \theta') \\
&\propto P(Z_t = z_i, X_t|X_{1:t-1}, \theta') \\
&= \sum_j P(Z_t = z_i, Z_{t-1} = z_j, X_t|X_{1:t-1}, \theta') \\
&= \sum_j P(Z_t = z_i, X_t|Z_{t-1} = z_j, X_{t-1}, \theta')P(Z_{t-1} = z_j|X_{1:t-1}, \theta') \\
&= P(X_t|Z_t = z_i, \theta') \sum_j P(Z_t = z_i|Z_{t-1} = z_j, X_{t-1}, \theta')\alpha(j)_{t-1}
\end{aligned}$$

Using the new variables $\psi_i = P(X_t|Z_t = z_i, \theta')$ and $\Psi_{i,j}(X_{t-1}) = P(Z_t = z_j|Z_{t-1} = z_i, X_{t-1}, \theta')$ we can define α_t consicely as

$$\alpha_t = \psi \odot \Psi^T \alpha_{t-1}$$

where $\odot$ is an element-wise vector product.

The M-Step. The purpose of the M-Step is to optimize the log-likelihood function formulated in the E-Step.

$$\theta'_{new} = \arg\max_\theta Q(\theta, \theta')$$

For some problems this optimization can be computed analytically. Often times though gradient descent, or some other numerical optimization method is used.

Olin Way, Olin College, Needham, Massachusetts 02492
E-mail address: patrick.varin@students.olin.edu

Olin Way, Olin College, Needham, Massachusetts 02492
E-mail address: arjun.aletty@students.olin.edu