Online Learning of Inertial Hidden Markov Models

George D. Montañez
Carnegie Mellon University
Pittsburgh, PA USA
gmontane@cs.cmu.edu

Saeed Amizadeh
Yahoo Labs
Sunnyvale, CA USA
amizadeh@yahoo-inc.com

Nikolay Laptev
Yahoo Labs
Sunnyvale, CA USA
nlaptev@yahoo-inc.com

Abstract

Building on recent advances in probabilistic temporal regularization for hidden Markov models, we develop an online learning extension for the inertial HMM framework, allowing for scaling to arbitrarily large datasets. In addition, we develop a robust delayed online prediction method, controlling for the trade-off between optimal and timely state prediction. Our method is tested on synthetic and real-world datasets, showing the effectiveness of our learning and prediction algorithms.

1 Introduction

2 Problem Statement

3 Inertial Hidden Markov Models

Montañez *et. al.* [2015] recently introduced the inertial HMM solution for learning temporally regularized hidden Markov models for segmentation and characterization of multivariate time series. The inertial HMM is a K-state hidden Markov model with a modified likelihood function that causes increased state persistence as a direct consequence of maximizing the likelihood function. Because of this, the inertial HMM allows for simple expectation maximization [Dempster *et al.*, 1977] training of the model.

3.1 Likelihood Function Modifications

Inertial HMMs propose redefining the likelihood function in one of two ways. The first is by applying a modified self-transition Dirichlet prior to the state transition matrix, which is scaled in proportion to sequence length in order to maintain consistent strength of regularization. This is referred to as the *MAP inertial HMM* by the authors. The second is to use pseudo-observations for temporal regularization, where the observations are added to the joint complete data likelihood through use of a set of binary indicator random variables. This second method is referred to as the *inertial pseudo-observation HMM*. The two methods lead to distinct, yet related, mathematical forms for the likelihood function and both allow for learning via expectation maximization. Furthermore, the authors report both methods to have similar performance on the tested datasets. Therefore, we extend the conceptually simpler MAP inertial HMM. Since the inertial regularization methods rely on standard EM learning, one can naturally incorporate online EM learning techniques into such systems.

3.2 Update Equation

For the MAP inertial HMM, the scale-free update equation for the state transition matrix A is defined as

$$A_{jk} = \frac{((T-1)^{\zeta}-1)\mathbb{1}(j=k) + \sum_{t=2}^{T}\xi(z_{(t-1)j}, z_{tk})}{((T-1)^{\zeta}-1) + \sum_{i=1}^{K}\sum_{t=2}^{T}\xi(z_{(t-1)j}, z_{ti})}, \tag{1}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, T the length of the time series and $\xi(z_{(t-1)j}, z_{tk}) = \mathbb{E}[z_{(t-1)j}z_{tk}]$. This modified update equation is what distinguishes the inertial HMM from a standard HMM, and thus requires derivation of a novel online update equation. We provide the required equations in the next section.

4 Online Learning of Inertial HMMs

We extend the work of Stenger *et al.* [2001] and Montañez *et al.* [2015] to provide an online learning algorithm for the regularized MAP inertial hidden Markov model, which allows scaling to arbitrarily large datasets. Theoretical justification for incremental online EM learning is given in [Neal and Hinton, 1999].

4.1 Parameter Update Equations

Define

$$D_{T,i} := ((T-1)^{\zeta}-1) + \sum_{t=2}^{T}\sum_{k=1}^{K}\xi(z_{(t-1)i}, z_{tk}).$$

The recurrence for $D_{T,i}$ is then formulated as

$$D_{T,i} = D_{(T-1),i} + [(T-1)^{\zeta} - (T-2)^{\zeta}] + \sum_{k=1}^{K}\xi(z_{(T-1)i}, z_{Tk})$$

where T is the current time-step. Since T is both the current and final time-step, we have $\beta(z_{T,k}) = 1$ for $k = 1, \ldots, K$,

and thus

$$\begin{aligned}\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) &= P(\mathbf{z}_{t-1}, \mathbf{z}_t|\mathbf{X}) \\ &= \frac{\alpha(\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t;\phi)p(\mathbf{z}_t|\mathbf{z}_{(t-1)})\beta(\mathbf{z}_t)}{p(\mathbf{X})} \\ &= \frac{\alpha(z_{(t-1)i})p(\mathbf{x}_t;\phi_j)A_{ij}^{(T-1)}}{\sum_{k=1}^{K}\alpha(z_{tk})}\end{aligned}$$

where

$$\alpha(z_{tj}) = \left[\sum_{i=1}^{K}\alpha(z_{(t-1)i})A_{ij}^{(t-1)}\right]p(\mathbf{x}_t;\phi_j).$$

An efficient online update equation for the regularized transition matrix is then given by

$$\begin{aligned}A_{ij}^{(T)} = \frac{D_{(T-1),i}}{D_{T,i}}A_{ij}^{(T-1)} + \frac{\xi(z_{(T-1)i}, z_{Tj})}{D_{T,i}} \\ + \frac{\mathbb{1}(i=j)[(T-1)^{\varsigma} - (T-2)^{\varsigma}]}{D_{T,i}}\end{aligned}$$

Given that $\beta(z_{T,k}) = 1$, we have

$$\gamma(z_{tk}) = \frac{\alpha(z_{tk})}{\sum_{i=1}^{K}\alpha(z_{ti})}$$

for the incremental update. The corresponding incremental update equations for a Gaussian emission model (as reported in [Stenger *et al.*, 2001]) are

$$\mu_j^{(T)} = \frac{\sum_{t=1}^{T-1}\gamma(z_{tj})}{\sum_{t=1}^{T}\gamma(z_{tj})}\mu_j^{(T-1)} + \frac{\gamma(z_{Tj})}{\sum_{t=1}^{T}\gamma(z_{tj})}\mathbf{x}_T$$

and

$$\begin{aligned}\mathbf{S}_j^{(T)} = \frac{\sum_{t=1}^{T-1}\gamma(z_{tj})}{\sum_{t=1}^{T}\gamma(z_{tj})}\mathbf{S}_j^{(T-1)} \\ + \frac{\gamma(z_{Tj})}{\sum_{t=1}^{T}\gamma(z_{tj})}\left(\mathbf{x}_T - \mu_j^{(T)}\right)\left(\mathbf{x}_T - \mu_j^{(T)}\right)'\end{aligned}$$

where $(\cdot)'$ denotes the matrix transpose operation and $\mathbf{S}_j$ is the covariance matrix for state j.

4.2 Initialization

The process begins by batch-learning initial parameter estimates from a small portion of the time-series. These estimates are used for $\mathbf{A}^{(1)}$, $\mu^{(1)}$, $\mathbf{S}^{(1)}$ and $\pi(\mathbf{z}_t)$. For the α values, we initialize $\alpha(z_{1j}) = \pi(z_{1j})p(\mathbf{x}_1;\phi_j)$ for each j. Using Equation 1 and the definition of $D_{T,i}$, we compute $D_{2,i} = \sum_{j=1}^{K}\xi(z_{1i}, z_{2j})$, and $A_{ij}^{(2)} = \xi(z_{1i}, z_{2j})/D_{2,i}$.

The estimates are then updated for each new observation, using the update equations given above. Algorithm 1 outlines the order in which the various terms are computed.

4.3 Robust Online Prediction

We now consider the problem of online prediction. If an observation at time t (the current time step) is an outlier, we cannot know whether the model should remain in the same hidden state, treating the outlier as an anomaly, or transition to a new hidden state. To overcome this limitation, we propose delayed prediction of state labels using a sliding window of length w. As the window moves through the observation sequence, the Viterbi algorithm is performed on the section of data within the window and a prediction for the second observation is output. The first observation is used to represent all past history, via the Markov property, and the remainder of the window allows for "future" observations to affect "past" observations, via the backtracking maximization performed by the Viterbi algorithm. We can begin to output delayed state label predictions as soon as w observations arrive.

The length of the sliding window controls the trade-off between optimal state prediction (which occurs when w equals the length of all future and past observations) and the need for timely predictions. This parameter can be set using cross-validation when labeled state data is available.

Algorithm 1 Incremental Learning

```
Batch learn initial parameter estimates.
Compute D_{2,i} and A_{ij}^{(2)} for all i, j.
for all T > 2 do
    Compute α values for observation at time T.
    Compute ξ(z_{(T-1)i}, z_{Tj}) values for all i, j.
    Compute γ(z_{Tj}) and D_{T,i} values for all i, j.
    Update A_{i,j}^{(T)} using incremental update rule.
    Update μ_j^{(T)} and S_j^{(T)} using incremental update rules.
end for
```

5 Experiments

5.1 Datasets

Our synthetic data is generated from a two-state three-dimensional hidden Markov model with transition matrix

$$\mathbf{A} = \begin{pmatrix} 0.9995 & 0.0005 \\ 0.0005 & 0.9995 \end{pmatrix},$$

having equal start probabilities and emission parameters equal to $\mu_1 = (-1,-1,-1)^\top$, $\mu_2 = (1,1,1)^\top$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \text{diag}(3)$. Using this model, we generated one hundred time series of length 100,000.

The second dataset we constructed using real-world human accelerometer data [Altun *et al.*, 2010], collected using Xsens MTx™ units attached to the torso, arms and legs of human volunteers, resulting in forty-five dimensional signals. The signals were recorded for volunteers performing five different activities, such as playing basketball, jumping, walking on a flat surface, rowing and ascending stairs. The signals consist of accelerometer, gyroscope and magnetometer data, which we consider as a single 45D multivariate time series.

From this human activity data we generated one hundred multivariate time series, with varying number of segments and varying activities, using a five-state HMM with 90% probability of self-transition, 2.5% probability of non-self-transition (equal for all states), equal start probability and emissions generated by using the actual sensor data in serial fashion for the five activities, modulo the length of the stream. One hundred time series of 100,000 time ticks were generated in this manner.

5.2 Experimental Methodology

5.3 Results

6 Discussion

7 Related Work

8 Conclusions

Acknowledgments

GDM is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1252522. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors alone and do not necessarily reflect the views of the National Science Foundation or any other organization.

References

[Altun *et al.*, 2010] Kerem Altun, Billur Barshan, and Orkun Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recogn.*, 43(10):3605–3620, October 2010.

[Dempster *et al.*, 1977] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[Montañez *et al.*, 2015] George D Montañez, Saeed Amizadeh, and Nikolay Laptev. Inertial Hidden Markov Models: Modeling change in multivariate time series. *The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*, 2015.

[Neal and Hinton, 1999] Radford M. Neal and Geoffrey E. Hinton. Learning in graphical models. chapter A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants, pages 355–368. MIT Press, Cambridge, MA, USA, 1999.

[Stenger *et al.*, 2001] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J.M. Buhmann. Topology free hidden Markov models: application to background modeling. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 294–301 vol.1, 2001.