Advanced Probabilistic Machine Learning and Applications

Pablo Sánchez and Isabel Valera

Tutorial 6: Hierarchical Dirichlet Hawkes Process (HDHP)

In this tutorial we will cluster the Twitter dataset used in previous tutorial sessions using the Hierarchical Dirichlet Hawkes Process (HDHP) (Mavroforakis et al. , 2017), a modeling framework for clustering continuous-time grouped streaming data. In order to do so, it combines the hierarchical Dirichlet process (HDP) and multidimensional Hawkes process (HP).

This document contains a summary of the notation necessary to understand the model and the description of the exercises proposed to get a fully understanding of the model. On the other hand, the implementation in Python of the HDHP is available in the Github repository of the course.

Introduction

Notation: Through this document we will use the following notation:

- L: total number of learning patterns.
- K_u: number of tasks of user u.
- $K = \sum_u K_u$: total number of tasks.
- $\varphi_l = \{\alpha_l, \boldsymbol{\theta}_l, \pi_l\}$: parameters of a learning pattern l. For each of the learning patterns, the parameter π_l represents the popularity among users in learning parttern l, $\boldsymbol{\theta}_l$ is the parameter of the mark distribution, and α_l controls the self-excitation (or burstiness) of the underlying Hawkes process.
- $e := (u, t, z, \mathbf{x})$: an event, i.e., a tweet, represented by a user u, timestamp t, latent variable for the table assignment z, and the content/mark (in our case "bag of words") $\mathbf{x}$.
- $H_u(t)$: history of events generated by user u.

Submission: Copy the Jupyter notebook and code folder available in the Github repository `https://github.com/APMLA/apmla_material/tree/master/L6` and complete the exercises proposed below. You will need to submit electronically the complete version of the Jupyter (together with the future exercises for Block I) by December 13th.

Exercise 1: Explore and understand the HDHP implementation

- Find the piece of code where the μ_u parameters are updated. If you cannot find it, explain the reason.
- Find the piece of code where the α_l parameters are updated. If you cannot find it, explain the reason.
- Find the piece of code where the $\boldsymbol{\theta}_l$ parameters are updated. If you cannot find it, explain the reason.
- Find the piece of code where the $z_{1:n}$ table's assignment variables are updated. If you cannot find it, explain the reason.

- Explore the code and explain how the final particle is chosen at the last iteration. Recall, we run the SMC with $|P|$ particles but at the end we only consider one sample per parameter/hidden variable to show the results.

Exercise 2: Coding task

- Implement the code to build the dataset in the events format. See the jupyter notebook for more information.
- Compute the log-likelihood of the training set.
- From the HDHP code provided extract the necesary information about the learning patterns to fill the following table in which each row refers to a learning pattern.

Learning pattern table				
l	m_l	π_l	α_l	words
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Hint: $\pi_l = \frac{m_l}{K}$

- From the HDHP code provided extract the necesary information about the users to fill the following table in each each row refers to a user.

Users table					
u	# of tasks	μ_u	# of patterns	patterns	$\{\pi_l\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Hint: $\pi_l = \frac{m_{ul}}{K_u}$

References

C. Mavroforakis, I. Valera and M. Gomez-Rodriguez. Modeling the dynamics of learning activity on the web. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.