A Bayesian Point Process Model for User Return Time Prediction in Recommendation Systems

Sherin Thomas
Indian Institute of Technology
Hyderabad, India
sarasherinthomas@gmail.com

P. K. Srijith
Indian Institute of Technology
Hyderabad, India
srijith@iith.ac.in

Michal Lukasik*
University of Sheffield
m.lukasik@sheffield.ac.uk

ABSTRACT

In order to sustain the user-base for a web service, it is important to know the return time of a user to the service. We propose a Bayesian point process, log Gaussian Cox process (LGCP), to model and predict return time of users. It allows encoding the prior domain knowledge and non-parametric estimation of latent intensity functions capturing user behaviour. We capture the similarities among the users in their return time by using a multi-task learning approach. We show the effectiveness of the proposed approaches on predicting the return time of users to *last.fm* music service.

CCS CONCEPTS

• **Information systems** → **Information systems applications**;
• **Computing methodologies** → **Machine learning**;

KEYWORDS

Return time, log-Gaussian Cox Process, recommendation systems

1 INTRODUCTION

Web services such as recommendation systems benefit a lot from learning the temporal dynamics of users. Modeling the return time of users is of great interest to these web services. It allows them to understand user engagement and provides an early feedback on user experience with the service. It helps the web service to devise measures to retain the customer base such as targeted marketing to those users who may not return soon.

In this work, we address the problem of modeling the temporal dynamics of music listening patterns of users, specifically their return times to the *last.fm* web service. Recently, this was modeled using recurrent neural networks with survival loss function in [3]. A hazard based approach based on Cox proportional Hazard model [4] considered features such as active weeks and visit number, while

---

*Michal Lukasik is now at Google.

recurrent activities of users were captured using self exciting point process in [2].

We propose to use a Bayesian point process model which could encode prior knowledge on user behaviour such as periodicity in return times. Specifically, we use log Gaussian Cox Process (LGCP) [6], which models user behaviour through a stochastic intensity function. It is genrally difficult to represent the latent intensity function modeling the user's complex temporal behaviour in point processes. LGCP learns the intensity function non-parametrically from the data by modeling it using a Gaussian Processes (GP) [8] prior. We also capture the similarities across the users using a multi-task learning approach based on GPs [1]. We compare the performance of the proposed approaches with various GP Kernels against several baselines and demonstrate their usefulness on the data from the *last.fm* music web service. We find that the RBF kernel which is often used in temporal applications is outperformed by alternatives[5].

Notations: We assume there are M users with each user associated with a set of sessions. Let $\mathbf{t}_m = \{t_n^m\}_{n=1}^{N_m}$ denote the sessions associated with a user m where t_n^m denotes the start time of the nth session and N_m the number of sessions. The task is to predict the next return time of a user given the previous session start times.

2 MODEL

User interest to start a new session and its duration changes over time. This can be captured using a point process with an intensity function $\lambda(t)$. Users exhibit complex temporal behavior and it is difficult to come up with an appropriate intensity function capturing their temporal behavior. This motivates us to use log Gaussian Cox Process (LGCP), where the logarithm of the time varying intensity function is assumed to come from a Gaussian Process prior [6]. This allows us to learn the intensity function non-parametrically from the data in addition to specifying the domain knowledge through the GP kernel. The intensity function for a user m at a session starting time t_n^m is defined as $\lambda^m(t_n^m) = exp(f^m(t_n^m))$, where $f^m(t) \sim \mathcal{GP}(\mu^m(t), cov^m(t, t')$, $\mu^m(\cdot)$ is the mean function, and $cov^m(\cdot, \cdot)$ is the covariance function of a GP for a user m. The covariance function is specified through a positive semi-definite kernel $k^m(t, t')$, which determines the properties of the intensity function such as periodicity and smoothness.

We capture the similarities existing across different users in order to learn better intensity functions for a user. This is achieved using a multi-task LGCP model which learns the user similarities from their temporal activities through a user covariance matrix B [1, 5]. The multi-task learning kernel is defined jointly over users and times as

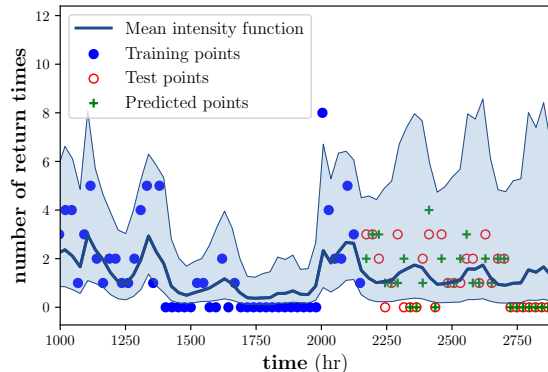$$k_{MTL}((m, t), (m', t')) = B_{m, m'} \cdot k(t, t') \tag{1}$$

Figure 1: Intensity function learnt for a user using LGCP-MULTITASK model with $RQ+Periodic$ **kernel on** *last.fm* **data. The x axis denotes the time and the y axis denotes the number of user returns within a** $24h$ **interval. The dark line denotes the predictive mean and the shaded region denotes the predictive variance. Note that we sample the predictions rather than using the predictive mean.**

The various model hyper-parameters such as the kernel parameters and the covariance matrix B are learnt by maximizing the marginal likelihood and a predictive distribution is obtained over the intensity function [5]. The return times of users are predicted using a mean intensity function and using the Ogata's thinning algorithm [7].

3 EXPERIMENTAL EVALUATION

Experimental setup We consider publicly available *last.fm* data which comprises of the music listening log of 992 unique users, with a total of 19,150,868 listening time stamps spanning from 2004 to 2009. For each user, the listening time stamps are split into different sessions if two consecutive time stamps differ by a gap of 1 hour or more [3]. The pre-processed session data is split into training and testing set by taking 3 consecutive months of data for training and 1 month data for testing purpose. Users with less than 100 listening events in training set and 50 events in test test are considered inactive and hence removed. This results in 394 active users with 243 sessions on an average per user. We consider various kernels in the LGCP model including the rational quadratic (RQ), periodic, and radial basis function (RBF) [8]. The RBF kernel models smooth functions, RQ models more complex functions with multiple length-scales and periodic kernels model the periodicity in the latent function. We also use LGCP with multitask learning kernel which captures similarities across users.

Evaluation Metrics In order to evaluate the model, we use mean absolute error (MAE) and root mean square error (RMSE) between the actual and the predicted times for each user in hours. [5] Since the data varies in size for each user, we take the micro average of the errors to obtain the final result.

Baselines The proposed models are compared against homogeneous Poisson process (HPP), linear regression, GP regression (RBF kernel), and recurrent neural networks (RNN) with one hidden layer and 50 neurons.

Method	Kernel	MAE	RMSE
LGCP	Periodic	9.37	19.45
	Rat Quad	8.68	18.74
	Periodic + Rat Quad	9.22	20.48
	RBF	15.89	22.44
LGCP-Multitask	Periodic	8.89	19.25
	Rat Quad	8.69	18.76
	Periodic + Rat Quad	8.90	19.07
	RBF	15.87	22.41
HPP		9.41	22.02
Linear Regression		10.25	22.56
GP Regression	RBF	10.30	22.98
RNN		11.05	20.46

Table 1: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between the actual and predicted user return time for proposed methods and baselines on *last.fm* **data.**

Results In Table 1 we compare the predictive performance of LGCP and LGCP-Multitask approaches against various baselines. We find that the standard kernel used in GP models, the RBF kernel, performs poorly due to the complex temporal patterns exhibited by users in their session start times. The RBF kernel typically models smoothly varying functions and is not suitable to model this situation. The RQ kernel could model such complicated behaviour patterns better than RBF by considering multiple length scales. This is evident from the experimental results where it outperforms other kernels and baselines. The periodic kernel could model the periodicity in the data (for instance, users tend to be more active on weekends) and are found to perform better than RBF but fails to capture other complex behavioral patterns captured by RQ. This is overcome by combining RQ with Periodic which improved the performance. The LGCP models with these kernels (except RBF) outperformed the baseline approaches such as HPP, linear regression, GP regression, and RNN. LGCP-Multitask brought significant improvements in performance with RBF, periodic and $RQ+Periodic$ kernels, while it retained the performance with RQ kernel. Figure 1 plots the intensity function of a user, learnt using LGCP-Multitask with $RQ + Periodic$ kernel. We found that considering other users temporal activity pattern through LGCP-Multitask could improve the predictive performance for a user.

REFERENCES
[1] Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. 2012. Kernels for Vector-Valued Functions: A Review. *Found. Trends Mach. Learn.* 4, 3 (2012).
[2] Nan Du, Yichen Wang, Niao He, and Le Song. 2015. Time-sensitive Recommendation from Recurrent User Activities. In *NIPS*. 3492–3500.
[3] How Jing and Alexander J. Smola. 2017. Neural Survival Recommender. In *WSDM*. 515–524.
[4] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava, and Tao Ye. 2014. A Hazard Based Approach to User Return Time Prediction. In *KDD*. 1719–1728.
[5] Michal Lukasik, P. K. Srijith, Trevor Cohn, and Kalina Bontcheva. 2015. Modeling Tweet Arrival Times using Log-Gaussian Cox Processes. In *EMNLP*. 250–255.
[6] Jesper Møller, Anne R. Syversveen, and Rasmus P. Waagepetersen. 1998. Log Gaussian Cox Processes. *Scandinavian Journal of Statistics* 25, 3 (1998), 451–482.
[7] Yosihiko Ogata. 1981. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory* 27, 1 (1981), 23–30.
[8] Carl E. Rasmussen and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.