Modeling Tweet Arrival Times using Log-Gaussian Cox Processes

Debolena Basak - AI20RESCH11003

Abstract

1. This paper focuses on predicting when the next tweet event will occur.
2. It is modeled and predicted using **Log Gaussian Cox Process(LGCP)**.
3. LGCP - an inhomogeneous Poisson Process which captures the varying rate at which the tweets arrive over time.
4. Incorporating textual features further improves predictions.

Motivation

1. Modeling the temporal dynamics of tweets provides useful information about the evolution of events.
2. Inter-arrival time prediction is a type of such modeling and has application in many settings like real-time disaster monitoring, journalism(tracking rumours) and advertising on social media.
3. Modeling the inter-arrival time of tweets is a challenging task due to complex temporal patterns exhibited.
4. Tweets associated with an event stream arrive at different rates at different points in time.
5. So the authors address the inter-arrival time prediction problem with **log-Gaussian Cox process(LGCP)**.

Introduction

1. **LGCP - an inhomogeneous Poisson process(IPP)** which models tweets to be generated by an underlying intensity function which can vary across time.
2. The **intensity function** assumes a **non-parametric form** which allows the model complexity to depend on the data set.
3. They have also **considered the textual content** of tweets to model inter-arrival times.
4. Dataset - The model is evaluated using the twitter rumours from the **2014 Ferguson unrest**.
5. It has been demonstrated that the proposed model provides good predictions for inter-arrival times, beating the baselines.

Poisson Process

Definition

A **counting process** $\{N(t), t \geq 0\}$ is said to be a **Poisson process** with rate $\lambda > 0$ if:

1. N(0) = 0 and N(t) = 0,1,2,...
2. The process has independent increments
3. $\Pr(N(t+h) - N(t) = 1) = \lambda h + o(h)$
4. $\Pr(N(t+h) - N(t) \geq 2) = o(h)$

Poisson Distribution

A counting random variable X is said to have a **Poisson Distribution** if it's PMF is:

$$\Pr(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, k = 0, 1, 2, ... \quad (1)$$

where $\lambda > 0$

Inhomogeneous Poisson Process(*IPP*)

Definition

A **counting process** $\{N(t), t \geq 0\}$ is said to be an **Inhomogeneous Poisson Process** with intensity function $\lambda(t), t \geq 0$, if:

1. N(0) = 0 and N(t) = 0,1,2,...
2. The process has independent increments
3. $\Pr(N(t+h) - N(t) = 1) = \lambda(t)\,h + o(h)$
4. $\Pr(N(t+h) - N(t) \geq 2) = o(h)$

The intensity function $\lambda(t)$ of an Inhomogeneous Poisson Process is a deterministic function. The intensity function is a function of time t. Example, arrival of e-mails in a day, arrival of customers in a shop.

Gaussian Process

Definition:

For any set S, a **Gaussian Process** (GP) on S is a set of R.V's $\{Z_t : t \in S\}$, such that, $\forall n \in N, \forall t_1, t_2, ...t_n \in S$, $(Z_{t_1}, Z_{t_2}, ..., Z_{t_n}) \sim$ Multivariate Gaussian.

Illustration:

Let $\mathbf{f}(\mathbf{x})$ be a function following Gaussian Process with mean function $\mu(\mathbf{x})$ and covariance function $\mathbf{k}(\mathbf{x}, \mathbf{x}')$.

$$\mathbf{f}(\mathbf{x}) \sim \text{Gaussian Process } \left(\mu(\mathbf{x}), \mathbf{k}\left(\mathbf{x}, \mathbf{x}'\right)\right) \tag{2}$$

$$\implies \begin{pmatrix} f(x_1) \\ f(x_2) \\ . \\ . \\ . \\ f(x_n) \end{pmatrix} \sim \textit{Gaussian} \left(\begin{pmatrix} \mu(x_1) \\ \mu(x_2) \\ . \\ . \\ . \\ \mu(x_n) \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & .... & k(x_1, x_n) \\ k(x_2, x_1) & .... & k(x_2, x_n) \\ . & .... & . \\ . & .... & . \\ . & .... & . \\ k(x_n, x_1) & .... & k(x_n, x_n) \end{pmatrix} \right) \tag{3}$$

Cox Process

Definition

An **Inhomogeneous Poisson Process** $\{X(t), t \geq 0\}$ where the rate function $\{\lambda(t), t \geq 0\}$ is itself a stochastic process.

1. Also known as **doubly stochastic Poisson Process**.
2. Named after Sir David Cox.
3. The process increments over disjoint intervals are, in general, **statistically dependent** in a Cox process.

MODEL: Log-Gaussian Cox Process

1. The intensity function is assumed to be **stochastic**.
2. The intensity function $\lambda(t)$ is modeled using a latent function $f(t)$ sampled from a **Gaussian process**.
3. To ensure **positivity** of the intensity function, $\lambda(t) = \exp(f(t))$ is considered.
4. This provides a **non-parametric Bayesian approach** to model the intensity function, where the complexity of the model is learnt from the training data.
5. The functional form of the intensity function can be defined through appropriate **GP priors**.

Modeling Inter-arrival Time

1. The number of tweets y occuring in an interval $[s, e]$ is Poisson distributed with rate $\int_s^e \lambda(t)\, dt$.

$$\Pr\left(y|\lambda(t), [s, e]\right) = \frac{\left(\int_s^e \lambda(t)\, dt\right)^y \exp\left(-\int_s^e \lambda(t)\, dt\right)}{y!} \tag{4}$$

2. Let us assume that the n^{th} tweet occured at time $E_n = s$. We are interested in the inter-arrival time T_n of the next tweet.
3. The arrival time of next tweet E_{n+1} can be obtained as $E_{n+1} = E_n + T_n$.
4. $\Pr$(a tweet occurs by time $s + u$)

$$= F_{T_n}(u) = \Pr(T_n \leq u) \tag{5}$$
$$= 1 - \Pr(T_n > u|\lambda(t), E_n = s) \tag{6}$$
$$= 1 - \Pr(0 \text{ events in } [s, s+u]\,|\lambda(t)) \tag{7}$$
$$= 1 - \exp\left(-\int_s^{s+u} \lambda(t)\, dt\right) = 1 - \exp\left(-\int_0^u \lambda(s+t)\, dt\right) \tag{8}$$

Modeling Inter-arrival Time (Contd..)

5. The probability density function can be obtained as:

$$\Pr\left(T_n = u\right) = \lambda\left(s+u\right)\exp\left(-\int_0^u \lambda\left(s+t\right)dt\right) \tag{9}$$

(By taking derivative of (8) w.r.t u)

6.

$$\Pr\left(T_n = u\right) \approx \lambda\left(s+u\right)\exp\left(-u\lambda\left(s+\frac{u}{2}\right)\right) \tag{10}$$

7. Each rumour E_i have varying temporal profiles. So, distinct intensity function is taken for each:

$$\lambda_i\left(t\right) = \exp\left(f_i\left(t\right)\right) \tag{11}$$

Modeling Inter-arrival Time (Contd..)

8. $$\mathbf{f_i} \sim GP\left(\mathbf{0}, \mathbf{k_{time}}\left(\mathbf{t}, \mathbf{t}'\right)\right) \tag{12}$$

$$\text{where, } \mathbf{k_{time}}\left(\mathbf{t}, \mathbf{t}'\right) = a \exp\left(-\frac{(t-t')^2}{l}\right) \tag{13}$$

(Squared exponential(SE) kernel)

9. Likelihood of posts ${E_i}^O$ over the entire training period = product of Poisson distribution in (4) over equal length sub-intervals $[s, e]$.
10. The rate is approximated as:

$$\int_s^e \lambda_i(t)\, dt = \int_s^e \exp\left(f_i(t)\right) \tag{14}$$

$$\approx (e-s) \exp\left(f_i\left(\frac{s+e}{2}\right)\right) \tag{15}$$

11. The likelihood of posts in the rumour data is obtained by taking the product of the likelihoods over individual rumours.

Importance Sampling

To predict the next arrival time of a tweet given the time at which the previous tweet was posted.

Algorithm 1 Importance sampling for predicting the next arrival time

1: **Input:** Intensity function $\lambda(t)$, previous arrival time s, proposal distribution $q(t) = \exp(t; 2)$, number of samples N
2: **for** $i = 1$ **to** N **do**
3: Sample $u_i \sim q(t)$.
4: Obtain weights $w_i = \frac{p(u_i)}{q(u_i)}$, where $p(t)$ is given by (4).
5: **end for**
6: Predict expected inter-arrival time as $\bar{u} = \sum_{i=1}^{N} u_i \frac{w_i}{\sum_{j=1}^{N} w_j}$
7: Predict the next arrival time as $\bar{t} = s + \bar{u}$.
8: **Return:** $\bar{t}$

1. Sampling the inter-arrival time of occurrence of the next tweet using (10). $p(t)$ is given by (10)
2. Proposal distribution is $Exp(rate = 2)$ $\therefore Mean = 0.5$
3. Assuming the previous tweet occurred at time s, the arrival time of next tweet is obtained using Algorithm 1.
4. This algorithm is run sequentially until the end of the interval of interest is arrived.

Incorporating Text

1. Kernel for text is a linear kernel: $\mathbf{k_{text}}\left(\mathbf{x}, \mathbf{x}'\right) = b + c\mathbf{x}^{\mathbf{T}}\mathbf{x}'$.
2. The full kernel is:

$$\mathbf{k_{TXT}}\left(\left(\mathbf{t}, \mathbf{i}\right), \left(\mathbf{t}', \mathbf{i}'\right)\right) = \mathbf{k_{time}}\left(\mathbf{t}, \mathbf{t}'\right) + \mathbf{k_{text}}\left(\mathbf{x}, \mathbf{x}'\right) \tag{16}$$

Optimization: All model parameters (a, l, b, c) are obtained using Maximum Likelihood Estimate(MLE) of the marginal likelihood over all rumour datasets.

Experiments

Data pre-processing

1. First 2 hours of each rumour lifespan is considered.
2. Posts from first hour of target rumour is used for training.
3. The arrival times of tweets in the second hour are predicted.
4. They consider observations over equal sized time intervals of length 6 minutes in the rumour lifespan for learning the intensity function.

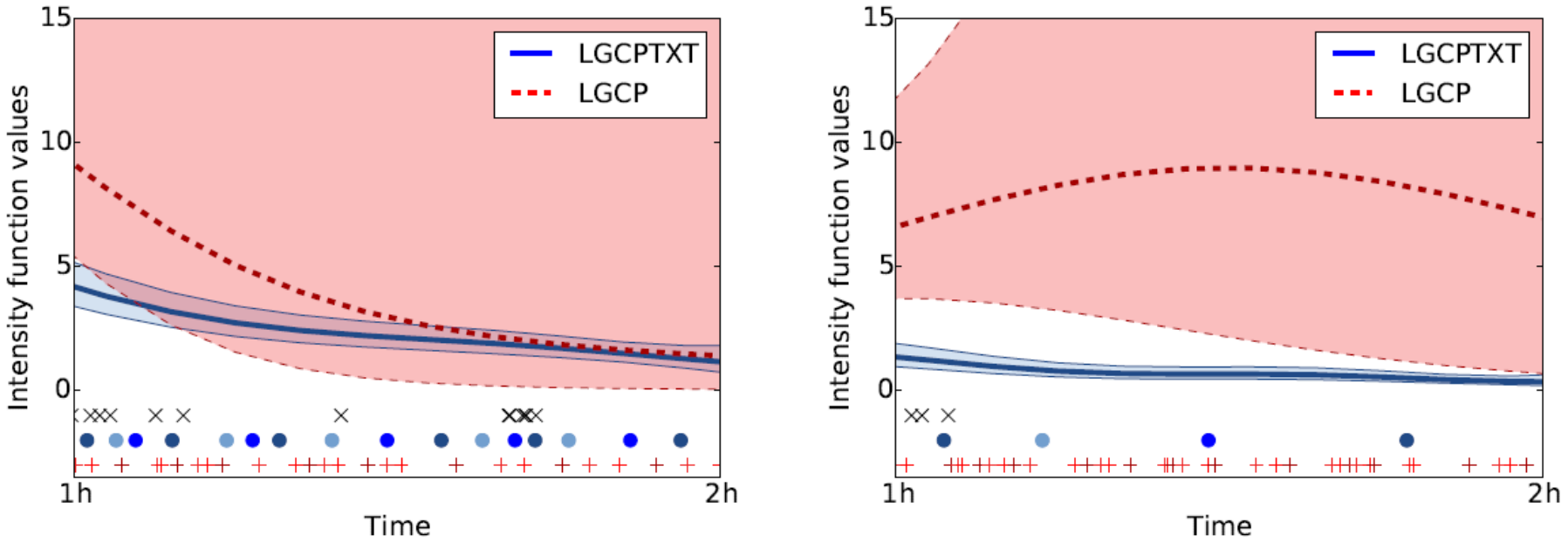
Evaluation metrics

Two metrics based on root mean squared error (RMSE): aligned root mean squared error (ARMSE), penalized root mean squared error (PRMSE).

Baselines

Homogeneous Poisson Process(HPP), Gaussian Process(GP) with linear kernel(GPLIN), Hawkes Process(HP).

Results



(a) rumour #39

(b) rumour #60

Figure 1: Intensity functions and corresponding predicted arrival times for different methods across example Ferguson rumours. Arrival times predicted by LGCP are denoted by red pluses, LGCPTXT by blue dots, and ground truth by black crosses. Light regions denote uncertainty of predictions.

Contributions and Conclusion

This paper makes the following contributions:

1. Introduces **log-Gaussian Cox process** to predict tweet arrival times.
2. Demonstrates how **incorporating text** improves results of inter-arrival time prediction.
3. Evaluation on a set of rumours from Ferguson riots showed **efficacy of the proposed methods** comparing to baselines.
4. Even though the central application is rumours, one could apply the proposed approaches to model the arrival times of tweets corresponding **to problems other than rumours**, e.g. disaster management, advertisement campaigns, discussions about politics