Statistical Language Models

Week 11.5

Stochastic Processes & Markov Chains

- Stochastic process $\{X(t), t \in T\}$, consider discrete case.
- Events could be {product works, product fails}
- Events for us could be Tweets or sentiments

- library(rtweet)
- library(tidyverse)
- library(tidytext)

- datascience = search_tweets('datascience', n = 10000,include_rts = FALSE)
- rstats = search_tweets('rstats', n = 10000,include_rts = FALSE)
- iamdavecampbellTweetw = get_timeline('iamdavecampbell',n=1500)

Exponential distribution model

- Time difference between events i and i-1, t_i
- Common (simple) model:
- $T_i \sim \text{expo}(\lambda)$
- $f(T \mid \lambda) = \lambda e^{-\lambda T}$
- $E(T) = 1/\lambda$, $\text{var}(T) = 1/\lambda^2$

Exponential Model

- $P(T > t) = \int_t^\infty \lambda e^{-\lambda T} dT = e^{-\lambda t}$
- $P(T > t) = \int_t^\infty \lambda e^{-\lambda T} dT = e^{-\lambda t}$
- $P(T > t + s \,|\, T > s) = \dfrac{P(T > t + s)}{P(T > s)}$, why?

Exponential Model

- $P(T > t) = \int_t^\infty \lambda e^{-\lambda T} dT = e^{-\lambda t}$
- $P(T > t) = \int_t^\infty \lambda e^{-\lambda T} dT = e^{-\lambda t}$
- $P(T > t + s \,|\, T > s) = \dfrac{P(T > t + s)}{P(T > s)} = \dfrac{\int_{t+s}^\infty \lambda e^{-\lambda T} dT}{\int_s^\infty \lambda e^{-\lambda T} dT}$

Exponential Model

- $P(T > t) = \int_t^\infty \lambda e^{-\lambda T} dT = e^{-\lambda t}$
- $P(T > t) = \int_t^\infty \lambda e^{-\lambda T} dT = e^{-\lambda t}$
- $P(T > t + s \,|\, T > s) = \dfrac{P(T > t + s)}{P(T > s)} = \dfrac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t - \lambda s + \lambda s}$

Exponential Model

- $P(T > t) = \int_t^{\infty} \lambda e^{-\lambda T} dT = e^{-\lambda t}$
- $P(T > t) = \int_t^{\infty} \lambda e^{-\lambda T} dT = e^{-\lambda t}$
- $P(T > t + s \,|\, T > s) = \dfrac{P(T > t + s)}{P(T > s)} = \dfrac{e^{-\lambda(t+s)}}{e^{-\lambda s}} == e^{-\lambda t} = P(T > t)$
- Memoryless: conditioning on the past does not matter

Poisson Process

- Model for event counts: Number of events occurring up to time t is N(t)
- {N(t),t=0,1…} is a Poisson process with rate β, it has these properties:
- 1) N(0) = 0
- 2) Process has independent increments
- 3) Number of events in an interval of length t is Poisson distributed with mean βt
- $P(N(t) = n) = \dfrac{e^{-\beta t}(\beta t)^n}{n!}$

Poisson Process

- Model for event counts: Number of events occurring up to time t is N(t)
- {N(t),t=0,1…} is a Poisson process with rate β, it has these properties:
- 1) N(0) = 0
- 2) Process has independent increments
- 3) Number of events in an interval of length t is Poisson distributed with mean βt
- $P(N(t+s) - N(s) = n) = P(N(t) = n)$ = βt
- small intervals have ~ no chance of containing any event counts

Poisson process with parameter β where event n occurs at time $\sum_{i=1}^{n} T_i$.

- Event 1 occurs at time T_1
- Find $P(T_1 > t)$
- $P\left[N(t) = 0\right] = \dfrac{e^{-\beta t}(\beta t)^0}{0!}$
-

Poisson process with parameter β where event n occurs at time $\sum_{i=1}^{n} T_i$.

- Event 1 occurs at time T_1
- Find $P(T_1 > t)$
- $P\left[N(t) = 0\right] = \dfrac{e^{-\beta t}(\beta t)^0}{0!} = e^{-\beta t} = P(T_1 > t)$
- Equivalence between the Poisson model for the number of events up to time t and the time between events.

Exponential distribution model

- Time difference between events i and i-1, t_i
- Common (simple) model:
- $T_i \sim \text{expo}(\lambda)$
- $f(T \mid \lambda) = \lambda e^{-\lambda T}$
- $E(T) = 1/\lambda$, $\text{var}(T) = 1/\lambda^2$

Time between events (units = seconds)

- Tweettimes = datascience$created_at
- #Tweettimes = rstats$created_at
- sort(Tweettimes)
- diff(sort(Tweettimes))
- plot(diff(sort(Tweettimes)))
- hist(diff(sort(Tweettimes)))
- hist(as.numeric(diff(sort(Tweettimes))),100)

Time between events

- T_i ~ expo(λ)
- $f(T \mid \lambda) = \lambda e^{-\lambda T}$
- E(T) = 1/λ,
- var(T) = $1/\lambda^2$

- #mean:
- lambda = 1/mean(as.numeric(diff(sort(Tweettimes))))
- x = seq(0,max(as.numeric(diff(sort(Tweettimes)))))
- hist(as.numeric(diff(sort(Tweettimes))),100,probability = TRUE)
- lines(x, lambda * exp(-lambda * x),col="#8A7AF9",lwd=3)

Poisson Model

- N(t) ~ Poisson(λt)
- $P(N(t) = n) = \frac{(\lambda t)^n}{(\lambda n)!} e^{-\lambda t}$
- E(T) = λt,
- var(T) = λt

Number of tweets per time interval

- N(t) ~ Poisson(λt)
- $$P(N(t) = n) = \frac{(\lambda t)^n}{(\lambda n)!} e^{-\lambda t}$$
- E(T) = λt,
- var(T) = λt

- #mean:
- lambda = 1/mean(as.numeric(diff(sort(Tweettimes))))
- #Timeinterval
- t=60*60* 6
- t*lambda
- plot(sort(Tweettimes))
- #time increments:
- abline(h=min(sort(Tweettimes))+ t*Z)
- abline(v=which(sort(Tweettimes)> min(sort(Tweettimes))+t)[1])