Flight Delay Prediction

Sirjan Kafle
sxkafle12@gmail.com
https://flightdelay.us

1 Introduction

We build an app that models the distribution of flight delays conditioned on variables like:

- Origin Airport
- Destination Airport
- Airline
- Time of Day

More specifically, we want users to provide a subset of those "conditional variables" and we'll output information on the probability that the delay time will exceed certain thresholds, e.g. 30 minutes, 1 hour, etc along with plotting those probabilities vs delay time.

We are resource constrained based on the container size of the web app we're deploying to, so we describe compression methods for these distributions by modeling them as exponentials - we store and load the parameters of that distribution for the app rather than full flight data. There are nearly 46M flights in the dataset from 2018 to 2024, so this is a significant reduction in size. In this doc, we explain a general framework for building algorithms to efficiently model these probabilities at scale.

2 Modeling Conditional Distributions as an Exponential

2.1 Definitions

Let D be the continuous random variable representing the delay time of a flight. Let C_i be the ith conditional variable, e.g. origin airport. In our cases C_i are categorical, although they don't have to be. Even for time of day, we've divided it into 4 categories with six hour intervals: morning, afternoon, evening, night. For the purposes of the theory, we'll derive all our results for generic C_i, which means these techniques are extendable to other conditioning variables in the future.

We denote the PDF of the conditional distribution as $f_D(\delta | \{C_i\})$ where δ is the delay time. What we actually care about, however, are the cases where $\delta > 0$, so we actually want to model $D' = \max\{D, 0\}$ well.

2.2 Proposed Model

Recall $D' = \max\{D, 0\}$. Observe that for any $\delta > 0$,

$$f_{D'}(\delta | \{C_i\}) = f_{D'}(\delta, D > 0 | \{C_i\}) = f_{D'}(\delta | \{C_i\}, D > 0) \Pr(\delta > 0)$$

As motivated by above, we model conditional distribution on $D > 0$ as an exponential distribution with parameter λ.

We model $\Pr(\delta > 0)$ as a Bernoulli distribution with parameter p.

Therefore, our final model for $f_{D'}(\delta|\{C_i\})$ is:

$$f_{p,\lambda}(\delta) = \begin{cases} 0 & \delta < 0 \\ 1-p & \delta = 0 \\ p\lambda e^{-\lambda\delta} & \delta > 0 \end{cases}$$

It's easy to see that this is a valid PDF since $\int_0^\infty p\lambda e^{-\lambda\delta} = p$.

2.3 Maximum Likelihood Estimate of Parameters

2.3.1 Derivation

We want to fit p, λ to best fit our data.

Take a set of conditional variables $\{C_i = c_i\}$. From our data, assume we have n samples such that the conditionals $\{C_i = c_i\}$ are satisfied. Denote these delays as $\delta_1, \delta_2, \ldots, \delta_n$.

Let $m = |\{i : \delta_i > 0\}|$ be the number of positive delays.

We make the assumption that the delays are iid. We first derive the log likelihood function of the data after transforming $\delta'_i = \max\{\delta_i, 0\}$.:

$$\begin{aligned} h(p,\lambda) = \log \mathcal{L}(p,\lambda;\{\delta'_1,\delta'_2,\ldots,\delta'_n\}) &= \log \prod_{i=1}^{n} f_{p,\lambda}(\delta'_i) \\ &= \sum_{\{i:\delta_i \le 0\}} \log(1-p) + \sum_{\{i:\delta_i > 0\}} \log p + \log\lambda - \lambda\delta_i \\ &= (n-m)\log(1-p) + m\log p + m\log\lambda - \lambda \sum_{\{i:\delta_i>0\}} \delta_i \end{aligned}$$

We want to find:

$$\arg\max_{p,\lambda} \mathcal{L}(p,\lambda;\{\delta'_1,\delta'_2,\ldots,\delta'_n\})$$

We compute partial derivatives:

$$\begin{aligned} \frac{\partial h}{\partial p} &= -\frac{n-m}{1-p} + \frac{m}{p} = 0 \\ \frac{\partial h}{\partial \lambda} &= \frac{m}{\lambda} - \sum_{\{i:\delta_i>0\}} \delta_i = 0 \end{aligned}$$

Yielding us our estimates:

$$\begin{aligned} \hat{p} &= \frac{|\{i:\delta_i > 0\}|}{n} \\ \hat{\lambda} &= \frac{|\{i:\delta_i > 0\}|}{\sum_{\{i:\delta_i>0\}} \delta_i} \end{aligned}$$

2.4 Delay Survival Function

Our ultimate goal is to display the survival function or $P(D > \delta|\{C_i\})$ for $\delta > 0$ to user queries. Based on our model we can derive this (for $\delta > 0$):

$$\begin{aligned} \Pr(D > \delta) &= \int_\delta^\infty p\lambda e^{-\lambda x} dx \\ &= -pe^{-\lambda x}\Big|_\delta^\infty \\ &= pe^{-\lambda\delta} \end{aligned}$$