EM Algorithm

Noah Gblonyah, Seth Okyere, Michael Throolin

4/14/2022

The EM Algorithm

The EM Algorithm has its roots in work done in the 1950s but really came into statistical prominence after the seminal work of Dempster, Laird, and Rubin, which detailed the underlying structure of the algorithm and illustrated its use in a wide variety of applications. (Casella and Berger 2002).

Another common tool used for getting a maximum-likelihood estimation of censored data is called the EM Algorithm. Here, the 'E' canonically stands for the 'Expectation' step and 'M' represents the 'Maximization' step. Hence the EM Algorithm takes the expectation of the log-likelihood function, then maximizes that quantity. It repeats that process until the parameter converges to a specified value.

Formally, if we let $\theta^{(p)}$ represent the p^{th} iteration of the algorithm to estimate the parameter θ. these two steps can be written out as follows:

Expectation (E-Step): Compute $Q\left(\theta^{(p)}|\theta^{(p-1)}\right) = E\left[\log\left(f(\mathbf{x}|\theta^{(p)}\right)|\mathbf{y}, \theta^{(p-1)}\right]$

Maximization (M-Step): Maximize $Q\left(\theta^{(p)}|\theta^{(p-1)}\right)$

The EM algorithm can often lead to functions that are tricky to evaluate. However in special cases, such as the exponential family family case, the algorithm becomes much easier to evaluate. Specifically a function $f(\mathbf{x}|\theta)$ is an exponential family if it can be written as $f(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(\mathbf{x})\right)$. It has been shown that we can use the complete sufficient statistic $T(\mathbf{X}) = \sum_{i=1}^{k} t_i(\mathbf{x})$ to estimate the parameter θ. This is done as follows:

Expectation (E-Step): Estimate $\mathbf{t}(\mathbf{x})$ by finding $\mathbf{t}^{(p-1)} = E\left(\mathbf{t}(\mathbf{x})|\mathbf{y}, \theta^{(p-1)}\right)$

Maximization (M-Step): Determine $\theta^{(p)}$ as the solution to $E\left(\mathbf{t}(\mathbf{x})|\theta\right) = \mathbf{t}^{(p-1)}$

Example

For demonstration purposes, we will simulate data from an exponential distribution, censor the data, and use the EM algorithm to estimate θ.

Now, noting that a random sample for the exponential family has a complete sufficient statistic of $\sum_{i=1}^{n} x_i$, we must find the expectation of $E(\sum_{i=1}^{n} x_i|\mathbf{y}, \theta^{(p-1)})$.

We can begin by separating the known data from the unknown data. Will call use an indicator, d_i to indicate presence of data at i, and if a variable is missing we will call it z_i.

Supposing for each insured person we get $(c_{(1,i)}, c_{(2,i)}, x_i)$, where $c_{(1,i)}$ represents the deductible and $c_{(2,i)}$ represents the policy limit.

Hence,

$$E\left(\sum_{i=1}^{n} x_i|\mathbf{y}, \theta^{(p-1)}\right) = E\left(\sum_{i=1}^{n} y_i d_i + z_i(1-d_i)\right)$$

To estimate z_i, we will use the memoryless property of the exponential distribution. Hence,

$$z_i = \begin{cases} \min\{\theta^{(p-1)}, c_{(1,i)}\} & 0 < z_i \leq c_{(1,i)} \\ c_{(2,i)} + \theta^{(p-1)} & z_i > c_{(2,i)} \end{cases}$$

This means

$$\begin{aligned} E(z_i) &= \frac{\int_0^{c_{(1,i)}} \min\{\theta^{(p-1)}, c_{(1,i)}\} \frac{1}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}\, dx_i + \int_{c_{(2,i)}}^{\infty} (c_{(2,i)} + \theta^{(p-1)}) \frac{1}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}\, dx_i}{\int_0^{c_{(1,i)}} \frac{x_i}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}\, dx_i + \int_{c_{(2,i)}}^{\infty} \frac{x_i}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}} \\ &= \frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}} \end{aligned}$$

Therefore, we can simplify our expectation step down to

$$E\left(\sum_{i=1}^{n} x_i | \mathbf{y}, \theta^{(p-1)}\right) = E\left(\sum_{i=1}^{n} y_i d_i + z_i(1-d_i)\right)$$

$$= \sum_{i=1}^{n} y_i d_i + (1-d_i)\left(\frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}\right)$$

And our maximization step is the solution to

$$\begin{aligned} E(\mathbf{t}(\mathbf{x})|\theta^{(p)}) &= E(\sum_{i=1}^{n} x_i | \theta^{(p)}) = n\,\theta^{(p)} \\ &= \sum_{i=1}^{n} y_i d_i + (1-d_i)\left(\frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}\right) \end{aligned}$$

$$\implies \theta^{(p)} = \frac{1}{n}\sum_{i=1}^{n} y_i d_i + (1-d_i)\left(\frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}\right)$$

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

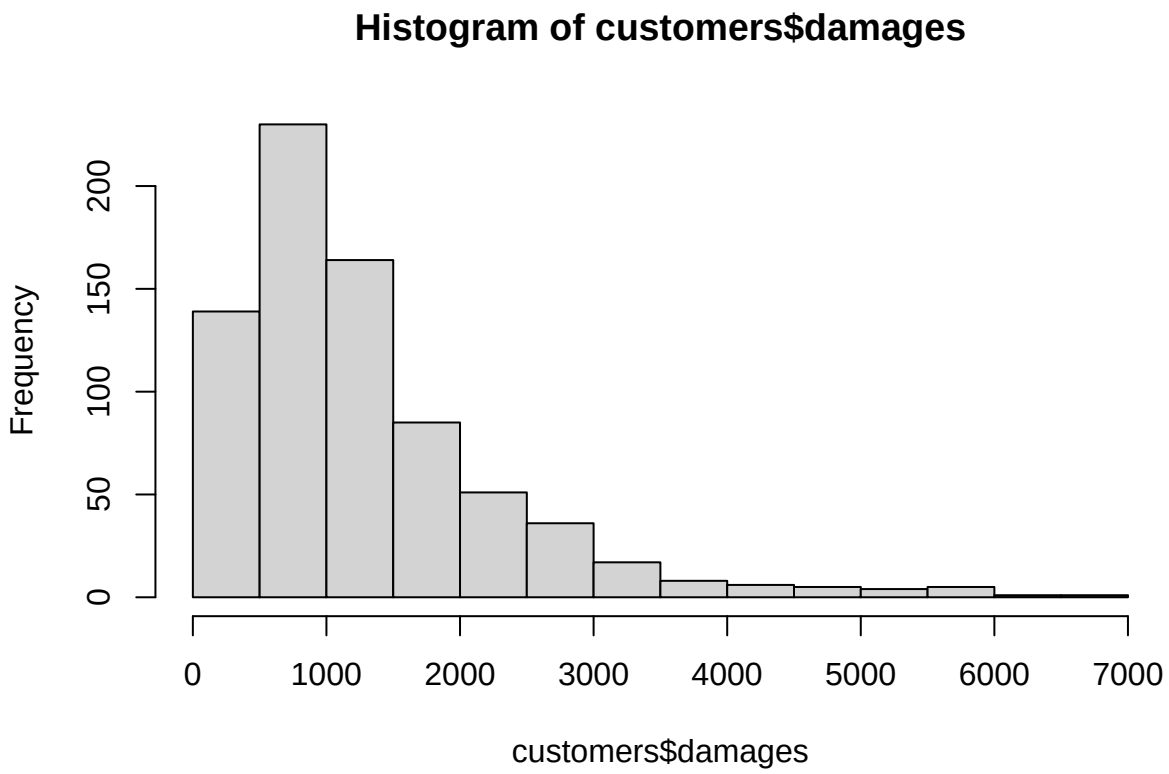
```
library(naniar)
```

```
## Warning: package 'naniar' was built under R version 4.1.3
```

```r
set.seed(53523)
c <- c(500,6000)

#Simulate Data
n_customers <- 1000
damages <- round(rexp(n_customers, 1/1000),2)
deductible <- round(runif(n_customers, min = 0, max = 600))
limit <- round(runif(n_customers, min = 5000, max = 20000))

customers <- as.data.frame(cbind(deductible, limit, damages))
customers[(damages < deductible),]$damages = NA
customers[(limit < damages),]$damages = NA
hist(customers$damages)
```




```r
observed_data <- customers[!is.na(customers$damages),]
unobserved_data <- customers[is.na(customers$damages),]

sum_observed <- sum(observed_data$damages)

theta_new <- 500
theta <- 0

i = 0
#Expectation step
while((theta - theta_new)^2 >0){
```

```
  theta <- theta_new
m <- min(theta, unobserved_data$deductible)
  numerator <- m-m*exp(-unobserved_data$deductible/theta) +
    (unobserved_data$limit +theta)*exp(-unobserved_data$limit/theta)

  denominator <- theta -(unobserved_data$deductible + theta)*exp(-unobserved_data$deductible/theta)
  + (unobserved_data$limit + theta)*exp(-unobserved_data$limit/theta)

expectation <- sum_observed + sum(numerator/denominator)

theta_new <- (expectation)/n_customers
 i<- i +1
}
```

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. Duxbury.