Project 3: Analyses of daily COVID-19 cases across nations

The pandemic of COVID-19 is the biggest challenge that the world is facing right now. Our lifes are all deeply affected by this public health crisis. The White House has pulled together multiple open research data sets, and called for data scientist to assist the modeling of the growth trajecties of confirmed cases across the worlds, and to help prediction future cases and to identify risk factors. I would like to encourage you to be part of the force;

The attached covid19-1.csv, is a subset of the open data, which recorded the following variables:

Id: Record ID

Province/State: The lcoal state/province of the record; 54% records do not have this info;

Country/Region: The country/regionoof the record;

Lat: Lattudiute of the record;

Long: Longitude of the record;

Date: Date of the record; from Jan 21 to March 23;

ConfirmedCases: The number of confirme case on that day;

Fatalities: The number of death on that day;

Task 1: Fit a logistic curve to the cumulative confirmed cases in each region;

Logisitc curves could be one way to model the trajectory of cumulative cases; It is a parametric function with the form

$$f(t) = \frac{a}{1 + exp\{-b(t-c)\}},$$

where t is the days since the first infection; a is the upper bound, i.e. the maxium number of cases a region can reach, b is growth rate, and c is the mid-point, where the curve changes from convex to concave; Each curve is uniquely defined by (a, b, c). By design a logistic curve increases exponentially at begining and slows down at the end;

Task 1.1 Develop an optmization algorithm to fit a logisitc curve to each region, and find a way to visualize your fitted curves effectively; What you learn from your fitted models? e.g. how many regions have passed the midpoint? how many regions are approaching to the end of the virus speading; Which regions have faster growth rates and which regions have more "flat" growth?

Task 1.2 You can find daily reports after March 23 from the following github site

https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports

From those data, do you think if the logistic curve is a reasonable model for fitting the curmulative casesa and predicting future new cases?

Task 2: Clustering your fitted curves

clustering is an effective data exploring tools; It helps develop hypothesis and identify potential risk factors; Apply K-mean and Guassian mixture model (with EM algorithm) to cluster the fitted parameters $(\hat{a}, \hat{b}, \hat{c})$; Which algorithm does a better job in clustering those curves? Are the resulting clusters related to geogrpahic regions, or the starting timeing of the local virus speading, or the resources of the regions? You may use external informations to help understand the clusters, i.e. find plausible explanations why some regions have similar (a, b, c)?

Task 3: Write a summary report to share your findings;

```r
#Includes uour R codes
```