電機13 崔浩堂109511068

September 26, 2022

Introduction to Data Science HW1

Introduction

The first homework is let us choice some dataset to study for future homework or project, and these datasets should have bright feature so that I can discuss in the future.

Datasets

- Stock – FAANG
- Life – Human Typing
- Computer Vision – MOT Dataset
- Medical Science – MIMIC III

FAANG

WHY FAANG?

The first dataset that comes to my mind is stock. But the stock market is so big, and it's hard to analysis without more information. This try to focus on part of them, and let stock dataset won't be indiscriminate use.

After survey, I think S&P 500 is a good stock to analysis. But the trouble is S&P500 update the company every season. However, only analysis the S&P500 stock is not enough to predict, it will let us lost a lot of information. Thus, the next idea is FAANG, which is the abbreviation of Facebook, Apple, Amazon, Netflix and Google. The common point of these companies is all of them is about technology company.

FEATURE

The feature of stock is it will update everyday and the price is cause by news of the day. There are pros and cons to this, because use old data and predict past future is not valuable, and predict future price can create a lot of value! However, update all data set is time consuming. So I write a script in python to update the stock data everyday, but append to old data(that is, only get the day's stock price). The code is available at here: (https://github.com/henrytsui000/DataScienceProject/blob/main/

dataset/update_data.py). The code avoid computer consume time to get entire data. After compare different method to get stock price, I think FFN package could through yahoo's API to get the price easily.

The second issue is stock would affect by news. Therefore I think in the next homework, I should use NLP to merge the news and the past stock price. Before looking for more information, I think the transformer (BERT) will be a good model to judge whether the news will affect the stock price!

MIMIC III

WHY MIMIC III?

MIMICIII is a dataset about the patient's physical condition in ICU, a lot of medicine machine learning problems would train on this database. It include almost all patient's feature, this things make it be the best dataset to analysis. And what I want to do is predict whether the patient get severe symptoms like sepsis in the next 12 hours.

FUTURE

Although MIMIC III is a well-known database, it still has some problems. The first problem is access rights. Everyone who wants to access the MIMIC database should pass the test and promise not to spread their information. So I need to apply for access again. Another problem is the large number of gaps in the data, which can lead to inaccurate predictions.

Apart from the above issues, I think this would be a good dataset to predict patient health over the last few hours!

MOT dataset

WHY MOT?

MOT is Multiple-Object-Tracking, which is a computer vision problem. Participate in Scattering Let the model keep track of everyone in the scene. An interesting fact is that next semester, I will go to Germany for an exchange. The tutor is MOT16, the founder of MOT17dataset. I see this as an opportunity to do more research on this dataset.

FEATURE

Unlike other datasets, the MOT dataset is a 3D dataset (2D pictures and temporal information). Which clue makes it hard to get the bounding box of the person. For example, a 60FPS, 10-second, 20-person video requires tens of thousands of bounding boxes! So how to make the dataset bigger or more useful is the topic I want to discuss.