Introduction to practical data science course

Haochen

7th January 2019

1 What is data science

New terms have always been created to motivates new direction of society advancement or to reflects new fashion of academia, So is " data science ". It is almost emerge in 2010 and hit the

People from different field with different background usually have different answers to the definition of data science. So to avoid the further inconvience, lets try to define it from " what is not data science ", and then go back to tankle this issue.

First, data science is not just machine learning. In machine learning realm, people focus more on design or analysis algorithms to deal with challenges of pattern recognition. What is worth mentioning is that almost 50% problems do not imply any patterns to predict, such as using twitter texts to predict stock market.

Second, data science is not just statistics. Traditional statistics emphasis more of theoretic work, like asymptotic behaviour and efficient estimator.

Third, data science is not big data. Admittedly, big data is needed for better solve the problem. But not always. So if it can be done in a single computer memory, we shouldn't create extra work, say, using spark or haloop.

A possible definition combine multiple aspect of above discussion comes as following: data science is the application of **computational** and **statistical** techniques to address or gain insight of **real-world** problems.

2 Learning objectives

So in this course, we pay less attention on theory such as linear algebra, they may be helpful though. The emphasis is on implemting shelf-off algorithms in python related packages. However, a good impletattion must go with a firm grasp and understanding of the algorithms We also need to understand at least some details of algorthm, better understanding help us better master these methods.

Anthoe emphasis is on the entire pipline of data science work.It generally contains five procedure. **Data collecting and management , form a problem(data exploration and visualiation), solve that problem imaginatively(usually using statistical or machine learning model), interpretation the results and presentation.**

Some people as well as Kaggle tends to neglect two sides and put more weight on the machine learning modeling which is somewhat misleading. So in this course we try to balance all the procedures, and may be focus more on data collection and processing and presentation.

3 Jupyter notebook lab

- The data science language is python3, it's in-negotiatable. Since all the package update for python2 is stoped in 2020. And usually, R is only well for statistical tasks. We use anaconda3 as a platform including virtue environment ,package management and Jupyter notebook.
- Typically there are two ways to install extra packages:conda install pkname pip install pkname
- notice serveal things to master Jupyter notebook.
  - **mixing codes and text** to be intrepretative enough.
  - make code cells run sequencially.
  - using **shift+tab** to see help, using tab to complete.

4 python data structures for data scientist

- list
  - **list comprehension**.
  - open close or with open.
  - list impletmentation: **cheaper for appending, costly for insert**.
    - list access $O(1)$
    - append $O(1)$
    - insert $O(n)$
- dictionary (keys $\mapsto$ values)
  - hash function and collision
  - **insert, delete, search** $O(1)$
  - dict comprehension
  - unique words count
- numpy array
- pandas dataframe(maybe)