HUDK 4050 Reflection Essay---Siyuan\_Gu

--Introduction

This essay briefly talks about my learning experience of HUDK 4050: Core Method Educational Data Mining. I will start from my background and expectation before taking this course. And then describe my learning experience of this course at this semester. Finally, ending up with my future goal according to what I have learned through this class.

--Pre-course experience

Since it is not the first time I take coding course from Teachers College, Columbia, I can’t wait to learn some new information about data processing and manipulation through diverse programming terminals. I have had internship experiences about applying for different packages of python about machine learning such as decision tree, neural networks, casual inference, and so on. But my superior just taught me how to use them by simply plug in dataset rather than understand how they worked. In other words, I know literally nothing about the function of each line of the original code from each package.

However, last semester I took some coding course using R studio as a terminal. The professor gave us some tasks about basic applications like plotting and programming structure. I felt like I was getting started on programming but there were more questions need to be studied.

Coming up with many questions about data analysis, I added this course to my schedule. At the beginning I was looking for moving further on deeper comprehension than before. So, I read the syllabus to learn the structure of this course. It was good to know that the class started from very fundamental study on data manipulation as I need to review and improve my basal skills of coding after summer break.

--Course experience

Although I did not build up a clear overall systematic schema, I kind of learned about how to tidy up dataset and choose proper model for specific question. For example, it’s better to use decision tree classification to deal with multiple Yes or No questions. And I can re-evaluate the tree model by Naïve Bayes’ model or even making up a neural system to let the machine to fit models including possible factors. But I have to be careful about the overfit problem as artificial intelligence cannot make judgement of the model based on the real-life issues. I got many tips like this example from accomplishing analysis challenge assignment by myself.

The requirement of the ACA mentioned that we may find other classmates to form a group to finish it, while I used to work individually. But the discussion before doing assignments during class procedure really helps me to construct some ideas about how to launch into the question. Sometimes their solutions were pretty surprising me. I guess that’s because of the different background experiences between us. Hence, I really enjoy to hear diverse sounds from people who were good at different fields.

Looking forward

Because I had taken part in several internships at many investment companies before took this course, I already tried some application of specific models we learned in this class. But now I can do more on investment analysis according to what we’ve learned this semester. Taking diagnostic metrics as an instance, without learning such technic, I could only run test through my neural system model by plugging in different training data. But now I am able to evaluate my prior models based on the original theorem or the fundamental architecture of the model. The Silhouette coefficient and elbow method would be a good choice of evaluation on clustering the neural cells. Or I may apply for prediction of linear regression. To do this, I am going to regard each neural unit as a factor (Dependent variable Xs) and the target value as independent variable Y.

Besides, the differences between R studio and Python decide their scope about data analysis. For my perspective, R has more package covering border aspect of topics from genetics to economics. And it has more function of data visualization as its’ stronger plotting like ggplot2. While Python does better on deep learning but we have to pay attention on the overfit problem.