Data Mining

1. Domain knowledge is very important in any kind of field, not only data science. Domain knowledge in data science, refer to the terms and the knowledge of the programs and systems. Without domain knowledge, it would be very hard to communicate to other data scientists, and very hard to perform tasks. Domain knowledge is essentially the knowledge needed to be familiar with each respective field, the skills and experience required from your job, or the knowledge of that particular domain.
2. Data preprocessing is very important due to it being the foundation of processing the data. Data preprocessing turns the data into a form of data that can be processed. Without the preprocess, data won’t be able to be processed at all.

Data preprocessing use methods like sampling and transformation. Which will either lessen the data needed to be processed or convert it into a tangible concept easily processable. In data science, this is very important, as there are limited resources, and the computer cannot understand abstract thoughts, most of the times, even we can’t understand abstract thoughts, which is why we need to convert the data into which something the computer can process.

1. Classification and Regression, classification is classifying the data into groups, it helps us see which data is similar and different with labels. Regression is prediction from previous data. Tools for Classification are vector machines, discriminant analysis, naïve bayes, nearest neighbor, and Neural Networks. Tools for Regression are Linear Regression, SVR,GPR, Ensemble Methods, Decision Trees, Neural Networks.
2. Parametric vs nonparametric prediction, Parametric methods have a limit of fit quality, while nonparametric methods keep improving as we add more data to fit. Parametric methods are often simpler to interpret. Parametric: you are looking at one part of the data, so you cannot see the whole picture when you predict, however it is very fast. Nonparametric: you look at every part of the data when you predict, which will take a long time.
3. Linear: data where it connects to another data, and only one data can be reached by another data. It can be traversed in a single run.

Nonlinear: can be connected to multiple data, cannot be traversed in a single run.

We cannot preform regression on nonlinear data because there are multiple data points to choose from.

All linear models are not useful because it restricts the data.

If the data isn’t linear, it tells us that the data isn’t simple to look at, it is very complex.

If the data isn’t separable, then we need to rethink our organization and find a new way to visualize the data.