Name: Chun-Chi Huang
Course: Data Modeling
Professor: Michael McKee
Date: 02/09/2018

What is machine learning?

Before answering what machine learning is, I want to mention one project that I had when I studied in National Taiwan University. The project was about non-destructive detection for the sweetness of mangoes. Instead of physical cutting mangoes and extracting their juice or flesh to measure by the sweetness meter, we employed the near infrared ray (NIR) machine to remotely scan the surface of mangoes and then got the sweetness value. Prior to getting the accurate result, we had to establish the detection model. First, we needed to have the data of the NIR spectrum and the corresponding sweetness value, and then regression analysis was introduced to obtain the relationship between the specific wavelength values and sweetness. We also called this "training". After training, we got the equation in which the specific wavelength values were the input data, and the output data was the sweetness of mangoes. In general, I used three specific wavelength values to progress the training model, and then I achieved 90% accuracy in our training data and also had similar accuracy in other data out of the training data. If I chose more than three wavelength values, I could get more precise accuracy than 90% in the training data. However, when I applied this model to predict the new coming samples, the accuracy would decrease. We called it "overtraining". It meant that this detection model excessively confined to the training data not to general samples.

Back to the topic of machine learning, it shows the method that people uses data, including input and output, and selects the learning model to get the predictive model. The predictive model is also the relationship between input and output data. As to the learning model, there are some models mentioned in this book, Predictive Analytics, such as decision tree, artificial neural networks, loglinear regression, support vector machines and TreeNet. I also used fuzzy logic in my undergraduate project to predict the behavior of purchasing new cars. I want to use an example in this book to explain the idea of machine learning. The example is about Dan and Chase Bank. Chase wanted to mitigate the risk of loan so they invited Dan to create a learning model to predict the possibility of the risk of prepayment. Dan used decision tree as the training model feeding 21816 samples as training data. Every sample included several characteristics. Then he used the learning model, 4 segments, 10 segments and 39 segments, to predict another 5486 samples, the result showed the same score in both training and test data. When he used 638 segments to proceed the training, although he got the higher score on training samples, the score on test samples was terrible. They called it "overlearning".

Another thing I want to mention is that people should not put two irrelevant things into the learning model. Machine learning is using inductive approach to obtain the principles between two events. For the example from Predictive Analytics, don't use Bangladesh's butter production to predict the U.S. stock market because these two events are irrelevant.