Calvary Gollwitzer, Tevin Julien, Kevin Walters

Amulya Yadav

IST 402

9/10/2019

Week 2 – Course Notes

This week's lectures consisted of a general introduction to machine learning. Machine learning was birthed out of the realization that humans could use computers to do pattern recognition for them. In fact, machines could handle pattern recognition much better than humans, and could analyze more features of a given data set and subsequently discover patterns that humans couldn't even recognize. Simply put, this realization is what helped start the creation of machine learning as we know it today.

In machine learning there lies four main categories. These categories are: Classical Learning, Neural Nets and Deep Learning, Reinforcement Learning, and Ensemble Methods. This week in class we focused primarily on Classical Learning, with some discussion about Reinforcement Learning as well. In classical machine learning there are two broad categories: Supervised and Unsupervised learning. Supervised learning consists of Classification Problems and Regression problems, which we discussed heavily in this week's class. Classification and Regression models make up the bulk of all the machine learning that occurs today.

Classification problems can be used to predict the outcome of something when the variable you want to predict is discrete. Generally, a classification problem is either a binary classification problem or a multi-class classification problem. A binary classification problem means that the variable you are trying to predict is either a yes or no, a 1 or a 0. An example of this could be a model that predicts whether a piece of email is spam or not. In this case, the machine would simply predict that yes, the email is spam, or no, it is not, making it a binary classification problem. Multi-class classification problems are more complicated because they must predict a variable that has more than two possibilities. An example of this would be a model that predicts the outcome of an election. One way that you can construct a multi-class classification model is by breaking it up into several binary classification models, so we spent a lot more time in class defining binary classification models and how they operate.

Binary classification models were some of the first forms of machine learning models. They are great because they can work with a lot of data, and they are generally easy to understand, as compared to a neural network. Binary classification models work off of data that has been labeled and given to them in order to find patterns within the features of the datapoints.

However, they do have a limit to their performance, and cannot learn arbitrarily well. Another problem that can exist in classification models and supervised learning in general is that these models are dependent on the data put into them. If there is a data imbalance, or the data suffers from skew, a poorly constructed classification model could still be accurate. For example, a poor model could simply predict that every person run through its algorithm has cancer. However, if all the data that it received was only from people who did have cancer, the model would be 100% correct. This is where concerns over the accuracy of a binary classifier come in. One way we can solve this is through assessing a classifiers precision and recall.

Precision and recall can be assessed for both the positive and negative class, or the 1 and the 0, of the classifier. Precision, in the positive class, asks "for the items that were returned by the classifier as being positive, how many were actually positive". In the case of our cancer example, this is asking how many people did our classifier predict to have cancer that actually had cancer. Recall, in the positive class, asks "out of the items which were actually positive, how many were predicted to be positive". In our example, this is saying that for the people that had cancer, how many were predicted to have it. If a classifier has been constructed poorly, it will show as having a poor factor in precision or recall for the positive or negative class. Properly constructed classifiers will demonstrate suitable outcomes when evaluating all four of these factors.

The other form of supervised learning within classical machine learning is regression models. Regression models are used when the variable you are trying to predict is continuous, like the price of a stock, or crop yields. Continuous data means that the prediction can take a value on a number line. Within regression models there lies linear regression and polynomial regression. Linear regression assumes that your data has a linear property and that you can fit a line to that data. Polynomial regression assumes you can fit a polynomial line to the data, or you can fit a line to the data that is curved. The assigned reading in the vas3k blog described regression models as being "classification where we forecast a number instead of category. Examples are car price by its mileage, traffic by the time of the day, demand volume by growth of the company etc,.".

We wrapped up this week's lectures with a brief discussion into unsupervised learning and reinforcement learning. Unsupervised learning means that none of the data that the model is using has labels. Labels equates to supervision. What unsupervised learning models try to do is find patterns between the data. An example of this is the facial recognition an iPhone or Google Photos can do. These programs are able to group together and find images of specific people by looking for patterns in facial features, and therefore clustering together images of a certain person for easy access.

Finally, we concluded with an introduction to reinforcement learning. In reinforcement learning you don't have data to put into the model, but rather an environment for the model to exist and learn in. This environment interacts with the model through a rewards system, so that the model can learn over time what actions prove more fruitful for it to do. An example of this would be a reinforcement model that is tasked with balancing a baseball bat on a point. The environment it exists in adheres to all the general rules of physics and gravity. Over time, the model will learn how its different reactions to how the bat falls plays out in the long run. Eventually, after enough simulations, the model would be proficient at balancing a baseball bat because it has learned what works best in each scenario.

This concludes the topics we discussed in class during the week of September 2nd-6th. Overall it encompassed a general overview of the history of machine learning, as well as a more in-depth look at the forms of classical machine learning, specifically supervised learning methods. The assigned readings were helpful in providing more context to the lessons and in creating these lecture notes. We wrapped up the week with a transition into unsupervised and reinforcement learning that transitioned into the next week's learning.