D.J. Anderson (dra2zp) 11/20/2017

Special Lecture Report

Introduction

SWIFT was created in 1973 when 239 banks from 15 countries met to solve the problem of how to communicate about cross-border payments. They completed the Global Payments Initiative (GPI) in less than a year that allows customers to track their payments and provides visibility about payments for instances where payments are not instantaneous.

Data Science

Data science is an interdisciplinary field, involving mathematics, statistics, computer science/engineering, business analytics, and information science. It is used to extract knowledge and insights from data like structure, volume, velocity, and type. More people are getting involved in data science now because "big data," which is the explosion and rapid growth in computing power, and explicit programmatic solutions are no longer feasible in all cases. According to IBM, ninety percent of current global data were created in the past two years. Complex interplay of numerous variables can be best studied via identification of pattern to the data. According to the McKinsey Report, by 2018, the United States will experience a shortage of 190,000 skilled data scientists and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge. This prediction from 2013 has been shown to be accurate when looking at the current data science trends. Data science involves many technical skills, including programming (R, Python, Scala), machine learning (scikit-learn, nltk, MLLib), tools (KNIME, WEKA, H_2O, Orange, Tableau), framework (Hadoop/HDFS, Apache Spark), and database (Oracle/MySQL, MongoDB, Cassandra, GraphX). A typical data science team involves a data scientist, data engineer, data solution architect, data platform administrator, and possibly a designer and/or a product manager/owner.

Machine Learning

Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed. There are two types of machine learning: supervised and unsupervised. There is also semi-supervised and reinforcement learning. A common problem is how to determine if the model is fully trained. Another problem is how to best fit the data. The cost function is $\frac{1}{2m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$, where m is the number of samples. Confusion metrics can be used to determine how accurate the model is.

Artificial Intelligence

PCS++ is an artificial intelligence approach. There have been at least five known cases of fraud worth over $1 billion in the last two years. SWIFT has been using machine learning to determine which transactions should be flagged.

Jupyter Demonstration

The guest lecturers from SWIFT showed their preprocessing techniques and how they used machine learning to analyze data. They used just about everything that we had either talked about or learned in class.

Questions

Q: How do you know if problem is suitable for machine learning?

> A: You know when it is suitable to use machine learning if your goal is to recognize patterns in the data.

Q: How do you deal with live streaming data?

> A: Online training occurs when you want to use machine learning on live streaming data. The new, incoming data becomes part of your training data.

Q: After deployment of a model, when do you update the model?

> A: You only tolerate a certain number of false positives, and at that point, you update the model to improve its accuracy. It also depends on how critical the application is and whether it should be updated weekly, monthly, etc.