Week Four Lecture Videos

Indiana University Bloomington: Big Data Applications

Naimesh Chaudhari: naichaud@iu.edu

Assignment 4 – October 10, 2019

1. Abstract

In this paper, I provide a brief introduction about the information presented in the Week 4 lecture videos. I also provide additional information on a subtopic that I am interested in from the videos, ideas that could improve some of the existing work, and finally a deeper dive into a specific video segment I am interested.

2. Summary

In the lecture videos of week 4 we start discussing how big data applications have been utilized in the field of Physics. We specifically discuss the Higgs particle study. To do this experiment many countries in the EU collaborated to develop a very large accelerator underground. The accelerator is 17 miles in circumference and 175 meters underground. The project is generating close to 15 petabytes of data per year! This was a cloud collaborative project with 3,600 people working on it simultaneously. Close to 500M was spent to build the required infrastructure. There are four main projects within the larger set they are CMS, Alice, Atlas, and LHC. This undertaking started close to 15 years ago and was built on a 350,000 cores computing grid.

Inside the tunnel there are strong magnets that accelerate the electron and proton particles. There are sensors within the tunnel that measure information when particles collide. One such sensor/magnet measures the magnitude of a collision. This information is then taken into a scripting language to analyze. We reviewed multiple scripts in both Java and Python to see effects on randomly generated data. We learned that and IID event is called a random variable in the terms of statistics. We learned about the normal distribution and how to use statistical data to bring insight.

In section 4 we dived a bit further to learn how generators create random numbers and what seeds are used for. We also learned about two other statistical distributions. The binomial and the poisson distribution. We learned how we can accept/reject methods to generate random variable with arbitrary distribution. Monte Carlo method is an example of a method that uses the accept reject method.

We discuss the central limit theorem which states that sampling distributions of sample mean approach a normal distribution as the size gets larger. The shape of the population does not matter, and the size gets increasingly larger.

Finally, we discuss two major trends of thought in probability, Bayesian statistics vs frequentists. In Bayesian approach we believe we can predict future events based on historical events which in the frequency approach we look at how often events have occurred. Usually in Bayesian approach we have an estimate. We discuss the Bayes Law that helps us derive “belief probability” through the frequency data.

3. Interested Subtopic

On thing that I found interesting was the Monte Carlo method. I have head of this

method but did not have much familiarity with it. Wikipedia states, "*Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use other approach."*

A simple example of this in mathematics is to the figure out the integral of the function sin(x) from o to pi. We know this is 2. We can simulate this result using the Monte Carlo method. We would generate many sets of random numbers uniformly distributed between 0 and pi. Then we can push them through the function to get a vector that we can sum and calc the volume. If we do this 1000 times, you would see that the distribution of results will be normally distributed with the number 2 being in the middle.

Monte Carlo has been used in other fields as well. Two interesting fields are sports betting and stock prices, where generating a Monte Carlo simulation x times can give you the probability distribution of an event happening which can be used for future predictions.

There are other methods of the monte Carlo that are more directed. One example is the Markov Chain Monte Carlo. The method attempts to model a specific distribution by taking a directed random walk. Once it finds a region of higher distribution, it tries to stay on the same trajectory. By doing this it will over many iterations get very close to the initial probability distribution. This process is called the stochastic process.

4. Ideas that can improves existing work

Over researching Monte Carlo simulation, I got excited about all the different possibilities that are out there. I would love to see and read about how other people have used these methods to implement in production.

One thing that would help me better grasp the theory is to be able to see live examples of these processes. In researching online, I saw many theoretical papers highlight how this is possible, but very little focusing on the coding side.

I believe every individual has his strengths and weakness, for me personally, reading theoretical papers and converting them to code is not one of them. I learn through example and having concreate examples would help tremendously.

Also, I would like to see more research and implementation of these processes on the business side. I see example of implementation in education studies but not many towards the business. Being in a shipping and packaging business I often struggle in linking statistical methods to business problems. Having a wider range of articles that highlight business problems and the statistical methods used to solve them would help tremendously.

Overall, I think the work out there is very helpful in understanding the concepts. It helps pave the way for us students. I hope in the future I am better able to learn from these and apply towards my future projects.

5. Video Deep Dive

Unit 11.8 Part IV

Interpretation of Probability

Bayes VS Frequency: Minutes 0-12.38

As stated in the video there are two types of trend of thought out there when it comes to probabilistic approaches, frequentist vs Bayesian. What people most often forget is that one builds on top of another. I will be reviewing both approaches in dept in these final two pages.

Frequentist

As best stated by Aristotle, "*the probable is which for the most part happens*." In mathematical terms an events probability is determined "as the limit of relative frequency in many trials." The key statement to understand is "many trials". If measuring the probability of an event one needs to be able to do many trails to determine which result is most probable.

In sports we can think of this as the number of times and event happened. Let's set our event to be how often did the IU basketball team win last year. We can do this quickly, we can pull up the information on the number of games played last year, and the number of games won. If we take the ratio between the two, we get the frequency of how often IU's basketball team won last year. In the 2018-19 season the IU team played 19 games and won 16 of them. The ratio between the two would be 84.21%.

Let's say now we wanted to estimate what is the probability that the IU will win this year. The most naïve way would be to say well since IU won 84% of the time last year, they are most likely to win 84% of the time this year. This is my basic understanding of using frequency to predict the future. approach. We can further extrapolate this by taking win percentages of many games. Let's say we take the win percentage over the last 1000 games. If we break the percentages out over a range like say 10-20, 20-30, 30-40 and so on, we can select a range that has happened most often. This would be the frequentist approach with many experiments. In our case the event is IU winning and the experiments are historical games already played.

The second scenario I would like to set up is let's say, we were watching an IU game live. It is the end of the third quarter, and we wanted to know what the odds are for IU to win this game. Similarly, to before we can blindly say well IU wins 84% of the time so there is an 84% probability that they will win this one.

Many sports fan would argue that this is a very bad prediction. Why you ask? Well there can be many different events that can affect the output of our final event. Examples are, which players are on the team, which players have scored so far, which players have not scored, how many points has IU scored in the 1^{st} 2^{nd} and 3^{rd} quarter, how many points has the opposing team scored so far, is it a home or away game, the list goes on and on.

Bayesian

This is where the Bayesian approach of things comes in. We can predict the future based on past events. We can generate a "belief" like the probability of winning this game. I will review some of the basic concepts of the Bayesian approach and then visit our above example again to so how we can improve our accuracy.

To start with Bayesian statistic was introduced to the world by Thomas Bayes in 1763. The Bayes theorem sates that the conditional probability of event A occurring given that event B has occurred is probability of event B occurring given event A has occurred multiplied to the ratio of the probability of event A over the

probability of event B. Below is the theorem.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

I would like to set our problem before in these terms. Let's define out parameters. Let's say today the different between the two games is +10. Historically there have been 11 games where the difference has been between +7 and +10. and out of those 11 games we have won 6. With this information we can define our problem below

Event A = Probability of winning = 84%

Event B = Points difference at the end of 3rd quarter between +7 and +10 = 11/19 = 57.89%

Event(B/A) = 6/16 = 37.5%

Now if we apply the above information to the Bayes theorem, we get an answer of 54.54%. Which is much lower than our initial naïve prediction! If for example our actual wins where event B occurred goes up to 10 then we have a probability of winning at 90%.

This intuitively makes senses to me. Knowing the relationship between such event makes us better predictors. This is were we can see how we use the frequency data to determine our belief on the future outcomes.

Once the Bayes theorems (Thomas Bayes) and the Bayesian interpretation of probability (Pierre-Simon Laplace) were published in the scholarly articles, they were adopted by many individuals from the mid-18th to 19th century. In the 20th century this approach took a downturn due to its nature of computation. If we can imagine our simple example before with 1000's of parameters, it would take a very long time to predict the probability of winning by hand given those parameters. Theoretically we can do it but doing this computation would be very difficult by hand. In the 21st century Bayes approach was revisited and started to get adapted more often. This was due to the development of very powerful computers and new algorithms like the Markov Chain and Monte Carlo.

Bayesian statistics has expanded to many different fields today. The one Big Data area I would like to discuss where Bayesian statistic inference is new is in Deep Learning. One of the common problems we face in deep learning is understanding model certainty. Due to the nature of the architecture there is no real measure that lets us know what the uncertainty of a prediction is. This can be very dangerous in the medical field, for example if I am classifying whether a person has cancer or not, I would like to be very certain that if I tell my patients they don't have cancer its true. In recent years there has been, and emerging idea called Bayesian Deep Learning, which tries to predict the model uncertainty of and event.

5. References

1. "Monte Carlo Algorithm." *Wikipedia*, Wikimedia Foundation, 19 July 2019, en.wikipedia.org/wiki/Monte_Carlo_algorithm.

2, "Markov Chain Monte Carlo." Wikipedia, Wikimedia Foundation, 10 Oct. 2019, en.wikipedia.org/wiki/Markov_chain_Monte_Carlo.

3 "Bayesian Statistics." Wikipedia, Wikimedia Foundation, 28 Aug. 2019, en.wikipedia.org/wiki/Bayesian_statistics.