Discussion Summary

Michael Wacey

For Doctor Zachary Dennis

SEAS 8510: Analytical Methods for Machine Learning

June 1, 2024

George Washington University

# Request

In at least 500 words in an essay format, describe the major topics covered in the discussion groups. What were some of the most useful insights from your class members? What are key takeaways that will be useful in later machine learning courses?

Do not use generative AI technologies to produce your essay.

# Overview

The discussion questions were aligned to but extended the material being covered in class. It gave the students an opportunity to extend their understanding. I found this interesting and very useful. Oftentimes each student would find a subtly or even extremely different way of addressing the question. This was very important learning for me. I sometimes get focused on the idea that there is only one way to do something.

Below is a copy of the discussion topics, my insights and takeaway, and conclusions.

# Topics Covered

This is the list of discussion questions that were presented.

## Discussion 1

Describe your past experience and level of knowledge in machine learning, linear algebra, statistics, and Python. How have you used machine learning to help you in a professional role? Comment on how you think improving your knowledge of machine learning will aid you in your career. Provide thoughtful feedback to at least two other students in your discussion group.

## Discussion 2

1. Describe a real world example of a system of linear equations.

2. Explain the concept of a Vector Subspace

## Discussion 3

Describe how each of the following algorithms work from a high level and describe at least one use case for each – (a) singular value decomposition, (b) principal component analysis, (c) linear least squares regression, and (d) network analysis.

For people with practical experience in machine learning, feel free to provide real world examples of implementations on real or fake data sets besides what was shown in class.

## Discussion 4

Explain why and how Bayes Theorem is useful for making better predictions in statistics. Use your own examples to illustrate your points. Also explain how Bayes Theorem is useful for Naïve Bayes and Decision Trees algorithms. Students experienced in programming and machine learning are welcome to share their own algorithm implementations and show where Bayes Theorem is used.

## Discussion 5

Implement a Naïve Bayes classifier. It’s okay to leverage examples from online. Explain how it works and give feedback to other students in the discussion group on their implementations. Use different data sets from other students in your group.

## Discussion 6

Using Python, make up an example of a real-valued random variable X using a cumulative distribution function. Plot the PDF and CDF of your function. Please use different functions and examples from other students in your group. Provide feedback to two other students on their implementations.

## Discussion 7

Create a data set using a random distribution function of your choice. Create another data set based on the first data set that is positively correlated, and create a third data set based on the first that is negatively correlated. Find the covariance between the first and second, between the second and third, and between the first and third. Explain your results and using graphs of the data as necessary to explain.

# Insights and Take Aways

The first discussion was a good chance for the members of the discussion group to get to know one another. In my group, two of the other members had very similar backgrounds to me in that we worked on data intensive solutions. The other members had varied backgrounds that I found to be interesting.

The next discussion was on systems of linear equations and vector sub spaces. I enjoyed seeing everyone’s real world examples. When I took linear equations in undergraduate school, the thing that stuck with me was that they can have 0, 1, or many solutions. This has stuck with me, and I see it in everything I do. People seem to think that there is always a single solution to any problem. But most problems in life really are sets of linear equations and I think most have zero or many solutions. Very few have a single solution.

For the next discussion, most of the responses were focused on providing detailed descriptions of SVD, PCA, Least Squares Regression, and Network Analysis. In contrast, the real-world examples were brief. For me, this reiterated the importance of these techniques and understanding them in detail.

Discussion four was our first foray into Bayes and Naive Bayes. I learned a lot about these from the other students’ posts. But it made me wonder what I did not know. So, I spent considerable time reading about Bayes and Naive Bayes. Everyone talks about how important these are to probability and machine learning. Even with my reading, I do not think I fully appreciate their importance. So, I need to do more to fully appreciate them. It seems to me that the two best ways to learn something are by using it and by teaching it. Hopefully, I will get to do both in the future.

We continued a deeper dive into Bayes in discussion five. Here the request was to “implement” a Naïve Bayes Classifier. There were two distinct approaches to what the word implement meant. For some, I mean to use a Naive Bayes library and implement it for some data. For others it meant to implement the Naive Bayes algorithm in Python. I took the later approach. It was a fun exercise and gave me a chance to better understand Naive Bayes and Python. I also found it interesting that you need to select a distribution when using Naive Bayes.

In the next discussion we were to use Python to ‘make up an example of a real-valued random variable X’. There seemed to be two interpretations of this. I think it stemmed from the use of the term real valued. I took that to mean that it is a continuous random variable and not binary, integer, or such. Others seemed to take it to mean that it was a real-world example. I found it interesting that we were starting with a CDF and determining the PDF from that. It took me a little bit to make sure that I was doing it right.

The last discussion involved creating random variables with specific correlations. Everyone in my group generated data rather than using real world data. This ensures we can create the requested correlations. I used normal distributions for the data but had the mean values vary linearly. Other people used different schemes. In general, the first two cases where A and B or A and C were plotted had a well-defined scatter plot. However, the scatter plot for B and C was less well defined. I found that interesting.

# Conclusion

I found the discussions to be beneficial in this class for several reasons. First, they reinforced the material that we were covering in class. Second, they gave me an opportunity to see how other people address the same problem. This was very useful for broadening my approach to problem solving. Lastly, I enjoyed interacting with the other students both in the online forums and the breakouts at the start of class. Overall, a great learning experience.