IS 507 - Data, Statistical Models, and Information - Discussion 5

1. Explain one new concept you learned in the second half of the course (since Week 7).

- Some very interesting concepts were covered during the lectures since Week 7. However, I found the concept of 'Principal Component Analysis' interesting.
- In the current century, data is being generated at a very fast rate and it is voluminous. Analyzing large datasets having a large number of features or descriptors can be a complex and cumbersome process.
- In such cases, it becomes necessary for us to reduce the size of dataset in a manner such that maximum mount of the information in the original dataset is retained in the compressed version using a smaller set of features.
- This is termed as Dimensionality Reduction and Principal Component Analysis is a techniques which can help us achieve it.
- This concept is exciting because the rate at which data is generated has increased and its volume is also high. Working with large datasets can be difficult as it can lead to issues related to data storage. By employing PCA, we can atleast reduce the dataset dimension which means the same data in a compressed form.

2. How can the concept in Question 1 be used in your field of interest? What research questions could it answer?

- Principal Component Analysis can be used to reduce the number of features or attributes present in the dataset so that it becomes easy to work with and at the same time retain maximum amount of information in original dataset.
- Let us take the problem from the healthcare industry, where the clinical studies utilize electronic healthcare records (EHRs). These datasets are huge in size and can contain a large number of features.
- There is a high chance that there can be features like Vitamin A, B, C and so on which can be highly correlated.
- Or there can be features which has been split into different categories and represented as an encoded variable like chest pain can be of 4 types - typical angina, atypical angina, non-anginal, asymptomatic. These can be 4 different features but can be reduced using PCA.
- After reduction, it can be used to interpret and name the new set of features obtained based on which columns contribute into the different principal components. Like all the chest pain types can be clubbed to 1 single variable - Chest Pain Type.
- It can be used in the areas where healthcare works with images. Images related to scans of the brain, MRI scans can be of high dimensions and to reduce the image size for quicker analysis is where PCA can help.

3. Where do you see Statistics, Data Science, AI, or Machine Learning going in your field of interest in the next 5 years?

- The modern healthcare system just faced a huge crisis when the entire world was hit by the Covid-19 pandemic.
- Artificial Intelligence and Machine Learning have been employed in the Healthcare field. For example, based on the dataset of any specific disease like cancer, cardiovascular issues etc, robust classification models and predictive algorithms have been built to identify if the patient has any disease or not.
- AI can also be used in automated disease identification given the statistics of the current patient condition and also predict the best possible treatment for specific condition.
- AI can also be used for automating the routine work done by clinical staff including nurses which may include the routine follow-up for checking the vitals of the patients and providing them with medicines at specific time intervals.
- Talking about Deep Learning and Neural networks they can be used in Computer Vision techniques for identifying the possible issues and sources based on the different body scans taken.
- Natural Language Processing can be used in developing chatbots which can provide an overall service to the patients and also act like a Virtual Assistant.