Dimensionality Reduction

*Note: Sub-titles are not captured in Xplore and should not be used

Asadujaman Nur
Electronic Engineering
Hamm-Lippstadt University of Applied Science
Lippstadt, Germany
Asadujaman.nur@stud.hshl.de

***Abstract*—Dimensionality Reduction is a part of machine learning. which involve methods to make Machine learning process possible. In this paper we are going to talk about what machine learning is, its history, how it works. and about Dimensionality reduction what it is, how helps machine learning process. Its Advantage and Disadvantages. And finally about its future and how it will help to shape the future of AI and Machine Learning.**

***Index Terms*—Machine learning, Dimensional reduction, unsupervised, research paper**

I. Introduction

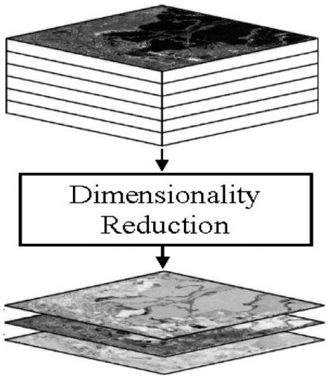


Fig. 1. Dimensionality Reduction. [6]

We know human learns from their past experiences. and machine follows the instruction given by human [1]. What if we could train the machines to do the same how human does? well we can, and this is called machine learning. just like human learn form experiences [1], machine learns from its provided data. by carefully feed,annualized and train the data to the machine, it can do it analysis and predict outcome of an task.

People creates huge amounts of data, in fact 90% of data has been produced in last 3-4 years thanks to the globalization and IoT,right now not only human generates data but also the Devices and smart tools. such as smartphones, computer, sensors etc. and this data could be used to train AI and feed to machine learning to achieve various task [1]. however there is a slight issue, because there are huge amount of data it also requires a lot of space and processing power to properly utilized the data. in-order to achieve certain goal we can use certain and useful type of data to teach the machine learning algorithm and the rest we can ignore. thus it will increase the precision and speed of the outcome as well as reduce storage requirement at the same time. now the question is how can we do that? among other methods dimensionality reduction is one of the best to methods for data reduction. lets have a quick look at dimensionality reduction.

Dimensionality Reduction is a part of machine learning where it helps the huge machine learning process to be possible among other other methods(such as Decision Tree, Random forest K-Means Clustering etc.). we are going to breakdown dimensionality reduction process later on in this paper. but for now in simple terms, dimensionality reduction mainly used to Reduce the load of unnecessary or irrelevant features/Data to save space and make faster data processing. To reduce irrelevant data, feature selection Methods, Matrix Factorization, Manifold Learning etc. are the name of the few techniques that dimensionality reduction uses.

II. History

Data is a vary powerful asset, therefore human started to record the data inform of information since the ancient time starting on wall painting, then clay tablet and later on paper. by recording the data it has allowed us to know about our past and history. And based on the data we have, we cloud predict the future and make efficient plan for the better future. every time the civilization got upgraded to a new technological era, the method of storing and analyzing data got updated as well. after the industrial revaluation the civilization witness many technological wonders being invented. one of the biggest and best technological invention was Digital computers. since then we have been generating more and more data. which can not be analyze by traditional way.

The idea of Machine learning came to light in 1950 [2], when a computer scientists "Alan Turing" publish a paper answering the question, "Can a Machine think?",There he stated that when a machine can successfully convenience a human,that it is not a machine will reach artificial

Identify applicable funding agency here. If none, delete this.

intelligence. in the year 1957 ”Frank Rosenblatt” first computer neural network which was designed to classify visual inputs into groups. In 1959 two computer scientist named ”Bernard Widrow” and ”Marcian Hoff” created two neural model which could detect binary pattern and eliminate echo on phone lines [2]. As Technology keeps evolving the machine learning method shifted from knowledge base approach to data based approach [2]. luckily because of the huge advancement in technology gathering data wasn´t an issue. since everything from books to business started to moved online the amount of data started to grow along with it.

Machine learning has come a long way since half a century ago from a concept to competing, human in games and events. IBM Watson beat 2 human in game of Jeopardy in the year 2002 [2]. after that googles Alpha Go which is the next generation of Machine learning and AI beat a professional Human Champion in the year 2016 [2]. since then a lot of companies started to invest and involved in machine learning technology. Google brain, open AI amazon are the name of the few [2]. among them google is heavily invested in machine learning and AI technology due to their search engine platform and Add business [2]. one of the greatest achievement of modern AI is Digital assistant. where google Assistant is being the dominant of them all. Google assistant uses AI technologies such as neural language processing, machine learning to analyze the data from user to provide replies.since goggle is one of the giant data collector and have the best minds of the world the machine learning and Artificially Technologies there are improving exponentially.

III. Methods

Before we talk about what kind of methods Dimensionality Reduction uses, lets get familiar with what dimensionality reduction really is?
Dimensionality reduction or Dimension Reduction is the transformation of data from higher dimension to lower dimension by trimming the unnecessary data. so that the transformed data can retain meaningful properties to the original data.Dimensionality is the part of Unsupervised Learning which is another part of machine learning. in unsupervised learning the data which used to train the model are unlabeled.Since the model is unaware of the label of the training data, there is no way to measure its accuracy. so this methods doesn’t follow accuracy. rather it uses clustering to analyze The method of dimensionalitly reduction [3] can be divided into 2 component or part [3].

1) Feature Selection.
2) Feature Extraction.

Feature selection is also known as variable selection. which is the process of selecting similar features to use in the model. where feature extraction is the process of transforming raw data to numerical input,which can be processed to preserve the data to its original source.
Since we have an idea of Dimensionality Reduction, there are lot of methods that can be applied for dimensionality reduction. and they are so depth that they deserve their own paper. however for the sake of this paper, We are going to discuss about Principal Component Analysis also know as PCA, since it is one of the most popular and widely use method.

A. Principal Component Analysis(PCA)

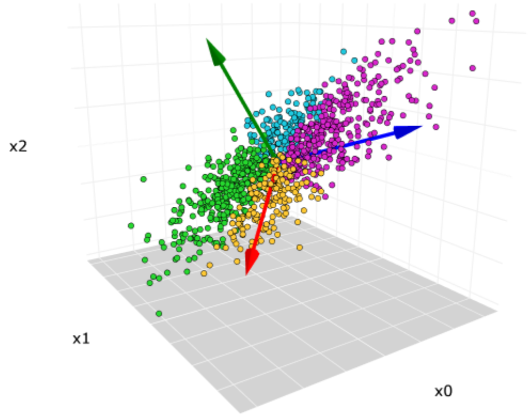
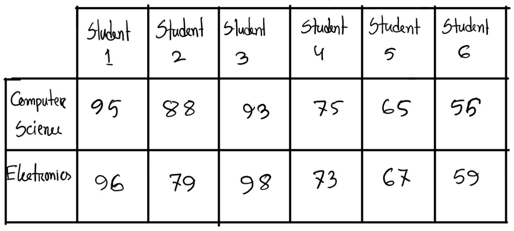


Fig. 2. Principle component Analysis [7]

In simple words The method principle component analysis [4] takes the data from higher dimension (3D for an example) and squeeze it into lower dimension(2D for example) on a liner path. PCA is around for a really long time since 1901 pearson paper [5]r. PCA is also known as Statistical interpretation [5]. It supplies us Data driven Hierarchical coordinate system that captures maximum amount of variance in data [5]. Its better to explain PCA method with and example according to [4], for that we are going to follow the data sheet bellow. where it is showing how many points did the student earned during the exam. Here in the Fig.3 we can see there are 6 students who participated on 2 exams which are ”Computer Science” and ”Electronics” and theirs score respectively. if we were to represent the data on ”one dimension”, it will look like the Fig.4 bellow. Here on the graph(fig.4) we can see that the student with relatively lower scores in the subject ”Computer Science” are clustered together on ”Lower Score”. who are student number: 4,5,6 and the student with higher scores are clustered together on ”Higher Score”.even tho this is a simple graph we can see that Student ’1,2,3’ are similar than student ’4,5,6’. With this method we can represent any data from a data sheet to a one dimension form if necessary.

Now lets have a look in case of 2 dimensional data representation . here on the Fig.5, we may analyze the data sheet again and carefully map the data into 2 dimensional Graph. Here we can map the subjects Computer science



	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6
Computer Science	95	88	93	75	65	56
Electronics	96	79	98	73	67	59

Fig. 3. Data sheet of Students Scores. [4]

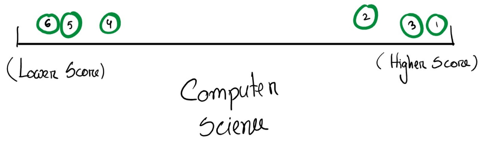


Fig. 4. 1 Dimensional Data representation of Computer Science. [4]

as "X"-Axis and Electronics as "Y-Axis". on "X-Axis" (Computer Science) the farther you move towards right the higher its score level. thus it is labeled as "Higher score" and the most left has been labeled as "Lower Score" now we can place the student on the graph based on their result.

For the subject Electronics which is mapped as "Y-Axis" on the graph, Same principle has been applied as before. the students with higher scores are placed on nearly top labeled as "Higher Scores" and the student with lower grades has been positioned on the bottom of the Axis labeled as lower scores.

Now if we carefully analyze the graph and data sheet (fig.3) we can place all the students based on their grade on this 2-Dimensional plane (fig.5). After relocating the data from the data sheet(Fig.3) to the graph(Fig.5) we can notice a pattern here. we can see that students with lower scores (student 4,5,6) with both subject(Computer science and Electronics) are closer to each other than the students with higher grades. we can see the similar pattern for the students with higher points as-well. the students (1,3,2) are also closer together than their counter parts. to make it even easier to grasp we can form two groups or clusters. they can be named as.

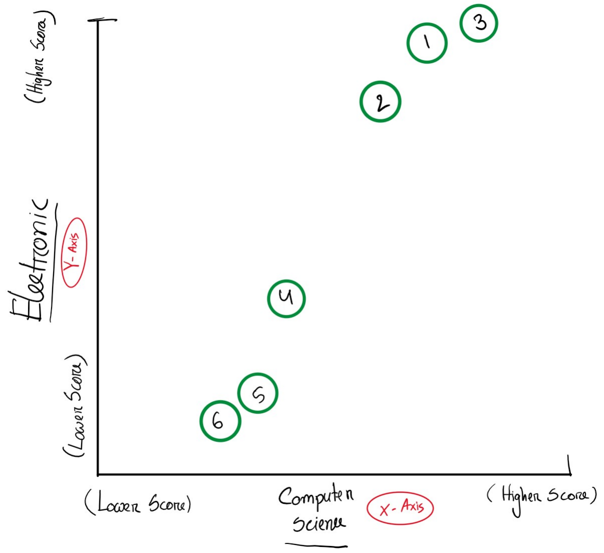


Fig. 5. 2 Dimensional Data Representation of Computer science and Electronics [4]

1) Students with Lower Grades.
2) Students with Higher Grades.

Since we have been able to successfully re-positioned the data from the "Data Sheet"(fig.3) to the Graph(Fig.5) and can read the data clearly, we no longer need the "Data Sheet" (fig.3) to follow the rest of the process of "Principle Component Analysis". From now on we will focus on the graph(Fig.5) and find a way to compress and represent the data from 2 dimension to one Dimension. To so we have to follow some the steps bellow.

First we have to find an average of the Data for the Subject "Computer Science" or The "X-Axis" and the average of "Electronics" Or "Y-Axis" and then mark it on the graph(fig.6). once we find our average, we can find the center of our data in the graph(fig.6). after that we can shift the data by placing it on the the origin point shown in fig.7. shifting the data doesn't change its orientation, it just changes its position. now since the data is positioned PCA [4] will Draw a line that best represent the data given that the line will go trough the origin(Fig.8). now lets have a look how PCA decides if the data is the best fit or not?

to specify how good the line fits the data [4] PCA projects the data onto it and then it could either measure the distance from the data to the line and to find the line that minimize or maximize those distances from the projected point to the origin(Fig.8). For the PCA mathematical explanation can be found on [4].but in-short it uses eigenvector,Singular vector and other mathematical methods.

The other liner and non liner methods like lasso

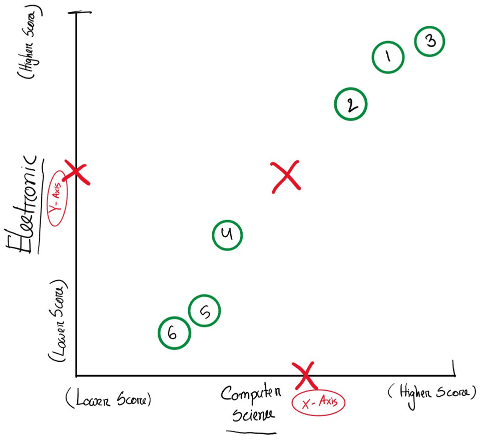


Fig. 6. Average score. [4]

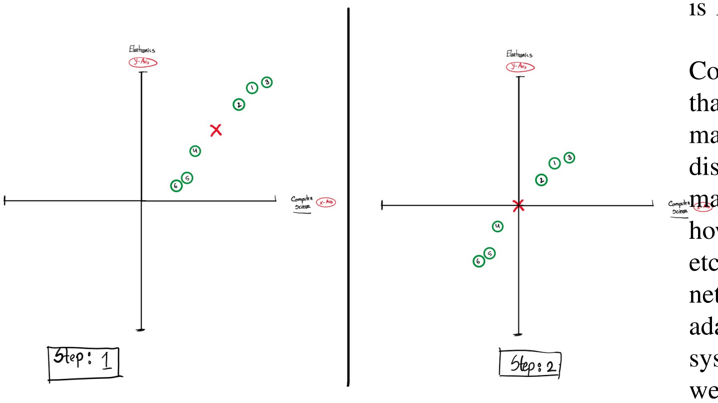


Fig. 7. Data Relocation. [4]

regression,Ridge regression, Linear regression, kernel pca, t-SNE, and MDS are also used in Dimensionality reduction [10]. for the sake of the paper there are some short definitions for few methods belwo. according to [8] ” Lasso regression is a type of linear regression that uses shinkage.shinkage is where data values are srunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (example: models with fewer parameters).” to talk about Ridge regression according to [9] ”Ridge regression is a model tuning method that it used to analyse any data that suffers from multicollinearity . this method performs L2 regularization”.

IV. Future of Dimensionality Reduction

We have Come a long way in a short time span from the era when a single computer would cover a whole room and could perform only the basic computation. to an era where the computers are being carried in our pockets.and the way we use computer has changed a lot as well. gone the days where the computers were only used for official purposes. now it has become a in separable part of our life. and assist us solving almost any problem imaginable by providing vital information, thanks to another revolutionary technology known as internet. which was born from as a next step and main communication medium between computers. The evolution of computers and technologies has not stop yet. It is keep evolving on an exponential rate.

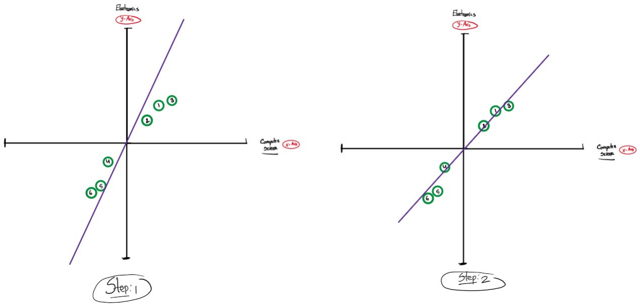


Fig. 8. Projecting Data to PCA. [4]

Computers are getting smarter and developing new skills thanks to the greatest minds in our world. and it has been made possible because of the machine learning just like we discussed before in the section ”Introduction & History”. the machine learning process was models after human behaviour, how they behave,learn from experience,try to solve problem etc. also it was mapped by following human brain or neural networks. Using machine learning computers can learn adapt evolve. for that they need data.Thanks to IoT,Embedded systems and Internet it is not an issue. we have more data than we can handle. in the past the data were mainly text based, numerical or graph based. However its not the same anymore. data can be found in all form. From simple numbers to photos and videos etc. Having this much of data is is hard to process since it takes a lot of processing power(Section:Introduction) and space.Therefor we use dimensionality reduction, where we keep only the data which is relevant to our purpose. Thus it makes our computation and Machine learning process faster.

since the industrial revolution happened, we´ve been trough some version of Industries. currently we are in ”Industry v3” some propose we are on ”Version 3.5”. And we are moving faster then ever before to the next Version of Industry.which is ”industry 4.0”. Where everything will be automated and connected to internet.From a coffee-machine to factories.To make it possible machine learning will be used almost everywhere.Since we will use a lot of automated processes there will be a lot of data which will need processing. not only the ”industry 4.0”, Web Version3, Augmented Reality, Virtual Reality,wearable Tech,Autonomous Vehicles etc. is on their way to be a part of our life. there we will generate even more data with different new kinds of information.(for example to run augmented reality application, we need to map our real life world with depth in addition to normal data.) which will enable machine learning and AI process even better and faster. Since we will be exposed to a lot of new kind of technology and data. And Machine learning is the future, To make this process faster and efficient we will need Dimensionality Reduction even more then ever before. There might be a new method along the way, only sky is the limit.

V. Conclusion

So Far we have been discussing about Dimensionality Reduction and one of its principle method named ”Principle Component Analysis, and got short glances at some others methods as well(Section:Method). We have learned its an part of unsupervised Machine Learning approach (section:Introduction) . We´ve also learned its history and how it parts with machine learning Algorithm(section:History).as well as had a glance of its possible future involvement to betterment of humanity and how its going to play a bigger role in future of AI and Machine learning techs as well as industry 4.0.

Since we´ve been talking about dimensionality reduction lets talk about Some of its Advantages: According to [11] the Advantages are:

- It makes data compression possible, therefore required reduce storage space. [11] [12]
- It helps Reduce computation and data processing time. [11] [12]
- It also assist removing redundant Feature, if there is any. [11]

Although it sounds some kind of miracle technology, Despite having a lot of benefits it have some disadvantages as well. just like a coin has two parts.We are going to address some of its disadvantages down below:

- Since it compresses data, There are some form of data loss. [12] [11]
- ”PCA tends to fin linear correlations between variables.That can be undesirable sometimes.”(according to [11])
- ”PCA fails in cases where mean and covariance are not enough to define data-sets.”(According to [11])
- ”We many not know how many ’Principal components’ to keep- in practice,some thumb rules are applied.”(according to [11])

These are some of the advantages and disadvantages among many, which gives us basic ideas about Dimensionality Reduction.

Finally in this paper we discuss about basics of dimesionality Reduction and got familiar with this topic along with machine learning and its origin. although this paper does fine job scratching the surface. it is a vast topic that can be expanded far beyond. more about this topic can be found by following the Reference below.

With this conclusion we draw an end to this paper.

References

[1] Machine Learning Basics — What Is Machine Learning? — Introduction To Machine Learning — Simplilearn. (2018, September 19). YouTube. $https : //www.youtube.com/watch?v = ukzFI9rgwfU\&ab_channel = Simplilearn$

[2] A Brief History Of Machine Learning — Machine Learning For Beginners — Simplilearn. (2019, August 23). YouTube. $https : //www.youtube.com/watch?v = lY1SELMvoVs\&ab_channel = Simplilearn$

[3] Machine Learning - Dimensionality Reduction - Feature Extraction Selection. (2017, April 21). YouTube. $https : //www.youtube.com/watch?v = AU_hBML2H1c\&ab_channel = CognitiveClass$

[4] StatQuest: Principal Component Analysis (PCA), Step-by-Step. (2018, 2 april). YouTube. Geraadpleegd op 8 mei 2022

[5] Principal Component Analysis (PCA)-Pearsonpaper. (2020, January 28). YouTube. $https : //www.youtube.com/watch?v = fkf4IBRSeEct = 144sab_channel = SteveBrunton$

[6] A. (2021, December 14). Importance of Dimensionality Reduction!! - Analytics Vidhya. Medium. https://medium.com/analytics-vidhya/importance-of-dimensionality-reduction-d6a4c7289b92

[7] Cheng, C. (2022, March 22). Principal Component Analysis (PCA) Explained Visually with Zero Math. Medium. https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

[8] S. (2021b, April 27). Lasso Regression: Simple Definition. Statistics How To. https://www.statisticshowto.com/lasso-regression/:

[9] A Brief History Of Machine Learning — Machine Learning For Beginners — Simplilearn. (2019, August 23). YouTube. $https : //www.youtube.com/watch?v = lY1SELMvoVsab_channel = Simplilearn$

[10] Pramoditha, R. (2022, January 7). 11 Dimensionality reduction techniques you should know in 2021. Medium. https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b

[11] GeeksforGeeks. (2018, February 8). Introduction to Dimensionality Reduction. https://www.geeksforgeeks.org/dimensionality-reduction/

[12] Dimensionality Reduction : Data Science Concepts. (2021, August 18). YouTube. $https : //www.youtube.com/watch?v = 6XGlqR6rcpUt = 286sab_channel = ritvikmath$