Linear Algebra in Data Analysis

Mohammed Alghudiyan

Department of Mathematics, King Fahd University of Petroleum and Minerals

MATH 225: Intro. to Linear Algebra

Dr. Ibrahim Al Rasasi

December 10, 2024

Contents

1 Introduction

Linear algebra plays a key role in data analysis by providing the mathematical foundation for many techniques used in processing and interpreting data. It helps with transforming data, solving systems of equations, and reducing dimensionality. Its applications allow data scientists to manipulate large datasets efficiently, making it an important tool for making accurate predictions.

Note, all the calculations and plots have been made using Python language with the help of pandas, numpy, matplotlib.pyplot and seaborn libraries.

2 The Correlation Matrix

The correlation matrix is a powerful tool in data analysis that helps to identify the relationships between different variables in a dataset. By analysing the matrix, one can see how strongly variables are connected, which is important for understanding patterns and making predictions.

2.1 Introduction

If we have a numerical dataset with n features and m index and we form an m x n matrix A, then if we want to know how i^{th} column is related to the j^{th} column, we may know this from the value of $cos(\theta) = \frac{\vec{x}_i^{\,T}\vec{x}_j}{|\vec{x}_i|\,|\vec{x}_j|}$. However, to get more information about the data, we may construct the n x n correlation matrix. First, we scale the columns of A and define an m x n matrix U where $u_i = \frac{\vec{x}_i}{|\vec{x}_i|}$, so that $cos(\theta) = \vec{x}_i^{\,T}\vec{x}_j$. Then the n x n correlation matrix is defined as follows: $C = U^T U$. The c_{ij} value represents how the i^{th} column of A is related to the j^{th} column of A. The entries of C would be between -1 and 1. Refer to the following table:

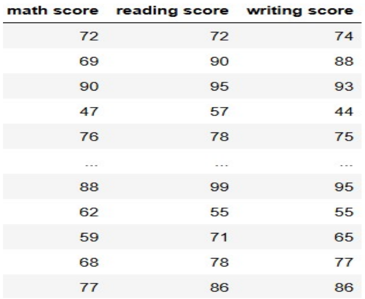
Correlation Value	Indication
± 0.8 to ± 1.0	High Correlation
± 0.6 to ± 0.79	Moderately High Correlation

± 0.4 to ± 0.59	Moderate Correlation
± 0.2 to ± 0.39	Low Correlation
± 0.0 to ± 0.19	Negligible Correlation

A perfect correlation, i.e. 1, between $\vec{x}_i$ and $\vec{x}_j$ would satisfy $\vec{x}_i = \alpha\vec{x}_j, \alpha > 0$.

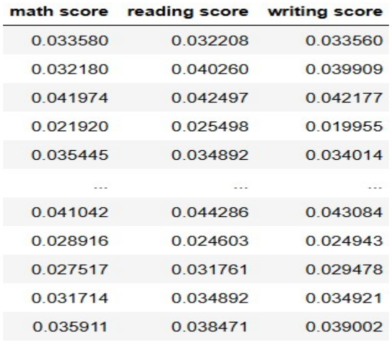
2.2 Example 1: School Grades

It is better for schools to know what affects students' grades. We will use a dataset containing students' math, reading and writing scores. Our matrix $A = [\vec{x}_1 \ \vec{x}_2 \ \vec{x}_3]$ would be:



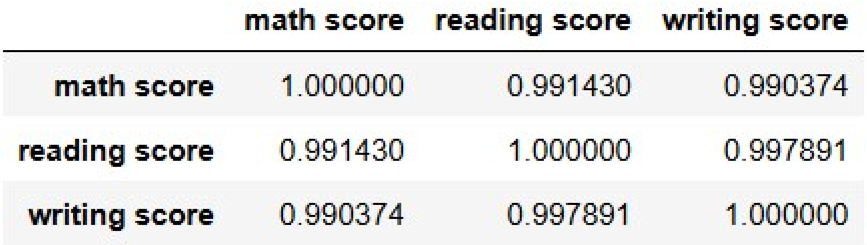
math score	reading score	writing score
72	72	74
69	90	88
90	95	93
47	57	44
76	78	75
...	...	...
88	99	95
62	55	55
59	71	65
68	78	77
77	86	86

Similar to above, we find U by scaling the columns of A as $u_i = \frac{\vec{x}_i}{|\vec{x}_i|}$. So that U is:



math score	reading score	writing score
0.033580	0.032208	0.033560
0.032180	0.040260	0.039909
0.041974	0.042497	0.042177
0.021920	0.025498	0.019955
0.035445	0.034892	0.034014
...	...	...
0.041042	0.044286	0.043084
0.028916	0.024603	0.024943
0.027517	0.031761	0.029478
0.031714	0.034892	0.034921
0.035911	0.038471	0.039002

Finaly, the 3 x 3 correlation matrix $C = U^T U$ is:



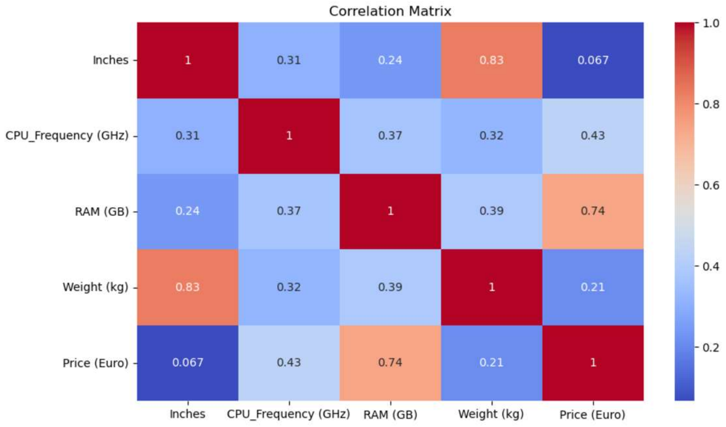
	math score	reading score	writing score
math score	1.000000	0.991430	0.990374
reading score	0.991430	1.000000	0.997891
writing score	0.990374	0.997891	1.000000

The correlation matrix shows that a student's score in mathematics is strongly correlated with his score in reading. Therefore, we can say that if a student is excellent in mathematics, he is most likely excellent in reading.

2.3 Example 2: Prices of Laptops

When a person wants to buy a laptop, it is important to know what affects the price of the laptop in order to know what the appropriate specifications are. We will consider a dataset [1] that contains a variety of laptop specifications and the price of each device in Euros.

We use the same approach as above to obtain the correlation matrix.



	Inches	CPU_Frequency (GHz)	RAM (GB)	Weight (kg)	Price (Euro)
Inches	1	0.31	0.24	0.83	0.067
CPU_Frequency (GHz)	0.31	1	0.37	0.32	0.43
RAM (GB)	0.24	0.37	1	0.39	0.74
Weight (kg)	0.83	0.32	0.39	1	0.21
Price (Euro)	0.067	0.43	0.74	0.21	1

From the correlation matrix we find that the RAM size and CPU frequency are the two factors that affect the price of the laptop the most.

3 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique useful for data visualization and exploration. It is defined as orthogonal projection of the data onto a lower dimensional linear space (principal space) such that the variance of the projected data is maximized. It helps to identify the most important features that explain the most variance in the data, making it easier to visualize and analyse.

3.1 Introduction

Theorem 1. If A is symmetric, then A is orthogonally diagonalizable and has only real eigenvalues.

Proposition 1. If A is any $m \times n$ matrix of real numbers, then the m × m matrix AA^T and the n × n matrix A^TA are both symmetric.

Proposition 2. The eigenvalues of the matrices AA^T and A^TA are nonnegative numbers, and they share the same nonzero eigenvalues.

Definition 1. The *mean* of n vectors in $\mathbb{R}^m$ as a single vector is:

$$\vec{\mu} = \frac{1}{n}\sum_{i=0}^{n} \vec{x}_i$$

Definition 2. Let $A = [\vec{x}_1 \ \vec{x}_2 \ \dots \ \vec{x}_n]$ and let B be the $m \times n$ matrix whose i^{th} column is $\vec{x}_i - \vec{\mu}$, the $m \times m$ *covariance matrix* S is:

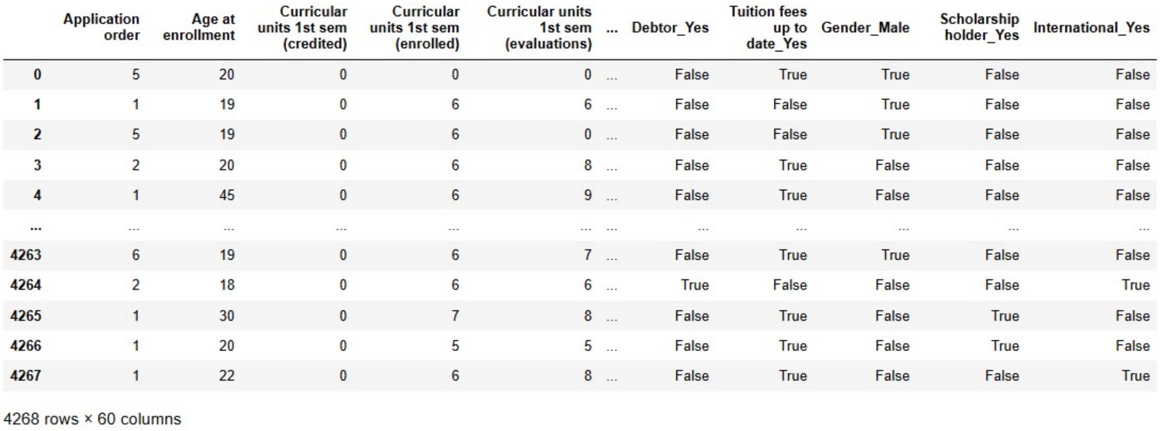
$$S = \frac{1}{n-1}BB^T$$

S is a symmetric matrix by Proposition 1; hence, it can be orthogonally diagonalized by Theorem 1. We let the eigenvalues of S to be such that $\lambda_1 \ \geq \lambda_2 \geq \ldots \geq \ \lambda_m \ \geq \ 0$ with corresponding orthogonal eigenvectors $\vec{u}_1, \ldots, \vec{u}_m$. These eigenvectors are called the principal components of the data set. Notice that the trace of S is equal to $T := \lambda_1 + \lambda_2 + \ldots + \lambda_m$. T is called the total variance. The direction in $\mathbb{R}^m$ given by $\vec{u}_1$ (the first principal direction) accounts for an amount λ_1 of the total variance, T. In fraction It is $\frac{\lambda_1}{T}$. And similarly, the second principal direction $\vec{u}_2$ accounts for $\frac{\lambda_2}{T}$ of the total variance, and so on. Therefore, the vector $\vec{u}_1 \in \mathbb{R}^m$ indicate the most direction of the data set.

If the variance captured by, say, the first 2 eigenvectors is high, we can consider these two eigenvectors in the data analysis instead of using the entire data columns. This is the benefit of PCA in reducing dimensions in data.

3.2 Example: Predict the Student's Status at the End of the Semester

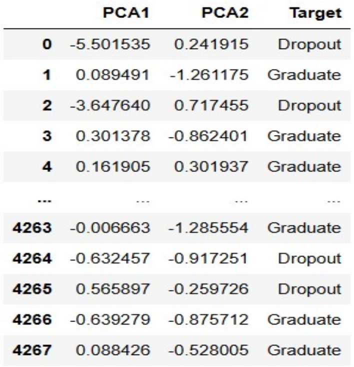
We have a dataset [2] that is about students enrolled in various undergraduate degrees that was collected by a higher education institution. The dataset contains data that was available at the time of enrolment as well as the academic standing of the students at the end of their first and second semesters. At the end of the course's regular term, the problem is presented as a three-category classification dropout, enrolled and graduate. The dataset includes key information on marital status, application details, course, attendance type, nationality, parental qualifications and occupations, scholarship status, age at enrolment, academic performance (curricular units), and economic factors like unemployment, inflation, and GDP. The total number of features (columns) of the data is 60.



	Application order	Age at enrollment	Curricular units 1st sem (credited)	Curricular units 1st sem (enrolled)	Curricular units 1st sem (evaluations)	...	Debtor_Yes	Tuition fees up to date_Yes	Gender_Male	Scholarship holder_Yes	International_Yes
0	5	20	0	0	0	...	False	True	True	False	False
1	1	19	0	6	6	...	False	False	True	False	False
2	5	19	0	6	0	...	False	False	True	False	False
3	2	20	0	6	8	...	False	True	False	False	False
4	1	45	0	6	9	...	False	True	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...
4263	6	19	0	6	7	...	False	True	True	False	False
4264	2	18	0	6	6	...	True	False	False	False	True
4265	1	30	0	7	8	...	False	True	False	True	False
4266	1	20	0	5	5	...	False	True	False	True	False
4267	1	22	0	6	8	...	False	True	False	False	True

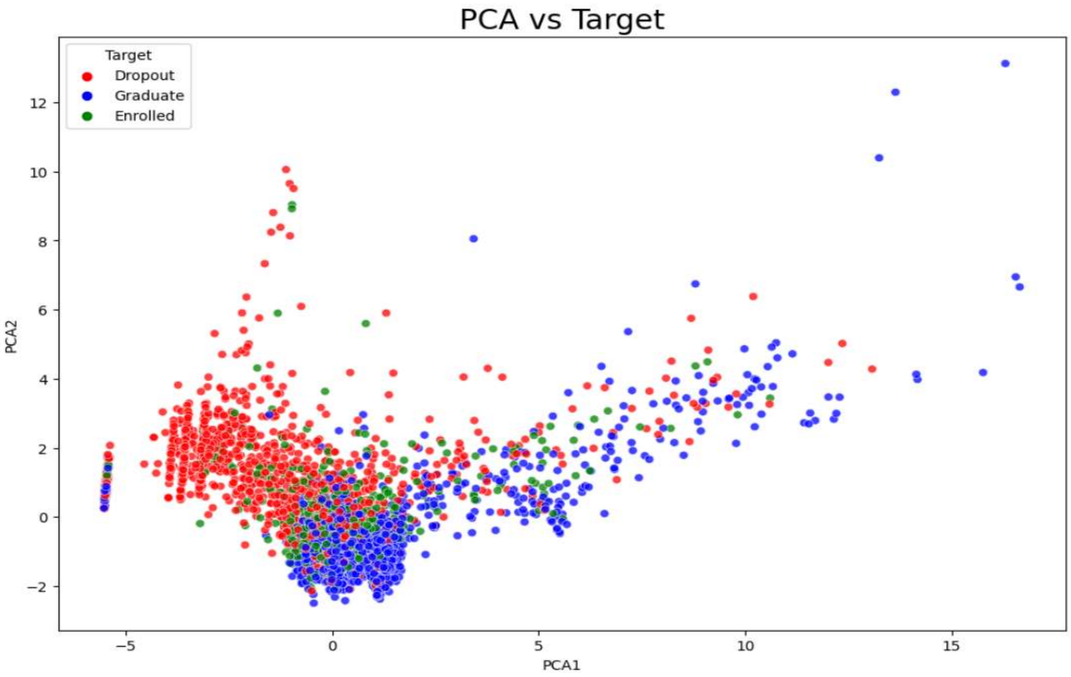
4268 rows × 60 columns

We will perform PCA on the data and we will consider the first two principal components. The first 2 eigenvalues of the covariance matrix are: $\lambda_1 = 76.06528281$ and $\lambda_2 = 58.94119564$. The variance captured by λ_1 is 38.43% and the variance captured by λ_2 is 29.78% so that the total variance captured by λ_1 and λ_2 is 68.21%. As shown below, one can see $\vec{u}_1$, $\vec{u}_2$ and the Target columns, respectively:



	PCA1	PCA2	Target
0	-5.501535	0.241915	Dropout
1	0.089491	-1.261175	Graduate
2	-3.647640	0.717455	Dropout
3	0.301378	-0.862401	Graduate
4	0.161905	0.301937	Graduate
...	...	...	...
4263	-0.006663	-1.285554	Graduate
4264	-0.632457	-0.917251	Dropout
4265	0.565897	-0.259726	Dropout
4266	-0.639279	-0.875712	Graduate
4267	0.088426	-0.528005	Graduate

Now, we construct a scatter plot using the first two PCA of the data, differentiating the points using 'Target' column:



As we can see, we were able to reduce the number of columns from 60 to 2 and still retain 68.21% of the data. As shown in the figure, the first two PCAs ware able to sort the data reasonably well. There were some data that were not considered in the model, but this does not negate the usefulness of the first two PCAs. We can increase the number of PCAs to get better accuracy.

4 Conclusion

In conclusion, we have examined two ways to use linear algebra directly in data science. First, we saw the use of matrices and their operations to extract the correlation matrix, which is important for knowing the relationship between some features. Second, we saw the use of eigenvalues and eigenvectors in PCA to reduce dimensions. Finally, linear algebra is a significant pillar of data science, as it is fundamentally involved in the basics of manipulate with, organizing, and simplifying data.

References

[1] https://www.kaggle.com/code/saifsalama/laptop-price-eda

[2] https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention

J. Jauregui, *Principal component analysis with linear algebra,* August 31, 2012

Leon, Steven J., *Linear Algebra with Applications*, 9th edition, Pearson, 2015.