Linear Algebra in Data Analysis

MATH 225 Project

Mohammed Alghudiyan

Agenda

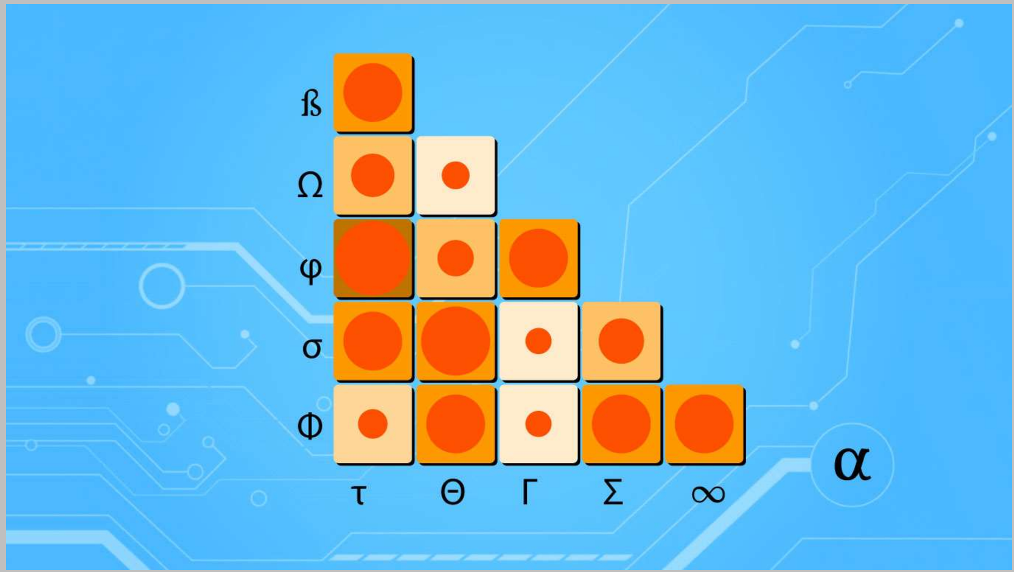
- Introduction
- The Correlation Matrix
  - Prices of Laptops
- Principal Component Analysis (PCA)
  - Predicting Student's Status

Introduction

- Mathematical foundation for data science
- Dealing with large datasets efficiently
- Reducing dimensionality for predictions

The Correlation Matrix

- n x n symmetric matrix
- How strongly variables are connected



The Correlation Matrix

If we have m x n matrix A with numerical entries, define an m x n matrix U where $u_i = \frac{\vec{x}_i}{|\vec{x}_i|}$. Then the n x n correlation matrix is

$$C = U^T U$$

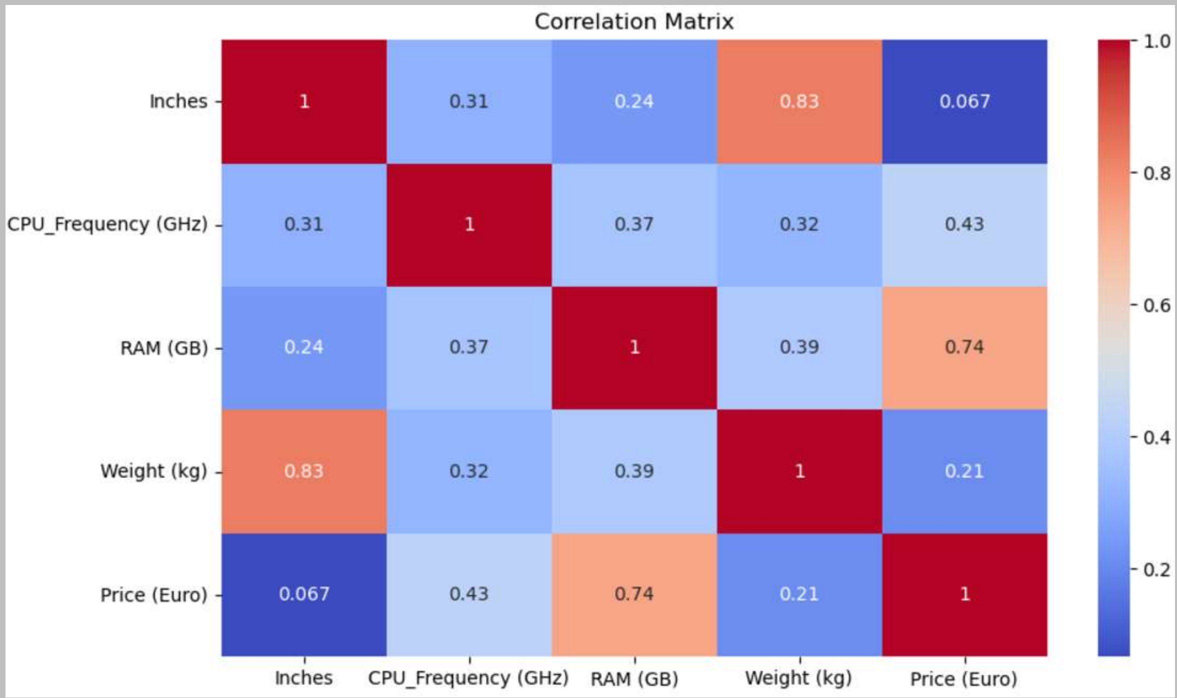
- c_{ij} represents how the i^{th} column of A is related to the j^{th} column

Correlation Values Interpretation

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

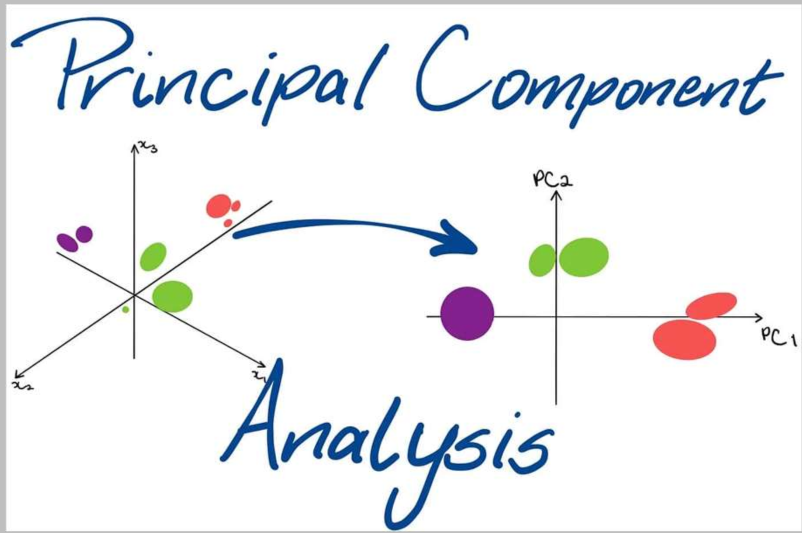
Example: Prices of Laptops

- The dataset contains a variety of laptop specifications and the price of each device.



Principal Component Analysis

- Dimensionality reduction technique
- Identify the most important features
- Visualizing and analyzing



Theorem and Propositions

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 $m \times m$ matrix AA^T and the $n \times n$ matrix A^TA are both symmetric.

Proposition 2: The eigenvalues of the matrices AA^T and A^TA are nonnegative numbers.

Definitions

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 $A = [\vec{x}_1 \ \vec{x}_2 \ \ldots \ \vec{x}_n]$ 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: $S = \frac{1}{n-1} BB^T$.

Principal Component Analysis

- S is a symmetric matrix by Proposition 1
- S can be orthogonally diagonalized by Theorem 1.
- Eigenvalues of S: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$
- Orthogonal eigenvectors $\vec{u}_1, \ldots, \vec{u}_m$ (principal components)
- $T := \lambda_1 + \lambda_2 + \ldots + \lambda_m$ (trace of S, total variance)

Principal Component Analysis

- $\vec{u}_1$ (the first principal direction) accounts for $\frac{\lambda_1}{T}$ of the total variance.

- The vector $\vec{u}_1 \in \mathbb{R}^m$ indicate the most direction of the data set.

- Consider two eigenvectors in the data analysis instead of using all data columns if the variance captured by the first 2 eigenvectors is high.

Example: Predicting Student's Status

- Dataset on students in undergraduate degrees.
- Classifies students into dropout, enrolled, and graduate categories.
- Contains data on demographics, course details, performance, and economic factors.

Example: Predicting Student's Status

- 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

Example: Predicting Student's Status

- Eigenvalues: $\lambda_1 = 76.06528281$ and $\lambda_2 = 58.94119564$
- The variance captured by λ_1 is 38.43%
- The variance captured by λ_2 is 29.78%
- The total variance captured by λ_1 and λ_2 is 68.21%

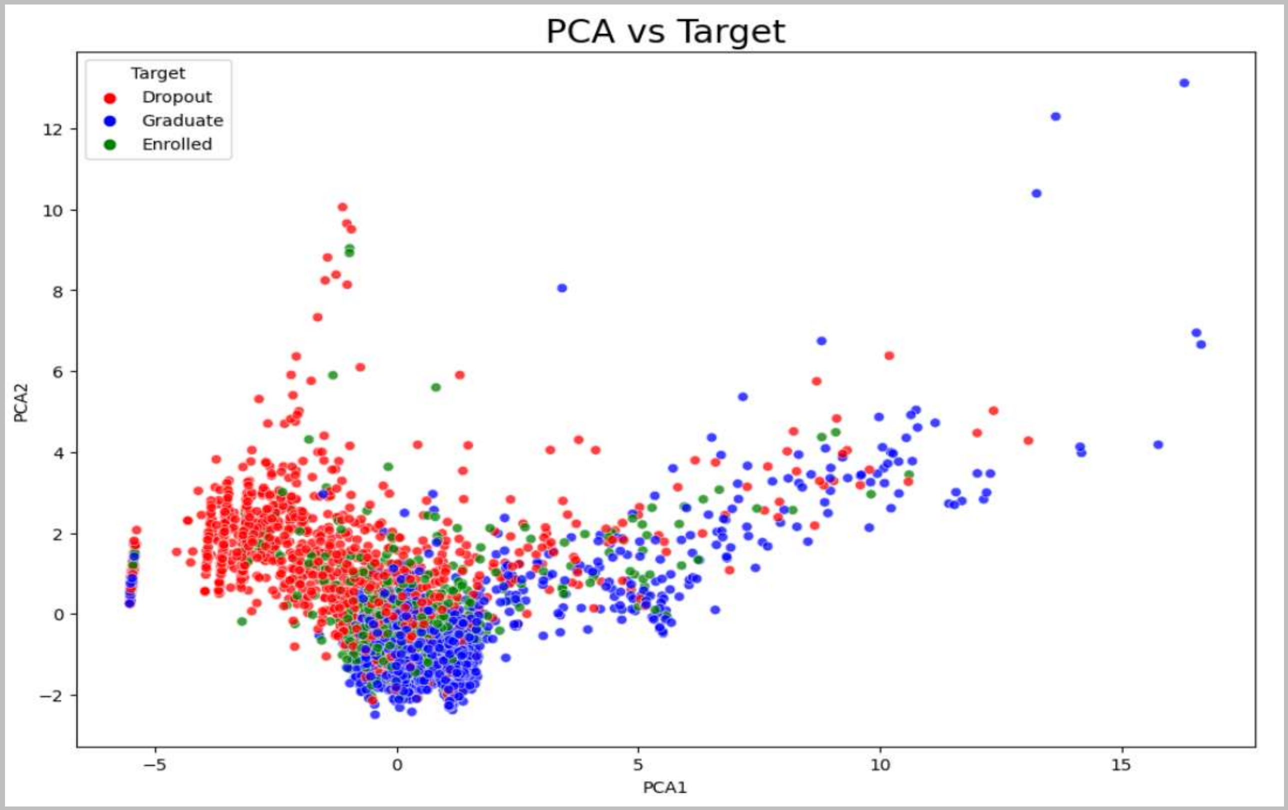
Example: Predicting Student's Status

- $\vec{u}_1$, $\vec{u}_2$ and the Target columns:

	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

Example: Predicting Student's Status

- Scatter plot using the first two PCA:



Thank you