# Chapter 1

# Introduction to Data Science

Data science is a multidisciplinary field that involves using various statistical, mathematical, and computational techniques to extract insights and knowledge from data. Data science encompasses a broad range of techniques, including data visualization, data mining, machine learning, and statistical analysis. The ultimate goal of data science is to use these techniques to make informed decisions and predictions. The data used in data science can come from a variety of sources, including social media, websites, sensors, and other digital devices. This data is often unstructured and requires cleaning and preprocessing before it can be analyzed. Once the data is prepared, it can be analyzed using statistical methods to identify patterns, relationships, and trends. One of the key aspects of data science is machine learning, which involves using algorithms to identify patterns in data and make predictions based on those patterns. Machine learning can be used for a variety of tasks, including image recognition, natural language processing, and fraud detection. Data science also involves data visualization, which is the process of creating visual representations of data to help identify patterns and trends. Visualization tools can be used to create charts, graphs, and other visualizations that make it easier to understand complex data sets.

**Defining Data Science**

Data Science is the systematic study of vast volumes of data to uncover patterns, trends, and insights that can inform decision-making processes across various domains. In essence, it's the art and science of transforming raw data into actionable knowledge. This interdisciplinary field encompasses a range of techniques and methodologies from statistics, computer science, mathematics, and domain expertise to extract meaningful information from data.

**The Role of Data Science**

At its core, Data Science revolves around collecting, processing, analyzing, and interpreting data to extract valuable insights. It involves a combination of statistical analysis, machine learning algorithms, data visualization techniques, and domain knowledge to make sense of complex datasets. By leveraging advanced computational tools and techniques, Data Scientists can uncover hidden patterns and correlations that can drive business strategies, scientific discoveries, and societal advancements.

**Where is Data Science Needed?**

1. **Business and Finance**: In the realm of business and finance, data science is indispensable for market analysis, risk assessment, fraud detection, and customer segmentation. By analyzing consumer behavior and financial trends, organizations can optimize pricing strategies, mitigate risks, and identify lucrative investment opportunities.

2. **Healthcare and Medicine**: Data science is revolutionizing healthcare by enabling predictive modeling, personalized medicine, and disease diagnosis. By analyzing electronic health records, genomic data, and medical imaging, healthcare providers can improve patient outcomes, optimize treatment protocols, and identify potential health risks at an early stage.

3. **Retail and E-Commerce**: Retailers leverage data science to enhance customer experiences, optimize inventory management, and personalize marketing campaigns. By analyzing purchase history, browsing patterns, and demographic information, retailers can tailor their offerings to meet the evolving needs and preferences of consumers.

4. **Transportation and Logistics**: Data science is reshaping the transportation and logistics industries by optimizing route planning, predicting demand, and minimizing operational costs. By analyzing traffic patterns, weather conditions, and delivery routes, companies can streamline supply chain operations and enhance the efficiency of transportation networks.

5. **Education and Research**: In the field of education and research, data science facilitates data-driven decision-making, student performance analysis, and academic research. By analyzing educational data, institutions can identify areas for improvement, personalize learning experiences, and enhance student outcomes.

6. **Energy and Utilities**: Data Science is playing a vital role in the energy and utilities sector by optimizing resource

## 1.1 Linear Algebra for data science

Linear Algebra is a fundamental branch of mathematics that plays a critical role in data science. It provides the tools necessary for representing and manipulating data in a structured and efficient way. Here are some of the key concepts of linear algebra that are essential for data science:

- **Vectors and Matrices:** Vectors are a set of numbers arranged in a specific order, while matrices are an array of numbers arranged in rows and columns. Vectors and matrices are used to represent data in a structured format.

- **Matrix Operations:** Matrix operations include addition, subtraction, multiplication, and inversion. These operations are used extensively in data science for tasks such as regression, classification, and clustering.

- **Linear Transformations:** Linear transformations involve transforming vectors and matrices by multiplying them with a transformation matrix. These transformations are used in data science for tasks such as image processing and feature extraction.

- **Eigenvectors and Eigenvalues:** Eigenvectors and eigenvalues are used to represent the properties of linear transformations. They are used in data science for tasks such as dimensionality reduction and principal component analysis.

- **Singular Value Decomposition (SVD):** SVD is a technique used to decompose a matrix into its constituent parts. It is used in data science for tasks such as image compression and data clustering.

## 1.2 Linear Equations

Linear equations play an important role in data science, as many real-world phenomena can be modeled using linear equations. In particular, linear regression is a commonly used technique in data science that involves fitting a linear equation to a dataset to describe the relationship between the input variables and the output variable.

Linear regression can be used for various tasks, including predicting housing prices, stock prices, and customer behavior. The linear equation used in linear regression can be represented as follows:

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$$

where $y$ is the output variable, $x_1, x_2, ..., x_n$ are the input variables, $b_0$ is the intercept, and $b_1, b_2, ..., b_n$ are the coefficients that determine the relationship between the input variables and the output variable.

To fit a linear regression model to a dataset, the coefficients $b_0, b_1, b_2, ..., b_n$ are estimated using a method called least squares. Least squares involves finding the values of the coefficients that minimize the sum of the squared differences between the predicted values and the actual values in the dataset.

Linear equations can also be used for other tasks in data science, such as optimization and constraint satisfaction. For example, linear programming is a technique used to optimize a linear objective function subject to linear constraints. Linear equations are used to represent the objective function and the constraints in this technique.

## 1.3 Distance

Distance measures in data science are algorithms that quantify the similarity or dissimilarity between two or more objects. These algorithms find common use across various data science applications like clustering, classification and recommendation systems.

Choosing the right distance measure is crucial for optimizing the performance of a data science model. It's essential to thoughtfully evaluate which distance measure best suits a specific problem, as the effectiveness of different distance measures varies depending on the data's characteristics.

Table 1.1: Common Distance Measures

| **Distance Measure** | **Formula** |
|---|---|
| Euclidean Distance | $d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$ |
| Manhattan Distance | $d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} \lvert q_i - p_i \rvert$ |
| Chebyshev Distance | $d(\mathbf{p},\mathbf{q}) = \max_i \lvert q_i - p_i \rvert$ |
| Minkowski Distance | $d(\mathbf{p},\mathbf{q}) = (\sum_{i=1}^{n} \lvert q_i - p_i \rvert^p)^{\frac{1}{p}}$ |
| Cosine Similarity | $\text{similarity}(\mathbf{p},\mathbf{q}) = \frac{\mathbf{p}\cdot\mathbf{q}}{\lVert\mathbf{p}\rVert\lVert\mathbf{q}\rVert}$ |
| Jaccard Distance | $d(\mathbf{p},\mathbf{q}) = 1 - \frac{\lvert\mathbf{p}\cap\mathbf{q}\rvert}{\lvert\mathbf{p}\cup\mathbf{q}\rvert}$ |
| Hamming Distance | $d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n}(p_i \neq q_i)$ |

**Applications of Distance Measures**
Distance measures are used in various applications in data science:

- **Clustering:** Distance measures help in grouping similar data points together.
- **Classification:** They assist in assigning labels or categories to data points based on their similarity.
- **Anomaly Detection:** Distance measures aid in identifying unusual or outlier data points.
- **Recommendation Systems:** They are used to recommend items or content similar to what the user has interacted with.
- **Dimensionality Reduction:** Techniques such as t-SNE and PCA use distance measures to map high-dimensional data into a lower-dimensional space while preserving distances.

When selecting distance measures, several factors should be considered:

- **Nature of Data:** The type of data being analyzed influences the choice of distance measure.
- **Scale and Units:** The scale and units of the data may require standardization or normalization.
- **Dimensionality:** Some distance measures perform better in high-dimensional spaces.
- **Computational Complexity:** The efficiency of the distance measure in large datasets is important.

## 1.4 Hyperplane

A hyperplane is a flat affine subspace of dimension $p-1$ in an $p$-dimensional space. The equation of a hyperplane in $p$-dimensional space can be represented as:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = 0$$

Where:

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of the hyperplane.
- $x_1, x_2, \ldots, x_p$ are the features of the data points.

A hyperplane is a fundamental concept in geometry and linear algebra. It represents a flat subspace of one less dimension than its surrounding space. In data science, hyperplanes are extensively used in support vector machines (SVMs) for binary classification.

**Example:** In a 2-dimensional space, a hyperplane is a line separating data points of different classes. SVMs aim to find the optimal hyperplane that maximizes the margin between the classes.

**Subspace** A subspace is a subset of a vector space that satisfies the properties of a vector space itself. It includes the zero vector and is closed under addition and scalar multiplication. In data science, subspaces are often encountered in dimensionality reduction techniques like principal component analysis (PCA).

**Example:** PCA identifies principal components (subspaces) that capture the most variance in the data, forming a lower-dimensional subspace while retaining essential information.

**Halfspace** A halfspace is a region of space bounded by a hyperplane. The equation of a halfspace is represented as:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \geq 0$$

or

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \leq 0$$

depending on which side of the hyperplane defines the halfspace.

A halfspace is a region of space divided by a hyperplane into two parts: the positive and negative halfspaces. In data science, halfspaces are integral to linear classifiers such as perceptrons and logistic regression.

**Example:** In a binary classification problem, a halfspace represents a linear boundary dividing the feature space into regions corresponding to different classes.

## 1.5 Eigen Values

Eigenvalues and eigenvectors are important concepts in linear algebra that are widely used in data science, particularly in the context of dimensionality reduction techniques such as principal component analysis (PCA), linear discriminant analysis (LDA), and singular value decomposition (SVD). Eigenvalues and eigenvectors are closely related to linear transformations. Given a linear transformation, an eigenvector is a vector that, when transformed by the transformation, is scaled by a scalar value known as the eigenvalue. In other words, the eigenvector retains its direction but is scaled by a certain factor.

## 1.6 Eigen Vectors

In PCA, the goal is to reduce the dimensionality of a high-dimensional dataset while retaining as much information as possible. The technique involves finding the principal components of the dataset, which are the eigenvectors of the covariance matrix of the data. The eigenvalues of the covariance matrix indicate the variance of the data along each principal component, and the principal components with the highest eigenvalues capture the most variance in the data. In LDA, the goal is to find a linear transformation that maximizes the separation between different classes of data points. The technique involves finding the eigenvectors of the scatter matrix, which is a measure of the spread of the data within and between classes. The eigenvectors with the highest eigenvalues indicate the directions that maximize the separation between the classes. In SVD, the goal is to decompose a matrix into its singular values, which are equivalent to the eigenvalues of the matrix's covariance matrix, and its singular vectors, which are equivalent to the eigenvectors of the matrix's covariance matrix. SVD is commonly used in data science for various tasks, such as image processing, recommendation systems, and natural language processing.

# PCA: Step-by-Step Example

**Step 1: Data Normalization**

First, we center the data by subtracting the mean of each feature from the respective feature values.

Compute the mean of Feature 1:

$$\bar{x}_1 = \frac{2+3+4}{3} = 3$$

Compute the mean of Feature 2:

$$\bar{x}_2 = \frac{3+4+5}{3} = 4$$

Subtract the means from the original data:

| Feature 1 | Feature 2 |
|---|---|
| $-1$ | $-1$ |
| 0 | 0 |
| 1 | 1 |

**Step 2: Compute the Covariance Matrix**

Next, we calculate the covariance matrix of the centered data.

The covariance matrix $\Sigma$ is given by:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Where $n$ is the number of data points, $\mathbf{x}_i$ is the $i$-th data point, and $\bar{\mathbf{x}}$ is the mean vector.

In our case, $n = 3$.

The covariance matrix for our centered data:

$$\Sigma = \frac{1}{3}\left(\begin{bmatrix}-1\\0\\1\end{bmatrix}\begin{bmatrix}-1 & 0 & 1\end{bmatrix}\right) = \frac{1}{3}\begin{bmatrix}1 & 0 & -1\\0 & 0 & 0\\-1 & 0 & 1\end{bmatrix}$$

**Step 3: Compute Eigenvalues and Eigenvectors**

Next, we find the eigenvalues ($\lambda$) and eigenvectors ($\mathbf{v}$) of the covariance matrix $\Sigma$.

Using the characteristic equation $\Sigma\mathbf{v} = \lambda\mathbf{v}$, we solve for $\lambda$ and $\mathbf{v}$.

**Step 4: Sort Eigenvalues**

Sort the eigenvalues in decreasing order and their corresponding eigenvectors.

**Step 5: Choose Principal Components**

Choose the top $k$ eigenvectors based on the explained variance you want to retain.

**Step 6: Project Data onto Principal Components**

Project the centered data onto the chosen principal components.

**Compute Eigenvalues and Eigenvectors**

Given the covariance matrix:

$$\Sigma = \frac{1}{3}\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

**Step 1: Eigenvalue Equation**

The eigenvalue equation is $\Sigma\mathbf{v} = \lambda\mathbf{v}$, where $\lambda$ is the eigenvalue and $\mathbf{v}$ is the corresponding eigenvector.

**Step 2: Solve the Characteristic Equation**

We need to solve the characteristic equation $|\Sigma - \lambda I| = 0$, where $I$ is the identity matrix.

Subtract $\lambda$ from the diagonal of $\Sigma$:

$$\begin{vmatrix} 1-\lambda & 0 & -1 \\ 0 & -\lambda & 0 \\ -1 & 0 & 1-\lambda \end{vmatrix} = 0$$

**Step 3: Solve the Equation**

Solve the determinant equation to find the eigenvalues $\lambda$.

$$(1-\lambda)((-\lambda)(1-\lambda)) - 0 - (-1)(0) = 0$$

$$(1-\lambda)(\lambda^2 - \lambda) = 0$$

$$\lambda^3 - 2\lambda^2 + \lambda = 0$$

This gives us a cubic equation. We need to find the roots of this equation.

**Step 4: Find Eigenvalues**

We can factor out $\lambda$ and solve the quadratic equation:

$$\lambda(\lambda^2 - 2\lambda + 1) = 0$$

$$\lambda(\lambda - 1)^2 = 0$$

This equation gives us eigenvalues $\lambda_1 = 0$ and $\lambda_2 = 1$.

**Step 5: Compute Eigenvectors**

For $\lambda_1 = 0$:

$$\Sigma - 0I = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

Perform Gaussian elimination to find the null space. We find the eigenvector corresponding to $\lambda_1 = 0$ is:

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

For $\lambda_2 = 1$:

$$\Sigma - 1I = \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

Perform Gaussian elimination to find the null space. We find the eigenvector corresponding to $\lambda_2 = 1$ is:

$$\mathbf{v}_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

Now, we have computed the eigenvalues and eigenvectors for the covariance matrix $\Sigma$.

**Step 6: Sort Eigenvalues**

Sort the eigenvalues in descending order.

For our example, the eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 0$. Thus, after sorting, the eigenvalues are:

$$\lambda_1 = 1$$

$$\lambda_2 = 0$$

**Step 7: Choose Principal Components**

After computing the eigenvalues and eigenvectors, we choose the principal components based on the explained variance or the proportion of total variance they capture. Typically, we select the top $k$ eigenvectors corresponding to the largest eigenvalues, as they explain the most variance in the data.

# Linear Discriminant Analysis (LDA) Example

**Step 1: Data Preparation**

Given:

- Class 1:
  - Feature 1: $[1, 2, 2]$
  - Feature 2: $[1, 3, 2]$
- Class 2:
  - Feature 1: $[4, 5, 6]$
  - Feature 2: $[5, 7, 6]$

**Step 2: Compute the Mean Vectors for Each Class**
For Class 1:

$$\bar{x}_1 = \frac{1+2+2}{3} = \frac{5}{3},$$
$$\bar{x}_2 = \frac{1+3+2}{3} = 2$$

For Class 2:

$$\bar{y}_1 = \frac{4+5+6}{3} = 5,$$
$$\bar{y}_2 = \frac{5+7+6}{3} = 6$$

**Step 3: Compute the Scatter Within Classes ($S_W$)**

For Class 1:

$$S_{w1} = \begin{bmatrix} \frac{4}{9} & -\frac{1}{3} \\ -\frac{1}{3} & 1 \end{bmatrix}$$

For Class 2:

$$S_{w2} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Thus, $S_w = \begin{bmatrix} \frac{13}{9} & -\frac{4}{3} \\ -\frac{4}{3} & 2 \end{bmatrix}$

**Step 4: Compute the Scatter Between Classes ($S_b$)**

$$S_b = \begin{bmatrix} \frac{100}{9} & \frac{40}{3} \\ \frac{40}{3} & 16 \end{bmatrix}$$

**Compute the Eigenvalues and Eigenvectors**
Given $S_w^{-1} S_b$, we compute its eigenvalues and eigenvectors.
Recall:

- $S_w$ is the within-class scatter matrix.
- $S_b$ is the between-class scatter matrix.

We need to solve the generalized eigenvalue problem:

$$S_w^{-1} S_b v = \lambda v$$

Where:

- $v$ is the eigenvector.
- $\lambda$ is the eigenvalue.

For this problem, we already calculated $S_w$ and $S_b$.
From previous calculations, we have:

$$S_w = \begin{bmatrix} \frac{13}{9} & -\frac{4}{3} \\ -\frac{4}{3} & 2 \end{bmatrix}$$

$$S_b = \begin{bmatrix} \frac{100}{9} & \frac{40}{3} \\ \frac{40}{3} & 16 \end{bmatrix}$$

We compute $S_w^{-1}$ first. Inverse of $S_w$ is:

$$S_w^{-1} = \frac{1}{|S_w|} \cdot \mathrm{adj}(S_w)$$

Where $|S_w|$ is the determinant of $S_w$ and $\mathrm{adj}(S_w)$ is the adjugate (or adjoint) of $S_w$.

For $2 \times 2$ matrix $S_w$:

$$|S_w| = \left(\frac{13}{9} \times 2\right) - \left(-\frac{4}{3} \times -\frac{4}{3}\right) = \frac{26}{9} - \frac{16}{9} = \frac{10}{9}$$

The adjugate of $S_w$ is obtained by swapping the elements on the main diagonal and changing the sign of the off-diagonal elements:

$$\mathrm{adj}(S_w) = \begin{bmatrix} 2 & \frac{4}{3} \\ \frac{4}{3} & \frac{13}{9} \end{bmatrix}$$

Thus,

$$S_w^{-1} = \frac{9}{10} \begin{bmatrix} 2 & \frac{4}{3} \\ \frac{4}{3} & \frac{13}{9} \end{bmatrix} = \begin{bmatrix} \frac{9}{5} & \frac{12}{10} \\ \frac{12}{10} & \frac{117}{90} \end{bmatrix}$$

Now, we compute $S_w^{-1} S_b$:

$$S_w^{-1} S_b = \begin{bmatrix} \frac{9}{5} & \frac{12}{10} \\ \frac{12}{10} & \frac{117}{90} \end{bmatrix} \cdot \begin{bmatrix} \frac{100}{9} & \frac{40}{3} \\ \frac{40}{3} & 16 \end{bmatrix}$$

Please perform the multiplication and subsequent calculations in order to obtain the eigenvalues and eigenvectors.

**Choose Discriminants** Choose the discriminants based on the eigenvalues. If there are $k$ classes, then the number of discriminants is $k - 1$.

**Project the Data** Project the data onto the discriminants to get the transformed dataset.

## 1.7 Exercises

### 1.7.1 Short and Long Answer Questions

1. Define the concept of data science and its significance in modern industries.

2. List the fundamental principles of linear algebra relevant to data science.
3. Explain the concept of eigenvalues and eigenvectors in the context of linear algebra for data science.
4. Discuss the importance of hyperplanes and half spaces in machine learning algorithms.
5. Given a set of linear equations, solve them using appropriate linear algebra techniques.
6. Apply the concepts of distance metrics to measure similarity between data points.
7. Analyze a dataset and determine the eigenvalues and eigenvectors using appropriate computational methods.
8. Compare and contrast different methods for solving linear equations in data science applications.
9. Evaluate the effectiveness of various distance metrics in capturing the similarity between high-dimensional data points.
10. Assess the significance of linear algebra concepts such as hyperplanes and half spaces in machine learning algorithms.

### 1.7.2 Fill in the Blanks

1. Linear algebra is essential for data science, especially in tasks like ______.
2. Eigenvalues are often used to understand how much ______ a transformation applies to its corresponding eigenvector.
3. In linear algebra, a ______ is a set of vectors that are linearly independent and span the vector space.
4. The m______ of a matrix determines whether it can be inverted or not.
5. Half spaces are regions of space that lie on one side of a ______.
6. In data science, linear equations are used to model relationships between ______.
7. The dot product of two vectors yields a ______, not another vector.
8. Distance between two points in Euclidean space can be computed using the ______.
9. Singular Value Decomposition (SVD) decomposes a matrix into ______.
10. In linear algebra, eigenvectors represent directions in space that are only scaled by a ______.

### 1.7.3 Multiple Choice Questions

1. What is data science?
   A) The study of computers
   B) The study of collecting and analyzing data to gain insights
   C) The study of rocket science
   D) The study of plants

2. Which branch of mathematics is essential for data science?
   A) Calculus
   B) Trigonometry
   C) Linear Algebra
   D) Geometry

3. What are linear equations used for in data science?
   A) To solve complex algorithms
   B) To model relationships between variables
   C) To draw graphs
   D) To create pie charts

4. What is a hyperplane in linear algebra?
   A) A plane with no dimensions
   B) A plane that divides a space into two half-spaces
   C) A plane in 3-dimensional space
   D) A plane with curved edges

5. What term describes the region of space on one side of a hyperplane?
   A) Quarter space
   B) Full space
   C) Half space
   D) Hyper space

6. What are eigenvalues used for in linear algebra?
   A) They represent the magnitude of eigenvectors
   B) They define the direction of eigenvectors
   C) They measure the size of matrices
   D) They solve linear equations

7. What are eigenvectors?
   A) Vectors that don't change direction when a linear transformation is applied
   B) Vectors with varying magnitudes
   C) Vectors with constant magnitudes
   D) Vectors that change direction when a linear transformation is applied

8. How is distance computed between two points in Euclidean space?
   A) Using the Manhattan distance
   B) Using the Pythagorean theorem
   C) Using the cosine similarity
   D) Using the dot product

9. Which matrix operation involves finding the transpose of a matrix?
   A) Addition
   B) Multiplication
   C) Transposition
   D) Inversion

10. In linear algebra, what does the rank of a matrix represent?
    A) The number of rows in the matrix
    B) The number of columns in the matrix
    C) The number of non-zero rows after row reduction
    D) The number of non-zero elements in the matrix

11. What is the determinant of a matrix used for?
    A) Finding the sum of elements in a matrix
    B) Determining if a matrix is invertible
    C) Finding the product of elements in a matrix
    D) Determining the shape of a matrix

12. Which of the following is not a type of matrix?
    A) Identity matrix
    B) Square matrix
    C) Rectangular matrix
    D) Scalar matrix

13. What is the main objective of Principal Component Analysis (PCA)?
    A) To reduce the dimensionality of data
    B) To increase the complexity of data
    C) To expand the dataset
    D) To perform clustering analysis

14. What is the role of singular value decomposition (SVD) in data analysis?
    A) It helps in clustering data points
    B) It assists in reducing noise in data
    C) It decomposes a matrix into singular vectors and singular values
    D) It performs polynomial regression

15. Which of the following is not a step in the data science workflow?
    A) Data acquisition
    B) smallsing
    C) Data visualization
    D) Data fabrication

16. What is the purpose of regularization in machine learning?
    A) To make the model overfit the training data
    B) To penalize large coefficients in the model
    C) To increase the bias of the model
    D) To decrease the variance of the model

17. Which statistical concept measures the spread of data points around the mean?
    A) Median
    B) Variance
    C) Mean absolute deviation
    D) Mode

18. What is the formula for the dot product of two vectors A and B?
    A) $|A| \times |B| \cos(\theta)$
    B) $|A| + |B|$
    C) $|A| \times |B|$
    D) $|A| - |B| \cos(\theta)$

19. Which algorithm is used to find the optimal solution for linear regression?
    A) Gradient Descent
    B) Breadth-First Search
    C) Depth-First Search
    D) K-nearest Neighbors

20. What is the purpose of the bias term in linear regression?
    A) To offset the influence of independent variables
    B) To add noise to the model
    C) To improve model generalization
    D) To create a curved relationship between variables