Data Science For Engineers
NPTEL PMRF Live Sessions

Teaching Assistant: Parvathy Neelakandan
PhD Student

05-08-2025

1 Representation of data

Data refers to a collection of observations, measurements or facts in different forms. For example, consider a table containing details of student scores. This could include information such as names of the students, the subjects they wrote exam for, their scores. This forms a dataset. Here, information about each student refers to an observation or sample. Names of students, scores are the features or characteristics of those samples/observations.

There are different ways in which such data can be stored

- Can you recall some of these that we discussed in the last class?
- When is each of it used?

1.1 What is a Matrix in Data Science Terms?

In data science terms, a matrix is a structured way to represent or store data similar to a table, where rows represent the samples and columns represent the features of each sample.

Examples

- Medical dataset: Rows may represent patients and columns might contain information such as blood pressure, sugar level, heart rate, age, etc.
- Images: A grayscale image can be represented as a two-dimensional matrix, where each element denotes the value of a pixel.

Matrix in R

How do you create a matrix in R?

2 Identification of independent attributes

When we organize data into a matrix, each column represents a feature (attribute) and each row represents an observation. Once we have created this matrix to store our data, this gives us access to tools that help us understand the data more deeply. One question could be: **"Do all the features in our data carry unique information or are all the features needed?"**.

Example

Consider a dataset about houses with columns for:

- Area in square feet
- Area in square meters
- Price in rupees
- Price in euros

The areas in different units and prices in different units carry similar information. This means that we do not have 4 independent pieces of information per house. We have only two unique information.

2.1 What is the benefit of knowing this?

It tells us how many variables are truly independent or carry unique information. Knowing this would help us identify the redundant features, remove the correlated features and optimize storage and computation.

2.2 Matrix Rank

How do we find relationships between features from the data?

Matrix Rank

Definition 2.1. The rank of a matrix A is the maximum number of linearly independent rows (or columns) in the matrix. It represents the dimension of the vector space spanned by the rows (or columns) of the matrix.

2.3 Practice questions: Rank

MCQ 1: What is the rank of the following matrix?

$$\begin{pmatrix} 2 & 4 & 6 \\ 1 & 2 & 3 \\ 3 & 6 & 9 \end{pmatrix}$$

1. 1
2. 2
3. 3
4. 0

MCQ 2: If a 4×6 matrix has rank 3, what is the maximum number of linearly independent columns it can have?

1. 3
2. 4
3. 6
4. Cannot be determined

MCQ 3: For an $m \times n$ matrix, the maximum possible rank is:

1. m
2. n
3. $m + n$
4. $\min(m, n)$

For an $m \times n$ matrix A, the maximum possible rank is $\min(m, n)$.

If the rank is less than the total number of variables (columns), then there must be some relationships among the variables. That is, some variables can be written as combinations of others. **How do we identify those relationships?**

3 Null Space

For matrix A, are there set of vectors $\mathbf{x}$ that satisfies the equation $A\mathbf{x} = \mathbf{0}$? This would mean that the set of vectors $\mathbf{x}$ forms the null space of A. If such a set $\mathbf{x}$ exists, then there are linear relationships among the variables (columns) of the matrix. The size of null space provides the number of such linear relationships among the variables and the set $\mathbf{x}$ gives the linear relationship between variables.

Null Space

Definition 3.1. For matrix A, the null space (or kernel) consists of all vectors $\mathbf{x}$ that satisfy the equation $A\mathbf{x} = \mathbf{0}$.

$$\text{Null}(A) = \{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$$

3.1 The Rank-Nullity Theorem

The number of columns of a matrix A is the sum of the rank of A and the nullity of A.

For an $m \times n$ matrix A:

$$\text{rank}(A) + \text{nullity}(A) = n$$

where $\text{nullity}(A)$ is the dimension of the null space.

3.2 Practice questions: Null space

MCQ 4: For a 3×5 matrix with rank 2, what is the dimension of its null space?

1. 2
2. 3
3. 5
4. Cannot be determined

MCQ 5: According to the rank-nullity theorem, for matrix A with n columns:

1. $\text{rank}(A) + \text{nullity}(A) = n$
2. $\text{rank}(A) - \text{nullity}(A) = n$
3. $\text{rank}(A) \times \text{nullity}(A) = n$
4. $\text{rank}(A) = \text{nullity}(A)$

MCQ 6: The null space of matrix $\begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}$ is:

1. $\{\mathbf{0}\}$
2. $\text{span}\{[2, -1]\}$
3. $\text{span}\{[1, -2]\}$
4. All of $\mathbb{R}^2$

MCQ 7: If the null space of a square matrix contains only the zero vector, then the matrix is:

1. Singular
2. Non-invertible
3. Invertible
4. Not full rank

4 Systems of Linear Equations

4.1 The Three Cases

When solving $A\mathbf{x} = \mathbf{b}$, the relationship between the number of equations (m) and variables (n) determines the nature of solutions:

Types of Linear Systems

1. **Case 1:** $m = n$
   - When number of equations (m) = number of variables (n)
   - Full rank (rank = n): Unique solution exists
   - Rank deficient: Either no solution or infinitely many solutions
2. **Case 2:** $m > n$ **(Overdetermined System)**
   - When number of equations (m) > number of variables (n)
   - All equations may not be satisfied with given variables -> no-solution case
   - Use least squares for best approximate solution: minimize $\|A\mathbf{x} - \mathbf{b}\|$
3. **Case 3:** $m < n$ **(Underdetermined System)**
   - When number of equations (m) < number of variables (n)
   - Results in infinitely possible solutions
   - How do we choose single best solution from infinite set?

Consistent System

Definition 4.1. A system of equations is said to be consistent if there is at least one set of values that satisfies the set of equations.

Moore-Penrose Pseudoinverse

What is Moore-Penrose Pseudoinverse? How do we calculate it?

5 Vectors

A vector represents a point in n-dimensional space:

- $n = 2$: Point in plane $\mathbb{R}^2$
- $n = 3$: Point in 3D space $\mathbb{R}^3$
- $n > 3$: Point in higher-dimensional space $\mathbb{R}^n$

5.1 Vector Magnitude Calculation

How do we calculate the magnitude of a vector? One way to calculate the magnitude is to take the Euclidean norm.

Vector Magnitude

Definition 5.1. For vector $v = [v_1, v_2, \ldots, v_n]$, the Euclidean norm is:

$$\|\mathbf{v}\|_2 = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

Unit Vector

Definition 5.2. A unit vector has magnitude 1

$$\hat{\mathbf{u}} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$$

5.2 Practice questions: Vectors

MCQ 8: The magnitude of vector $[3, -4, 12]$ is:

1. 11
2. 13
3. 15
4. 19

MCQ 9: A unit vector has magnitude:

1. 0
2. 1
3. Any positive value
4. Depends on the dimension

MCQ 10: To find a unit vector in the direction of $[6, 8]$, we:

1. Multiply by 10
2. Divide by 10
3. Multiply by $\frac{1}{10}$
4. Divide by 14

5.3 Orthogonal and Orthonormal Vectors

Two vectors are orthogonal to each other if their dot product is zero.

Dot Product

Definition 5.3. For vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$:

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$$

What do dot product between two vectors tell us?

Orthonormal Vectors

Definition 5.4. Orthonormal vectors are orthogonal vectors with unit magnitude.

5.4 Practice questions: Orthogonality

MCQ 11: Two vectors are orthogonal if their:

1. Magnitudes are equal
2. Dot product is zero
3. Cross product is zero
4. Sum is zero

MCQ 12: The dot product of $[1, 2, -1]$ and $[2, -1, 2]$ is:

1. -2
2. 0
3. 2
4. 4

MCQ 13: Orthonormal vectors are:

1. Orthogonal with unit magnitude
2. Parallel with unit magnitude
3. Only orthogonal
4. Only unit vectors

MCQ 14: Vectors $[2, 3]$ and $[6, -4]$ are:

1. Orthogonal
2. Parallel
3. Neither orthogonal nor parallel
4. Linearly dependent

5.5 Basis and Span

Span

Definition 5.5. The span of a set of vectors is the set of all possible linear combinations of those vectors.

Basis

Definition 5.6. A basis is a set of vectors that are linearly independent and span the entire vector space.

5.5.1 Interpreting span and basis

Span represents all points that are reachable by linear combinations of the set of vectors. This could be a line, plane, or any higher-dimensional subspace. Basis gives the minimum number of vectors needed to represent any vector in the space.

5.6 Practice questions: Basis and Span

MCQ 15: A basis is a set of vectors that are:

1. Independent and span the space
2. Orthogonal and span the space
3. Unit vectors that span the space
4. Any vectors that span the space

MCQ 16: The span of vectors $\{[1, 0, 1], [0, 1, 1]\}$ in $\mathbb{R}^3$ is:

1. All of $\mathbb{R}^3$
2. A plane through the origin
3. A line through the origin
4. Just the zero vector

MCQ 17: The vectors $\{[1, 2], [2, 4]\}$ form a basis for:

1. $\mathbb{R}^2$
2. A line through the origin
3. No vector space
4. A plane in $\mathbb{R}^3$

MCQ 18: Minimum number of vectors needed to form a basis for $\mathbb{R}^2$?

1. 1
2. 2
3. 3
4. Any number

5.7 Projection

Orthogonal Projection

Definition 5.7. The orthogonal projection of vector $\mathbf{u}$ onto vector $\mathbf{v}$ is:

$$\text{proj}_{\mathbf{v}}(\mathbf{u}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{v}\|^2}\mathbf{v}$$

5.8 Practice questions: Projection

MCQ 19: The projection of $[4, 3]$ onto $[1, 0]$ is:

1. $[4, 0]$
2. $[1, 0]$
3. $[4, 3]$
4. $[0, 3]$

6 Hyperplanes

Geometric entity whose dimension is one less than that of the ambient space. For example: the hyperplanes for a 3D space are 2D planes and hyperplane for a 2D space are 1D lines and so on.
Hyperplane is represented by the equation $\mathbf{x}^T\mathbf{n} + b = 0$

- $\mathbf{n}$: normal vector (perpendicular to hyperplane)
- b: bias term (shifts hyperplane from origin)

6.1 Halfspaces

Halfspaces on either sides of a line can have different characteristics that can help with classification problems. For eg:, the equation $\mathbf{x}^T\mathbf{n} + b = 0$ divides space into two halfspaces:

- $\mathbf{x}^T\mathbf{n} + b > 0$: One side
- $\mathbf{x}^T\mathbf{n} + b < 0$: Other side

This is useful for classification tasks.

6.2 Practice questions: Hyperplanes

MCQ 20: In 4-dimensional space, a hyperplane has dimension:

1. 4
2. 3
3. 2
4. 1

MCQ 21: In 2D space, hyperplanes are:

1. Points
2. Lines
3. Planes
4. Cubes

7 Eigenvalues and Eigenvectors

Consider the equation $A\mathbf{x} = \mathbf{b}$. What does this mean geometrically? When we multiply a matrix A by a vector $\mathbf{x}$, we usually get a vector pointing in a completely different direction. However, special vectors called eigenvectors only get scaled (stretched or shrunk) without changing direction.

Eigenvalue and Eigenvector

Definition 7.1. For a square matrix A, if there exists a non-zero vector $\mathbf{x}$ and scalar λ such that:

$$A\mathbf{x} = \lambda\mathbf{x}$$

then λ is called an eigenvalue and $\mathbf{x}$ is called an eigenvector.
The characteristic equation becomes:

$$|A - \lambda I| = 0$$

Why Eigenvalues and vectors matter?

7.1 Computing Eigenvalues

To find eigenvalues of matrix A, solve the characteristic equation:

$$\det(A - \lambda I) = 0$$

Properties of Eigenvalues for different matrix types

- **Real matrices:** Eigenvalues can be real or complex numbers. If eigenvalues are complex, then the corresponding eigenvectors are also complex.
- **Symmetric matrices:** Eigenvalues and eigenvectors are always real and there will be n linearly independent eigenvectors.
- **Matrices of the form A^TA or AA^T:** Have real and non-negative eigenvalues with linearly independent eigenvectors.

How are the null space and column space related to eigenvalues and eigenvectors?

- When eigenvalues are zero ($A\mathbf{x} = \mathbf{0}$), the eigenvectors corresponding to zero eigenvalues are in the null space of the matrix.
- If A is symmetric and if there are r eigenvalues which have a value of 0, then the dimensionality of the null space is r. If there are n real eigenvalues, then the remaining $(n - r)$ are non-zero independent vectors and these form the basis for column space.

7.2 Practice questions

MCQ 22: To find eigenvalues of matrix A, we solve:

1. $\det(A) = 0$
2. $\det(A - \lambda I) = 0$
3. $A - \lambda I = 0$
4. $A\mathbf{x} = \lambda$

MCQ 23: For real matrices, eigenvalues can be:

1. Only real numbers
2. Only complex numbers

3. Both real and complex numbers
4. Neither real nor complex

MCQ 24: The determinant of a 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is:

1. $ad + bc$
2. $ad - bc$
3. $ac - bd$
4. $ab - cd$

MCQ 25: The determinant of matrix $\begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix}$ is:

1. 2
2. 5
3. 6
4. 10

MCQ 26: When eigenvalues are zero ($A\mathbf{x} = \mathbf{0}$), the corresponding eigenvectors are in the:

1. Column space
2. Row space
3. Null space
4. Entire space