INTRODUCTION TO DATA I

- Data is a raw facts, characteristics, number, or quantity that can be measured or counted. Data is kept for variables.
- A variable may also be called a data item. Age, business income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables.
- It is called a variable because the value may vary between data units in a population, and may change in value over time.

Examples:

1. Income = 20,000
Here, income is a variable and 20,000 is a data.

2. Age = 25
Here, age is variable and 25 is a data.

3. Country = India
Here, country is a variable and India is a data.

4. Sales = 45.67
Here, sales is a variable and 45.67 is a data.

From Machine Learning point of view, we care only for variables not constants

INTRODUCTION TO DATA II

Data values of a variable:

There are different ways a variable can take the data based on how is studied, measured, and presented. A variable can take two type of data or data values.

1. **Quantitative data:** Here, the values taken by the variable are numerical in nature.
   - **Continuous:** It represents data value that is numerical and, has an infinite number of possible values.
     - **eg:** Age, Income, Marks, Temperature, Speed, Distance.
   - **Discrete:** It represents data value that is numerical and, has a finite number of possible values.
     - **eg:** Red flowers in the garden, Balls in the bag
2. **Qualitative data:** Here, the values taken by the variable are non-numerical in nature.
   - **Categorical:** It represents the data values that describe a 'quality' or 'characteristic'of a data unit, like 'what type' or 'which category'.
   - Categorical variables fall into two categories:
     - **Mutually exclusive** (in one category or in another)
     - **Exhaustive** (include all possible options).
   - **Categorical** variables are qualitative variables.
     - **eg:** Gender, Eye color, Marital Status, Breed of Dog

INTRODUCTION TO DATA III

Data set

- Collection of data represented in an organised way is a data set.
- Data set is generally represented in rows and columns.
  - Each row indicates an observation or an object against which data is recorded.
  - Column represents characteristics of objects.

There exist several terminologies for rows and columns of a data set such as:

Rows: observations, objects, data points, tuple

Columns: features, attributes, dimensions

Example illustration

Age	Patient Name	Weight
67	Ajay	67
62	Ravi	78
58	Sunita	56
51	Nidhi	66
59	Rita	56

Figure 1: Hospital dataset: 5 rows and 3 features

Name	Roll no.	Age	Course
Ravi	1234	23	BTech
Raj	1235	20	BSC
Heena	1236	21	BCA
Ajay	1237	22	MTech
Sunita	1238	23	BTech

Figure 2: College dataset: 5 rows and 4 features

INTRODUCTION TO DATA IV

Data Exploration

Data Exploration is a preliminary step to understand the key characteristics and behaviour of data. Data Exploration can aid in selecting the suitable models for knowledge discovery. The two most common data exploration methods are:

1. Summary statistics :

Using basic statistics, we can infer the general characteristics of data values taken by the variable. The most common statistics methods to summarise a variable are:

(a) Frequencies:

It computes the frequencies by which each data value occurs for a particular variable. It is used to summarise a categorical variable. Frequency of a categorical variable is represented in a frequency table.

Consider Table 2 that captures degree records of 185 students. For example, there are 100 students with BTech degree, 50 with MTech, 30 are MCA and 5 are PhD holders.

Degree	BTech	MTech	MCA	PhD	Total
Frequency	100	50	30	5	185

Table 2: Frequency table of 185 students

INTRODUCTION TO DATA V

(b) Mean and standard deviation

- **Mean** of a continuous variable indicates the central value around which most of the values taken by the variable will fall.

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

- The standard notation of representing mean is μ. Where x represents a variable, n are the number of observations given and x_i indicates ith value of the variable. Mean is calculated using Equation (1)

- **Standard deviation** measures the spread in the data values of a continuous variable. Higher spread indicates heterogeneous nature of the variable whereas, lower spread represents that all data values are similar and close to mean. Standard deviation Similar to measure called "variance".

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu_x)^2} \qquad (2)$$

- The standard notation to represent standard deviation is σ and variance is indicated by σ^2. Standard deviation is calculated using Equation (2) .

INTRODUCTION TO DATA VI

Example Illustration

Consider we are given age of 5 students. Table 3 details the age of students. Using Equations 1 and Equation 2 mean, standard deviation and variance is calculated as follows for data in Table 3.

Age	25	23	20	23	21

Table 3: One dimensional data set

$$\mu_{Age} = \mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i = 22.4$$

$$\sigma_{Age} = \sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_X)^2} = 1.94$$

$$\sigma^2_{Age} = 3.8$$

AMITY UNIVERSITY ONLINE

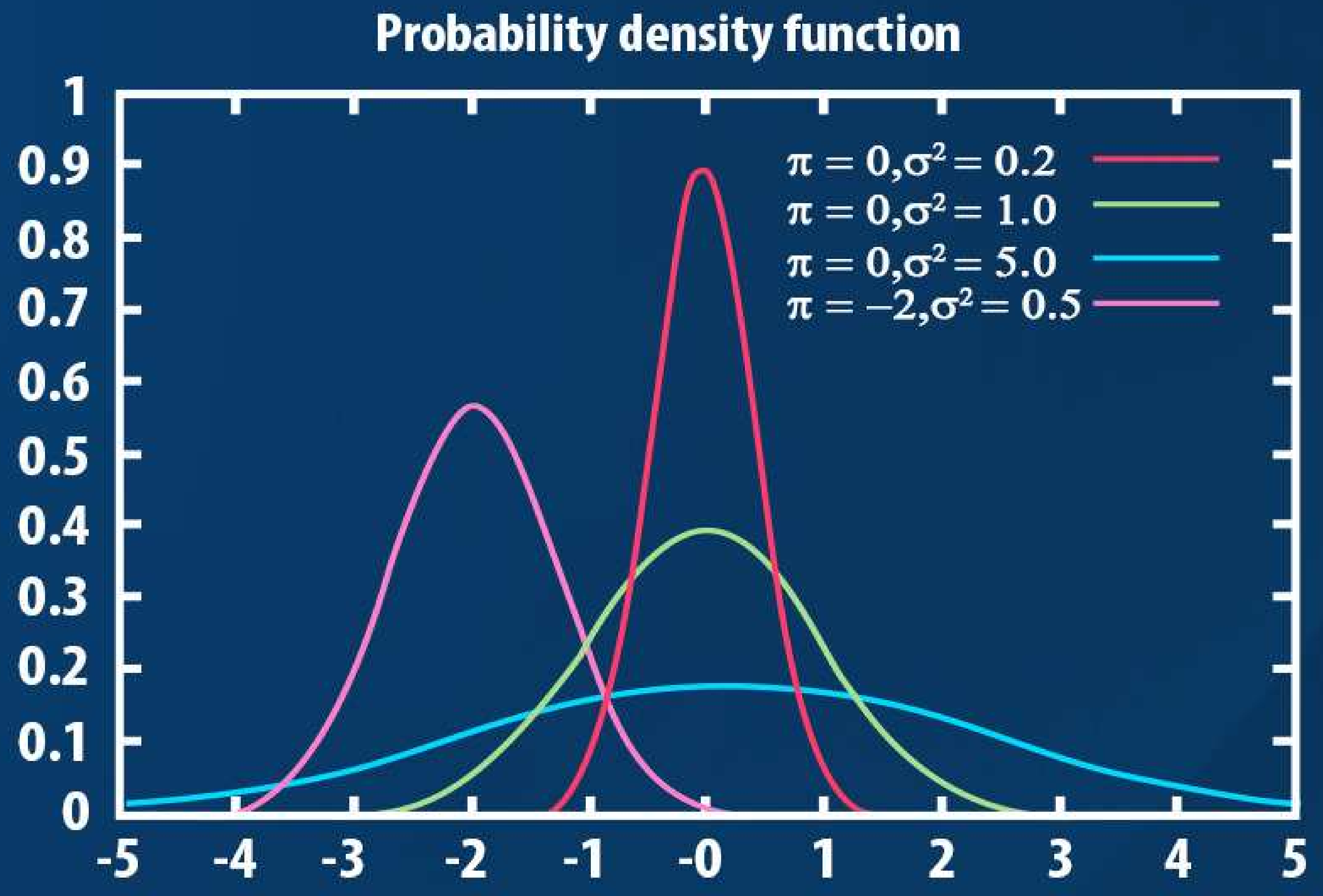
CAREERS OF TOMORROW

INTRODUCTION TO DATA VII

Data Distribution

- Data can be "distributed" (spread out) in different ways.
- Continuous data distributed here is called as "data distribution".
- For a given variable, if most of its data value is centred around mean follows "normal distribution". It is also called as "bell curve".
- The normal distribution is represented by μ and σ.
- Using these parameters, we can plot the distribution.

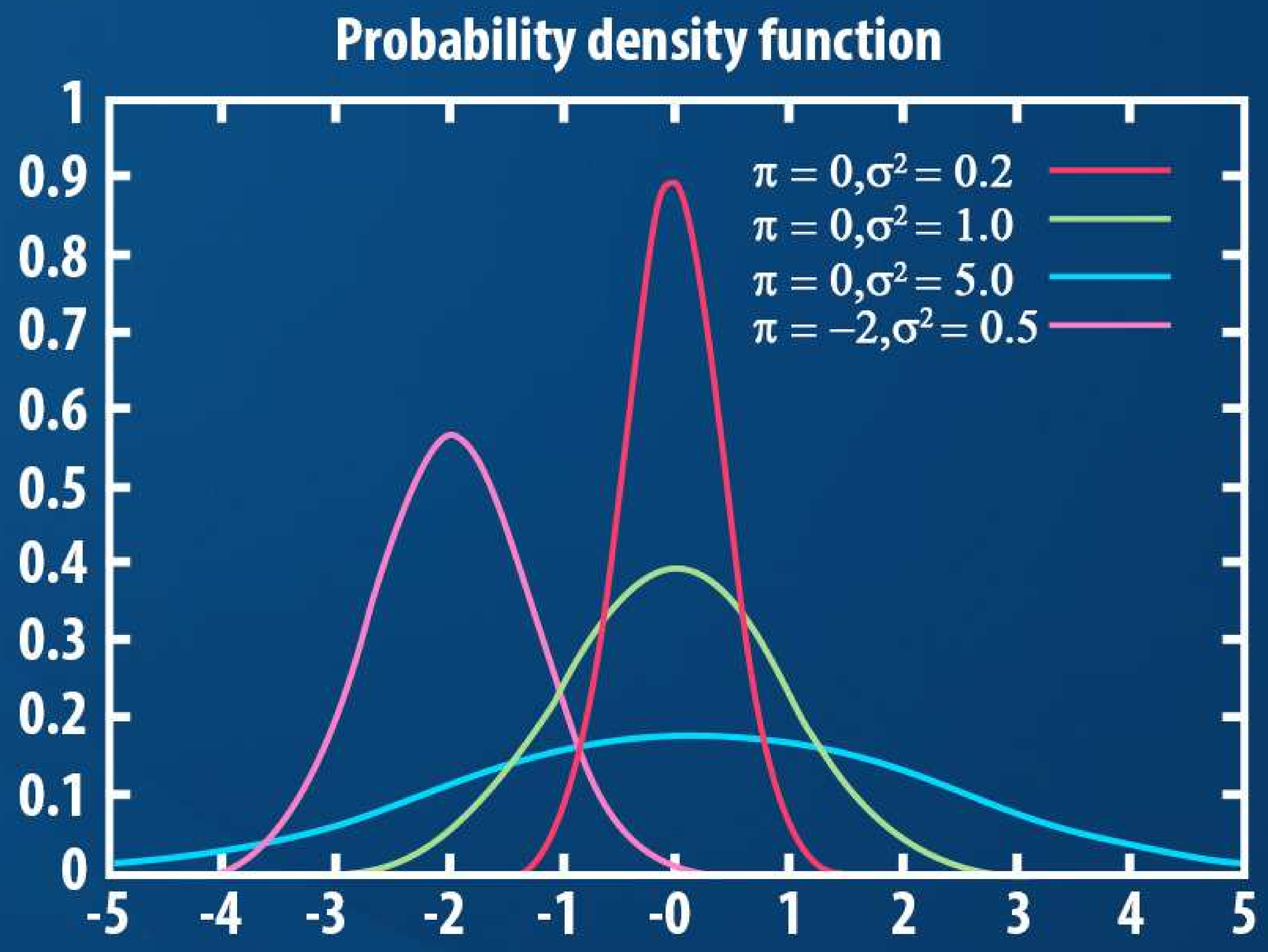
The density of the normal distribution (the height for a given value on the x axis) is computed using equation below. The parameters μ and σ are the mean and standard deviation, respectively, and define the normal distribution.



The green line is the standard normal distribution

Figure 3: Normal distributions differing in mean and standard deviation

INTRODUCTION TO DATA VIII



The green line is the standard normal distribution

$$\mathcal{P}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

(3)

Figure 4: **Normal distributions differing in mean and standard deviation**