

29.3 Classification vs Regression

Classification - Definition

Classification is the problem of identifying to which of a set of categories a new observation belongs to, on the basis of a training set of data containing observations whose category is already known.

Classification deals with predicting a qualitative (or) categorical response.

Below are the examples of Classification problems that were discussed starting from the timestamp 0.03.

The image contains handwritten mathematical notation and diagrams on a black background. At the top, it shows a dataset $\mathcal{D} = \{(\underline{x_i}, y_i)_{i=1}^n \mid x_i \in \mathbb{R}^d, y_i \in \{0, 1\}\}$. Below this, a diagram shows $y_i \in \{0, 1\}$ with arrows pointing to '-ve' and '+ve', and a bracket indicating '2: classes' and '2 class - classification / binary'. To the right, an arrow points to 'Amazon Food reviews'. At the bottom, it shows 'MNIST: $y_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} \rightarrow 10\text{-class / Multi-class classification}$ '.

In the dataset of Amazon Fine Food Reviews, the class labels are 0 and 1. If any query point is given, its label would be either 0 or 1. As it has only 2 classes, such a classification problem is called a **binary classification** problem.

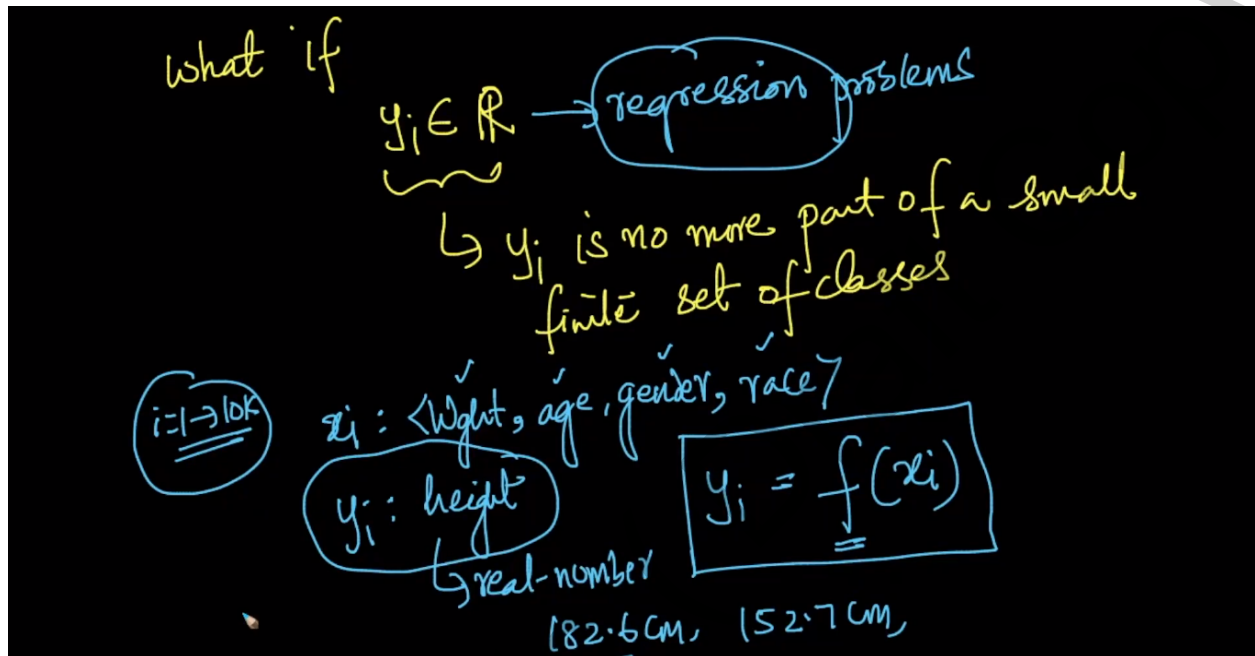
Similarly in the MNIST dataset, the class labels are the numbers from 0 to 9. If any query point is given, its label would be any one value from 0 to 9. As it has only 10 classes, such a classification problem is called a **10-class classification** problem (or) **multi-class classification** problem.

In both the examples mentioned above, the output class label comes from a finite set of values. Hence we call the problem of predicting such an output, a classification problem.

Regression - Definition

Regression is the problem of predicting a value for a new observation, on the basis of a functional relationship between two variables and on the basis of the training set of data containing observations whose value is already known.

Regression deals with predicting a quantitative (or) numerical response.



In the above example, we are predicting the height of an individual. The height value doesn't come from a finite set of values, but it comes from an infinite set of numerical values. As the output comes from an infinite set of numerical values, we call this problem a Regression problem.

Note: As the output of a regression problem comes from an infinite set of numerical values, we denote it as $y_i \in \mathbb{R}$.