ONLINE MASTERS IN **DATA SCIENCE**

DSC 255 - MACHINE LEARNING FUNDAMENTALS

NEAREST NEIGHBOR CLASSIFICATION

SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING

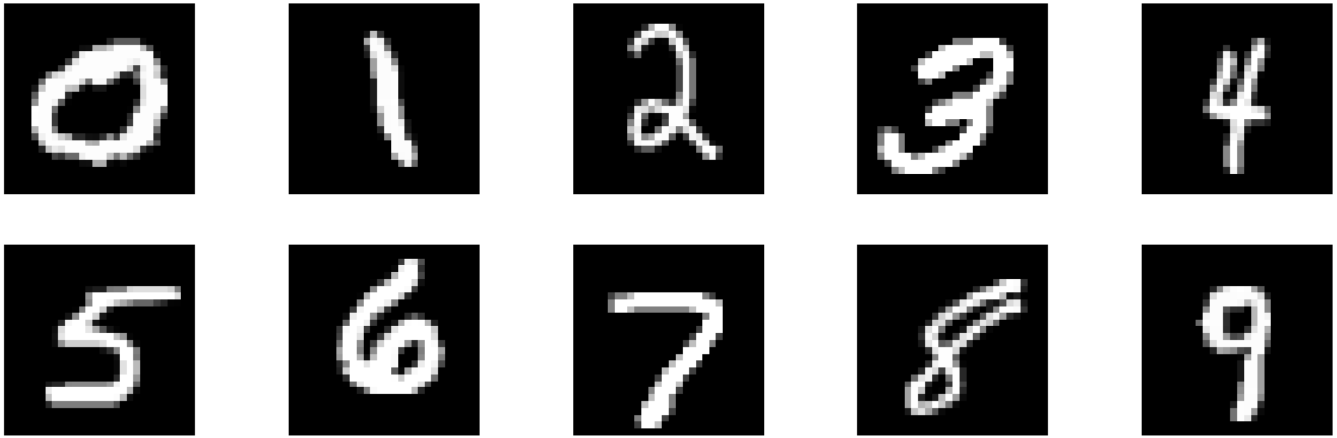
HALICIOĞLU DATA SCIENCE INSTITUTE

The problem we'll solve today

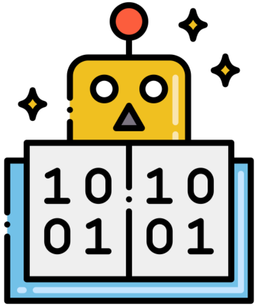
Given an image of a handwritten digit, say which digit it is.

$\Rightarrow$ 3

More examples



The machine learning approach



1. Assemble a data set

The MNIST data set of handwritten digits:

- **Training set** of 60,000 images and their labels.
- **Test set** of 10,000 images and their labels.

2. let the machine figure out the underlying patterns.

Nearest neighbor classification

Training images $x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(60000)}$

Labels $y^{(1)}, y^{(2)}, y^{(3)}, \ldots, y^{(60000)}$ are numbers in the range 0 - 9

How to **classify** a new image x?

- Find its nearest neighbor amongst the $x^{(i)}$
- Return $y^{(i)}$

How to measure the distance between images?

MNIST images

- Size 28 X 28 (total: 784 pixels)
- Each pixel is grayscale: 0-255

Stretch each image into a vector with 784 coordinates:

- Data space $\chi = \mathbb{R}^{784}$
- Label space $\gamma = \{0, 1, \ldots, 9\}$

Remember Euclidean distance in two dimensions?

$$z = (3, 5)$$

$$x = (1, 2)$$

Euclidean distance between 784-dimensional vectors *x*, *z* is

$$\| x - z \| = \sqrt{\sum_{i=1}^{784} (x_i - z_i)^2}$$

Here x_i is the ith coordinate of x.

Nearest neighbor classification

Training images $x^{(1)}, \ldots, x^{(60000)}$
Labels $y^{(1)}, \ldots, y^{(60000)}$

To classify a new image x:

- Find its nearest neighbor amongst the $x^{(i)}$ **using Euclidean distance** in $\mathbb{R}^{784}$
- Return $y^{(i)}$

How accurate is this classifier?

Accuracy of nearest neighbor on MNIST

Training set of 60,000 points

- What is the error rate on training points?

Accuracy of nearest neighbor on MNIST

Training set of 60,000 points

- What is the error rate on training points? **Zero**
  In general, **training error** is an overly optimistic predictor of future performance.

Accuracy of nearest neighbor on MNIST

Training set of 60,000 points

- What is the error rate on training points? **Zero**
  In general, **training error** is an overly optimistic predictor of future performance.

- A better gauge: separate test set of 10,000 points.
  Test error = fraction of test points incorrectly classified.

Training set of 60,000 points

- What is the error rate on training points? **Zero**
  In general, **training error** is an overly optimistic predictor of future performance.

- A better gauge: separate test set of 10,000 points.
  Test error = fraction of test points incorrectly classified.

- What test error would we expect for a *random classifier?*
  (One that picks a label 0 - 9 at random?)

Accuracy of nearest neighbor on MNIST

Training set of 60,000 points

- What is the error rate on training points? **Zero**
  In general, **training error** is an overly optimistic predictor of future performance.

- A better gauge: separate test set of 10,000 points.
  Test error = fraction of test points incorrectly classified.

- What test error would we expect for a *random classifier?*
  (One that picks a label 0 - 9 at random?) **90%**

Accuracy of nearest neighbor on MNIST

Training set of 60,000 points

- What is the error rate on training points? **Zero**
  In general, **training error** is an overly optimistic predictor of future performance.

- A better gauge: separate test set of 10,000 points.
  Test error = fraction of test points incorrectly classified.

- What test error would we expect for a *random classifier?*
  (One that picks a label 0 - 9 at random?) **90%**

- Test error of nearest neighbor: **3.09%**