NEAREST NEIGHBOR CLASSIFICATION (najstasza i najprostsza)

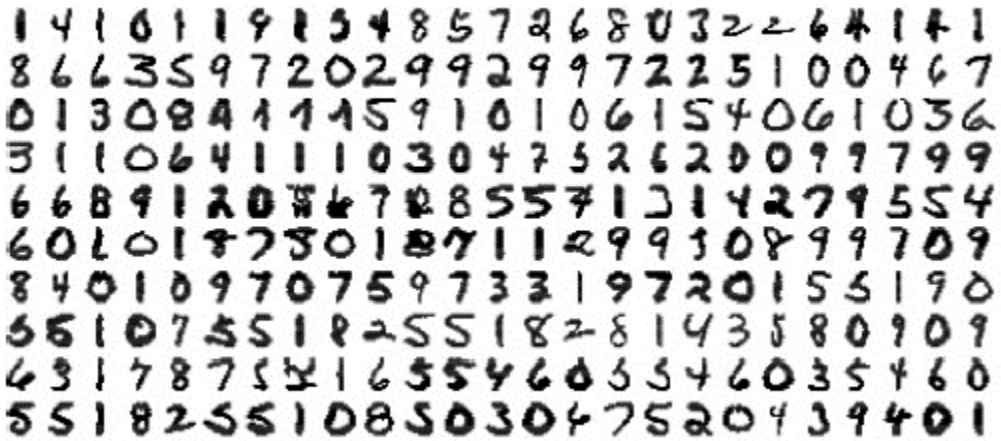
Zadanie: **Dany jest obraz ręcznie napisanej cyfry, określ jaka to jest cyfra**

$\Longrightarrow$ 3

Próbowano rozwiązać problem zapisując zestaw zasad (np. liczba pętli, odcinków itp.), ale to nie zdawało egzaminu (bo: pismo ręczne jest bardzo nieregularne i zmienne).

ML – (sposób podejścia za pomocą ML)

Utworzyć zbiór danych:



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.

Zlecić maszynie znalezienie „wzoru” rozpoznawania.

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 $y^{(i)}$

Nearest neighbor = najbardziej podobny = najmniej się różnią = najmniejszy dystans

How to measure the distance between images?

MNIST images:

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

Stretch each image into a vector with 784 coordinates:

- Data space $\mathcal{X} = \mathbb{R}^{784}$
- Label space $\mathcal{Y} = \{0, 1, \ldots, 9\}$

To classify a new image x:

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

EUCLIDIAN DISTANCE

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.

|| || - oznacza Euclidian Distance

Ocena dokładności:

- 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%.**

Ideas for improvement: (1) k-NN (2) better distance function.

K – nearest neighbours

To classify a new point:

- Find the k nearest neighbors in the training set.
- Return the most common label amongst them.

MNIST:

k	1	3	5	7	9	11
Test error (%)	3.09	2.94	3.13	3.10	3.43	3.34

W rzeczywistości często nie dysponujemy zbiorem testowym – operujemy wyłącznie zbiorem treningowym. Stosuje się metodę **Cross-walidacji** do wyznaczania **k** w metodzie **k-NN** , jak i wiele innych parametrów w innych metodach.

Dzielimy zbiór treningowy na n – kawałków (folds)