

k-nearest neighbors (kNN)

The problem we'll solve today

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



Some more examples:



MNIST dataset



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

Nearest neighbor classification

- Training images $x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(60000)}$
- Labels $y^{(1)}, y^{(2)}, y^{(3)}, \dots, y^{(60000)}$ are numbers from 0 – 9



1 4 1 6 1 1 9 1 5 4 8 5 7 2 6 8 0 3 2 2 6 4 1 4 1
8 6 6 3 5 9 7 2 0 2 9 9 2 9 9 7 2 2 5 1 0 0 4 6 7
0 1 3 0 8 4 1 1 1 5 9 1 0 1 0 6 1 5 4 0 6 1 0 3 6
3 1 1 0 6 4 1 1 1 0 3 0 4 7 5 2 6 2 0 0 9 9 7 9 9
6 6 8 9 1 2 0 5 6 7 2 8 5 5 7 1 3 1 4 2 7 9 5 5 4
6 0 2 0 1 8 7 5 0 1 8 7 1 1 2 9 9 3 0 8 9 9 7 0 9
8 4 0 1 0 9 7 0 7 5 9 7 3 3 1 9 7 2 0 1 5 5 1 9 0
6 5 1 0 7 5 5 1 8 2 5 5 1 8 2 8 1 4 3 5 8 0 9 0 9
4 3 1 7 8 7 5 2 1 6 5 5 4 6 0 5 5 4 6 0 3 5 4 6 0
5 5 1 8 2 5 5 1 0 8 5 0 3 0 4 7 5 2 0 4 3 9 4 0 1

Nearest neighbor classification

- Training images $x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(60000)}$
- Labels $y^{(1)}, y^{(2)}, y^{(3)}, \dots, y^{(60000)}$ are numbers from 0 – 9



1 4 1 6 1 1 9 1 5 4 8 5 7 2 6 8 0 3 2 2 6 4 1 4 1
8 6 6 3 5 9 7 2 0 2 9 9 2 9 9 7 2 2 5 1 0 0 4 6 7
0 1 3 0 8 4 1 1 1 5 9 1 0 1 0 6 1 5 4 0 6 1 0 3 6
3 1 1 0 6 4 1 1 1 0 3 0 4 7 5 2 6 2 0 0 9 9 7 9 9
6 6 8 9 1 2 0 5 6 7 2 8 5 5 7 1 3 1 4 2 7 9 5 5 4
6 0 2 0 1 8 7 5 0 1 8 7 1 1 2 9 9 3 0 8 9 9 7 0 9
8 4 0 1 0 9 7 0 7 5 9 7 3 3 1 9 7 2 0 1 5 5 1 9 0
6 5 1 0 7 5 5 1 2 5 5 1 8 2 8 1 4 3 5 8 0 9 0 9
4 3 1 7 8 7 5 2 1 6 5 5 4 6 0 5 5 4 6 0 3 5 4 6 0
5 5 1 8 2 5 5 1 0 8 5 0 3 0 4 7 5 2 0 4 3 9 4 0 1

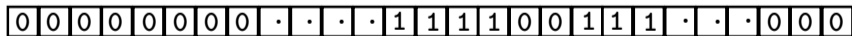
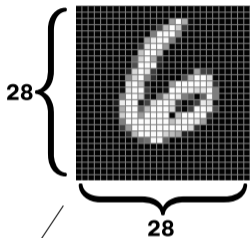
How to classify a new image x ?

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



Data as vectors

How to measure the distance between images?



Stretch each image into a vector with 784 coordinates

$$x^{(1)} = (0, 0, 0, \dots, 0.6, 1, 1, 1, 0, 0, 1, 1, 0.8, \dots, 0, 0, 0)$$

$$y^{(1)} = 6$$

The distance function

Euclidean distance in two dimensions is

Euclidean distance in higher dimension

Two images a and b :

$$a = (a_1, a_2, a_3, \dots, a_{784})$$

$$b = (b_1, b_2, b_3, \dots, b_{784})$$

The Euclidean distance between a and b is

$$\begin{aligned} \|a - b\|_2 &= \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_{784} - b_{784})^2} \\ &= \sqrt{\sum_{i=1}^{784} (a_i - b_i)^2} \end{aligned}$$

Nearest neighbor classification

Training images $x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(60000)}$ Labels $y^{(1)}, y^{(2)}, y^{(3)}, \dots, y^{(60000)}$

1 4 1 6 1 1 9 1 5 4 8 5 7 2 6 8 0 3 2 2 6 4 1 4 1
8 6 6 3 5 9 7 2 0 2 9 9 2 9 9 7 2 2 5 1 0 0 4 6 7
0 1 3 0 8 4 1 1 1 5 9 1 0 1 0 6 1 5 4 0 6 1 0 3 6
3 1 1 0 6 4 1 1 1 0 3 0 4 7 5 2 6 2 0 0 9 9 7 9 9
6 6 8 9 1 2 0 5 6 7 2 8 5 5 7 1 3 1 4 2 7 9 5 5 4
6 0 2 0 1 8 7 5 0 1 8 7 1 1 2 9 9 3 0 8 9 9 7 0 9
8 4 0 1 0 9 7 0 7 5 9 7 3 3 1 9 7 2 0 1 5 5 1 9 0
6 5 1 0 7 5 5 1 8 2 5 5 1 8 2 8 1 4 3 5 8 0 9 0 9
4 3 1 7 8 7 5 5 1 6 5 5 4 6 0 5 5 4 6 0 3 5 4 6 0
5 5 1 8 2 5 5 1 0 8 5 0 3 0 4 7 5 2 0 4 3 9 4 0 1

Nearest neighbor classification

Training images $x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(60000)}$ Labels $y^{(1)}, y^{(2)}, y^{(3)}, \dots, y^{(60000)}$

1 4 1 6 1 1 9 1 5 4 8 5 7 2 6 8 0 3 2 2 6 4 1 4 1
8 6 6 3 5 9 7 2 0 2 9 9 2 9 9 7 2 2 5 1 0 0 4 6 7
0 1 3 0 8 4 1 1 1 5 9 1 0 1 0 6 1 5 4 0 6 1 0 3 6
3 1 1 0 6 4 1 1 1 0 3 0 4 7 5 2 6 2 0 0 9 9 7 9 9
6 6 8 9 1 2 0 5 6 7 2 8 5 5 7 1 3 1 4 2 7 9 5 5 4
6 0 2 0 1 8 7 5 0 1 8 7 1 1 2 9 9 3 0 8 9 9 7 0 9
8 4 0 1 0 9 7 0 7 5 9 7 3 3 1 9 7 2 0 1 5 5 1 9 0
6 5 1 0 7 5 5 1 2 5 5 1 8 2 8 1 4 3 5 8 0 9 0 9
4 3 1 7 8 7 5 5 1 6 5 5 4 6 0 5 5 4 6 0 3 5 4 6 0
5 5 1 8 2 5 5 1 0 8 5 0 3 0 4 7 5 2 0 4 3 9 4 0 1

To classify a new image x

- Find its nearest neighbor **in Euclidean distance**, say $x^{(i)}$
- Return $y^{(i)}$



Accuracy of Nearest Neighbor on MNIST

1 4 1 0 1 1 9 1 5 4 8 5 7 2 6 8 0 3 2 2 6 4 1 4 1
8 6 6 3 5 9 7 2 0 2 9 9 2 9 9 7 2 2 5 1 0 0 4 6 7
0 1 3 0 8 4 1 1 1 5 9 1 0 1 0 6 1 5 4 0 6 1 0 3 6
3 1 1 0 6 4 1 1 1 0 3 0 4 7 5 2 6 2 0 0 9 9 7 9 9
6 6 8 9 1 2 0 8 6 7 2 8 5 5 7 1 3 1 4 2 7 9 5 5 4
6 0 2 0 1 8 7 5 0 1 8 7 1 1 2 9 9 3 0 8 9 9 7 0 9
8 4 0 1 0 9 7 0 7 5 9 7 3 3 1 9 7 2 0 1 5 5 1 9 0
6 6 1 0 7 5 5 1 2 5 5 1 8 2 8 1 4 3 5 8 0 9 0 9
6 3 1 7 8 7 5 2 1 6 5 5 4 6 0 5 5 4 6 0 3 5 4 6 0
5 5 1 8 2 5 5 1 0 8 5 0 3 0 4 7 5 2 0 4 3 9 4 0 1

Predictions on all points in the **Training set**

Question: What is the accuracy?

Accuracy of Nearest Neighbor on MNIST



1 4 1 6 1 1 9 1 5 4 8 5 7 2 6 8 0 3 2 2 6 4 1 4 1
8 6 6 3 5 9 7 2 0 2 9 9 2 9 9 7 2 2 5 1 0 0 4 6 7
0 1 3 0 8 4 1 1 1 5 9 1 0 1 0 6 1 5 4 0 6 1 0 3 6
3 1 1 0 6 4 1 1 1 0 3 0 4 7 5 2 6 2 0 0 9 9 7 9 9
6 6 8 9 1 2 0 8 6 7 2 8 5 5 7 1 3 1 4 2 7 9 5 5 4
6 0 2 0 1 8 7 5 0 1 8 7 1 1 2 9 9 3 0 8 9 9 7 0 9
8 4 0 1 0 9 7 0 7 5 9 7 3 3 1 9 7 2 0 1 5 5 1 9 0
6 6 1 0 7 5 5 1 2 2 5 5 1 8 2 8 1 4 3 5 8 0 9 0 9
6 3 1 7 8 7 5 5 1 6 5 5 4 6 0 5 5 4 6 0 3 5 4 6 0
5 5 1 8 2 5 5 1 0 8 5 0 3 0 4 7 5 2 0 4 3 9 4 0 1

Predictions on all points in the **Test set**

Question: What is the accuracy?

Examples of errors

Test set of 10,000 points

- 309 are misclassified

Examples of errors:

Test image					
Nearest neighbor					

Ideas for improvement: k -NN

k -nearest neighbor classification

To classify a new point:

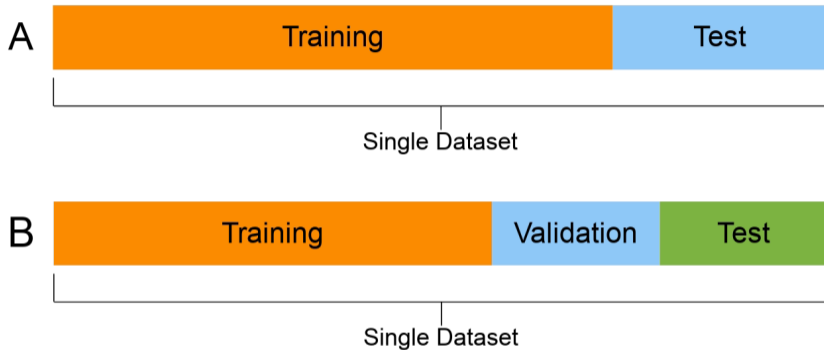
- Find the k **nearest neighbors** in the training set
- Return the most common label amongst them

MNIST:

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

need to find k before final eval on the test set

Validation



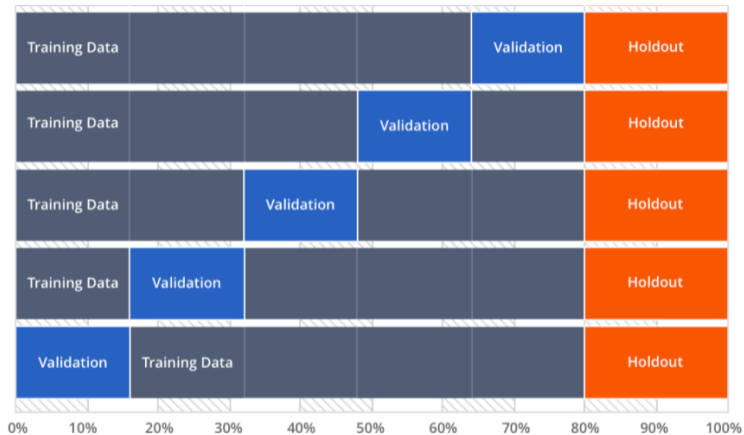
Train on Training set with $k = 1 \implies$ Evaluate on the Validation set

Train on Training set with $k = 3 \implies$ Evaluate on the Validation set

Train on Training set with $k = 5 \implies$ Evaluate on the Validation set

\vdots

Cross-validation



Other distance function

$$a = (a_1, a_2, \dots, a_m) \quad b = (b_1, b_2, \dots, b_m).$$

- Cosine similarity

$$d_{\cos}(a, b) = \frac{a \cdot b}{\|a\|_2 \|b\|_2} = \frac{a}{\|a\|_2} \cdot \frac{b}{\|b\|_2}$$

measures the angle between vector a and b .

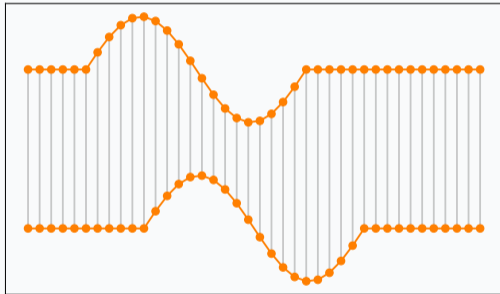
$$-1 \leq d_{\cos}(a, b) \leq 1.$$

Examples

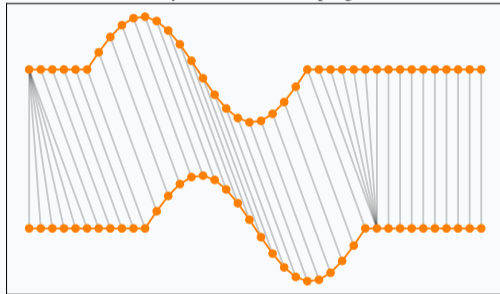
$$a = (1, 2, 2) \quad b = (3, 4, 0)$$

Distance between time series

Euclidean distance



Dynamic Time Warping



use **dynamic time warping**

k -NN regression

y is continuous $x =$ test data

$(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)$ are k -nearest neighbors of x .

Prediction:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^k y_i$$

Example