PREDICTION PROBLEMS

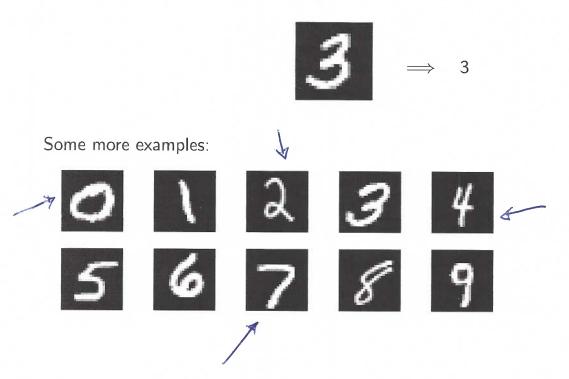
Nearest neighbor classification

Topics we'll cover

- Mhat is a classification problem?
- ② The training set and test set
- 3 Representing data as vectors
- Distance in Euclidean space
- **5** The 1-NN classifier
- 6 Training error versus test error
- 7 The error of a random classifier

The problem we'll solve today

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



The machine learning approach

Assemble a data set:

```
1416119154857268U32264141
86635972029929977225100467
0130844145910106154061036
3110641110304752620099799
6689120847285571314279554
60101778018711299308999709
8401097075973319720155190
5510755182551828143580909
```

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.

And let the machine figure out the underlying patterns.

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 in the range 0-9

> 1416119134857268U32264141 8663S97202992997225100467 0130841145910106154061036 3110641110304752620099799 6689120%67885571314279554 60101775018711299308999709 8401097075973319720155190 5510755182551828143580109



How to classify a new image x?

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

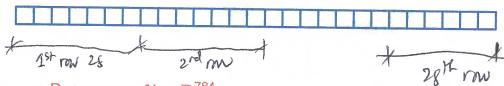
The data space

How to measure the distance between images?

MNIST images:

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

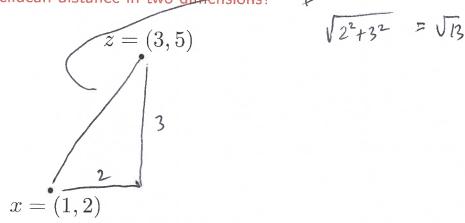
Stretch each image into a vector with 784 coordinates:



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

The distance function





Euclidean distance in higher dimension

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 *i*th coordinate of x.

Nearest neighbor classification

Training images $x^{(1)}, \dots, x^{(60000)}$, labels $y^{(1)}, \dots, y^{(60000)}$

1416119134857868U32264141 8663597202992997225100467 0130841145910106154061036 3110641110304752620077799 6684120867885571314274554 6010177301871129910899709 8401097075973319720155190 5510755182551828143580109 4317875246554605546035460



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? 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%.

Examples of errors

Test set of 10,000 points:

- 309 are misclassified
- Error rate 3.09%

Examples of errors:

query quy is 6 but theylor itwas form

5 8 7 8 9 9 9 5 en ar 8