Tomasz Heimowski

IHS Global

tomek.heimowski@gmail.com

@theimowski

Digit Recognizer Dojo

A Gentle Introduction to Machine Learning

The Goal

- Take a Kaggle data science competition

- Write some code and **have fun**
- Write a classifier, from scratch, using F#
- Learn some Machine Learning concepts

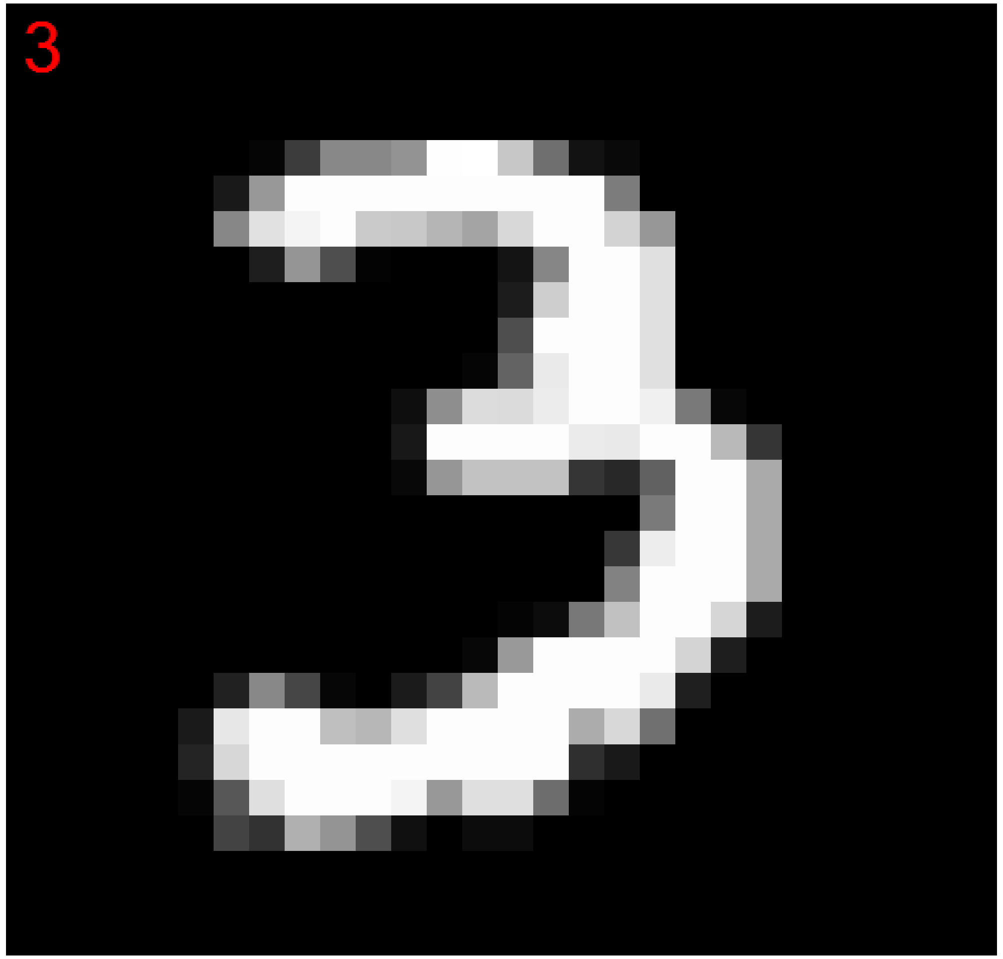
The format

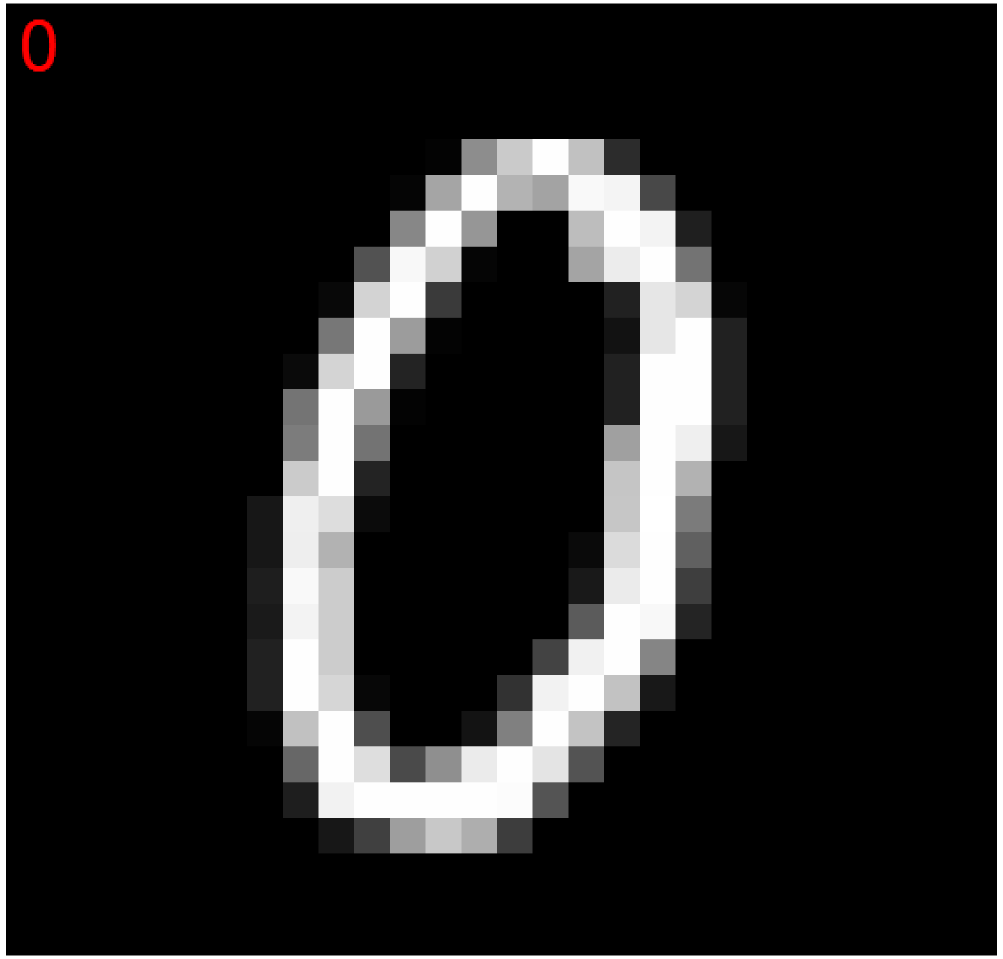
- Brief introduction to the problem
- You code in teams, I help out

Kaggle Digit Recognizer contest

- » http://www.kaggle.com/c/digit-recognizer
- » Dataset of hand-written digits
- » Goal = automatically recognize digits
- » Training sample = 50,000 examples
- » Contest = predict 20,000 “unknown” digits

The data “looks like that”

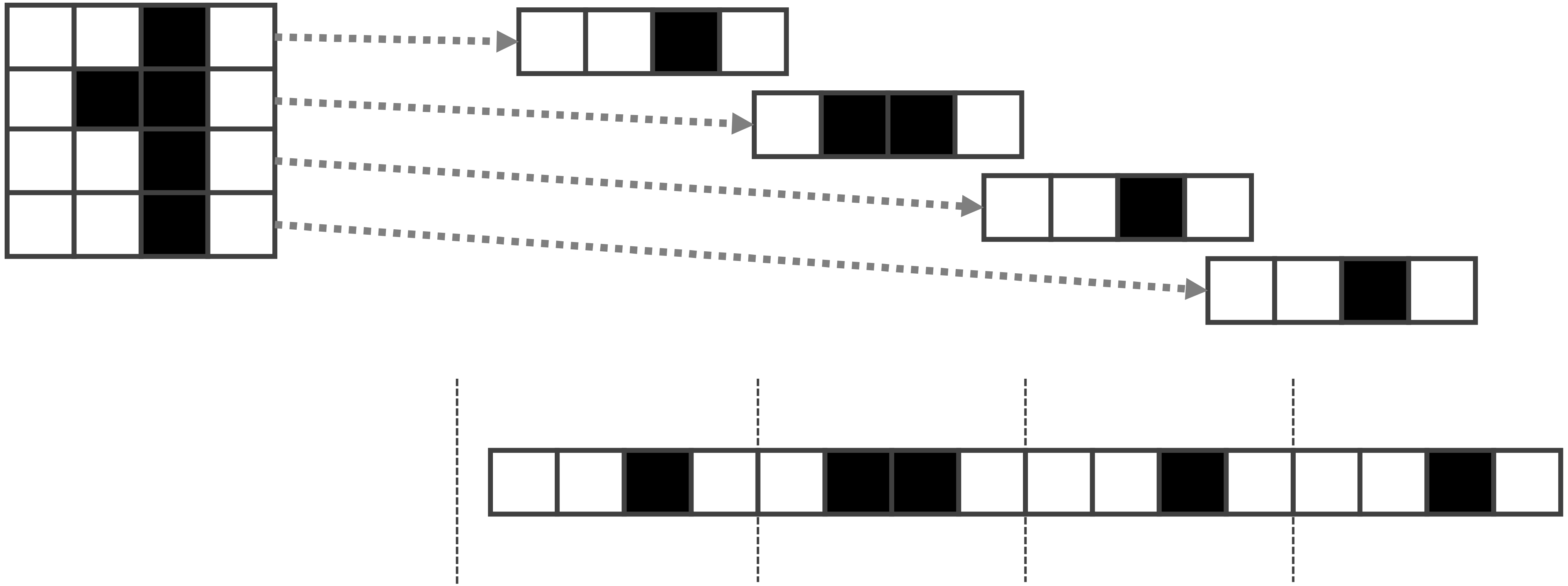




Real sample

- 28 x 28 pixels
- Grayscale (0 = black, to 255 = white)
- Flattened: each record = Number + 784 pixels
- CSV file
- Reduced dataset: 5,000 training, 500 validation

Illustration (simplified 4x4 data)



1,0,0,255,0,0,255,255,0,0,0,255,0,0,0,255,0

Actual number | **Each pixel, encoded from 0 to 255**

What’s a classifier?

- “Give me an unknown data point and I will predict what class it belongs to”
- In this case, classes = 0, 1, 2, … 9
- Unknown data point = scanned digit, without the class it belongs to

The KNN Classifier

» KNN = K Nearest Neighbors

» Given an unknown subject to classify,

» Lookup all the known examples,

» Find the K closest examples,

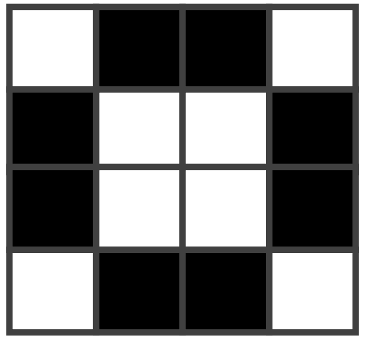
» Take a majority vote,

» Predict what the majority says

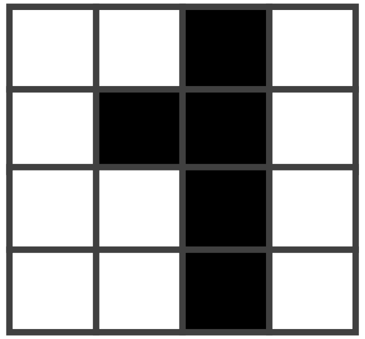
Illustration: 1-nearest neighbor

Sample

0

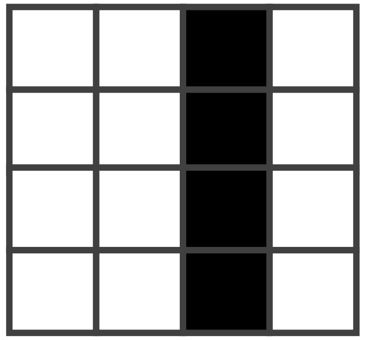


1



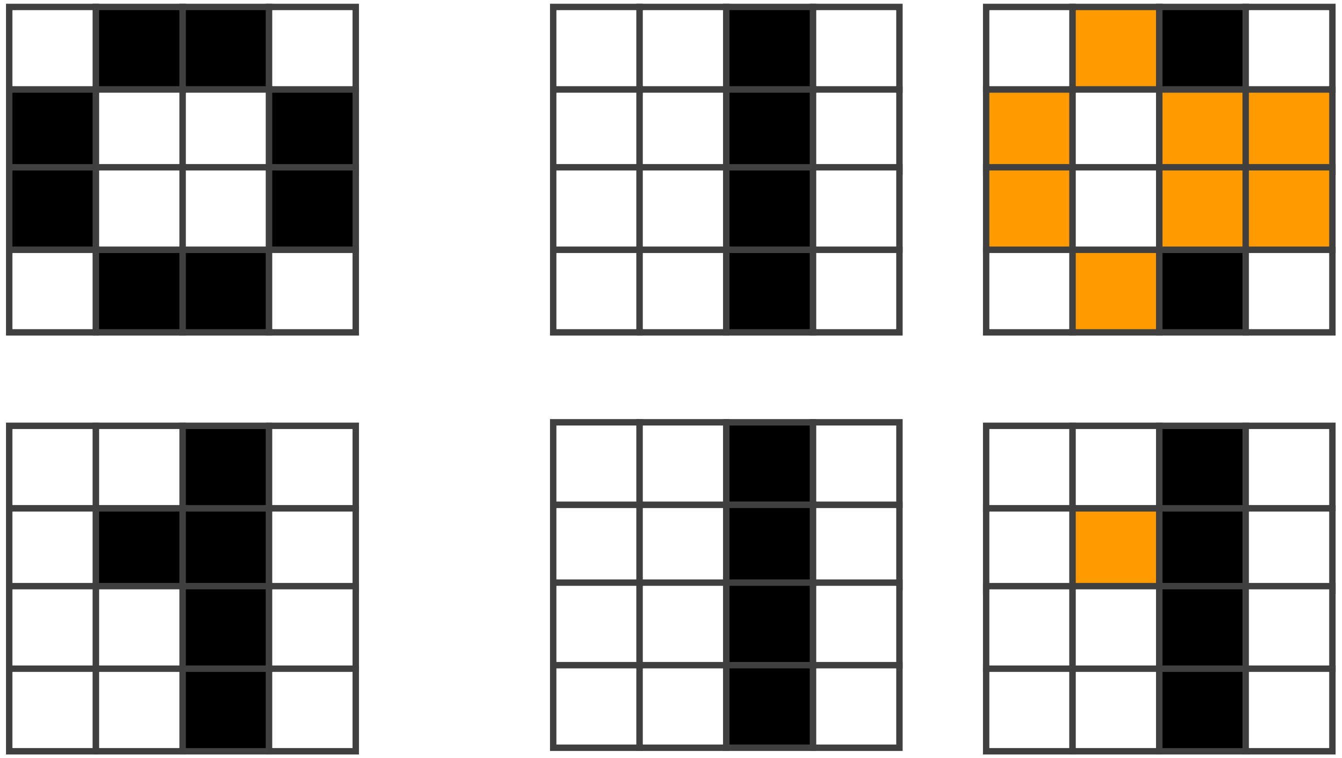
Unknown

?



Which element in the Sample is the most similar / closest to the Unknown item we want to classify?

Illustration: 1-nearest neighbor (2)



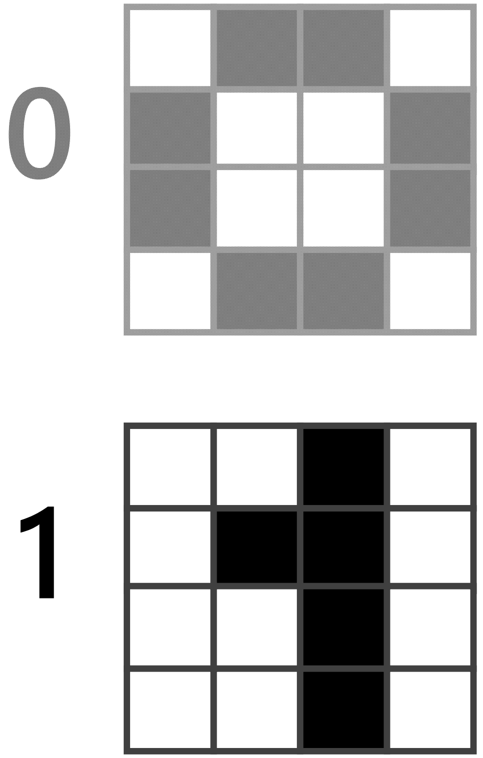
$$D = \sqrt{255^2 + 255^2 \ldots + 255^2}$$

$$D = \sqrt{255^2}$$

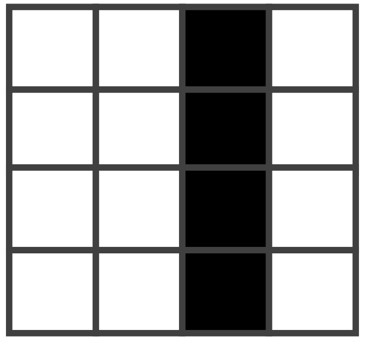
Compare images, pixel by pixel

We compute the distance between each element of the Sample, and the Item we try to classify

Illustration: 1-nearest neighbor (3)



The second example is closest, therefore we predict that the unknown Item has the same label, and is a 1



Questions?

Your mission

- Code a 1-nearest-neighbor classifier

- Guided script available at:
- www.github.com/c4fsharp/Dojo-Digits-Recognizer

A few recommendations

- » www.github.com/c4fsharp/Dojo-Digits-Recognizer
  - › No need to create new Library Project – just use .fsx
  - › „Alt + Enter" – Execute selected code in interactive
  - › Watch out for whitespaces
  - › Try to avoid red squigglies
  - › When in trouble - ask for help