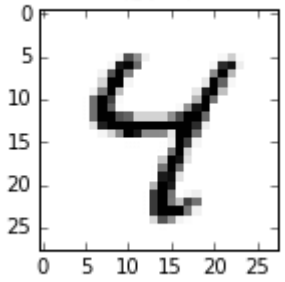
The MNIST Dataset of Handwritten Numbers

A brief introduction to the MNIST Dataset of Handwritten Digits.

Recognising human handwriting is an ideal challenge for testing artificial intelligence because the problem is sufficiently hard and fuzzy. It's not clear and defined as multiplying lots of lots of numbers.

Getting computers to correctly classify what an image contains, sometimes called the *image recognition* problem, has withstood decades of attack. Only recently has good progress been made, and methods like neural networks have been a crucial part of these leaps forward.

To give you a sense of how hard the problem of image recognition is, we humans will sometimes disagree on what an image contains. We will easily disagree what a handwritten character actually is, particularly if the character was written in a rush or without care. Have a look at the following handwritten number. Is it a 4 or a 9?



There is a collection of images of handwritten numbers used by artificial intelligence researchers as a popular set to test their latest ideas and algorithms. The fact that the collection is well known and popular means that it is easy to check how well our latest crazy idea for image recognition works compared to others. That is, different ideas and algorithms are tested against the same data set.

That data set is called the MNIST database of handwritten digits and is available from the respected neural network researcher Yann LeCun's website

http://yann.lecun.com/exdb/mnist/. That page also lists how well old and new ideas have performed in learning and correctly classifying these handwritten characters. We'll come back to that list several times to see how well our own ideas perform against professionals!

The format of the MNIST database isn't the easiest to work with, so others have helpfully created data files in a simpler format, such as this one http://pjreddie.com/projects/mnist-in-csv/. These files are called CSV files, which means each value is plain text separated by commas (comma separated values). You can easily view them in any text editor, and most spreadsheet or data analysis software will work with CSV files. They are pretty much a universal standard. This website provides two CSV files:

- *Training* set http://www.pjreddie.com/media/files/mnist_train.csv
- *Test* set http://www.pjreddie.com/media/files/mnist_test.csv

As the names suggest, the *training set* is the set of $60,000$ labeled examples used to train the neural network. *Labelled* means the inputs come with the desired output, that is, what the answer should be.

The smaller *test set* of $10,000$ is used to see how well our idea or algorithm works. This too contains the correct labels so we can check to see if our own neural network got the answer right or not.

The idea of having separate training and test datasets is to make sure we test against data we haven't seen before. Otherwise, we could cheat and simply memorize the training data to get a perfect, albeit deceptive, score. This idea of separating training from test data is common across machine learning.