# Representation and classification of data

Wojciech Czaja and David Levermore

January 25, 2016

# Outline

# Machine Learning Project

Project #1 is devoted to the problem of classification of handwritten digits collected by USPS. The origin of this problem comes from a series of experiments in the early 90s which were aimed at developing methods for automated handwriting recognition. One of the main players in this was the United States Postal Service. USPS, in collaboration with Center of Excellence in Document Analysis and Recognition (SUNY Buffalo), collected a large number of addresses and developed a number of databases. The details are described in:

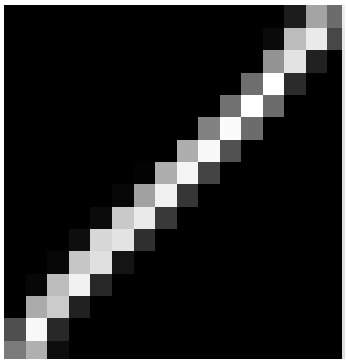
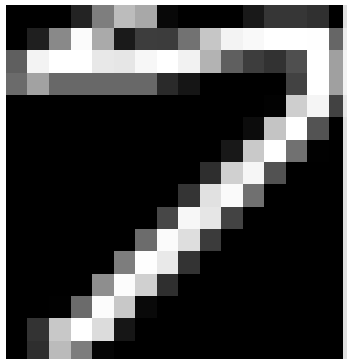
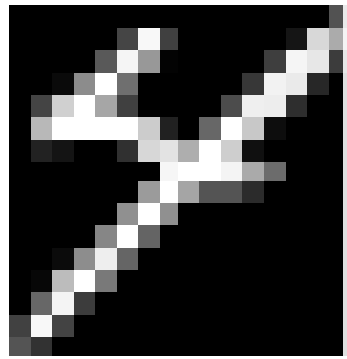
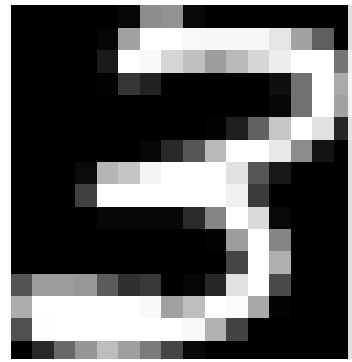
J. J. Hull., "A database for handwritten text recognition research", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 5 (1994), pp. 550–554.

# Project 1 Data

- Data for Project #1 comes from the website of late Prof. Samuel Roweis:
- http://www.cs.nyu.edu/∼roweis/data.html
- It consists of 1100 examples of each class of digits “0” through “9”.
- These examples are all in the form of $16 \times 16$ 8-bit grayscale images.

# Example

Examples of digits: 1, 7, 4, 3:

# Goals

The major goal of this project is to develop and test methods for classification of the handwritten data (to be discussed further in class), which are optimized on the training set and then applied to the testing set to assess the performance of the developed methodology.
By **classification** we understand an algorithmic method to assign any given new element of the dataset to one of a priori provided classes (categories).
A **training set** is a set of data used to discover potentially best parameters for the method selected for classification.
A **testing set** is a set of data used to verify how well these selected parameters (and the method) perform.

# Potential problems and issues

- One of the first issues you will encounter is the problem of representing images in a form that mathematical algorithms can understand. This will require vectorization and quantization of the imagery.
- Selection of testing and training sets from the given data is another problem that has to be solved early. Once you make a decision on what are these sets, you cannot make changes and substitutions. A good representative training set is crucial for successful classification of the testing data.
- Often, a new representation of the data is necessary to provide us with a chance to do well.
- There are many parameters, both in the data representation and in the classifications methods, that need to be optimized.
- Overfitting is a common problem in data classification.

# Problems

- We will learn at least one method of data classification: the **k-nearest neighbors classifier**. In your projects you may use other classification schemes that you may have learned from other classes.
- Projects will explore different structures of the training data. Some problems will have a balanced training, and some projects will have to work with unequal sets of training digits.
- In all projects you will need to find optimal parameters. Sometimes this may require subdiving the digits into further subclasses.
- In some of the projects you will work with the standard format of the vectorized data and in some projects you will need to apply further data representation techniques to obtain better classification results.