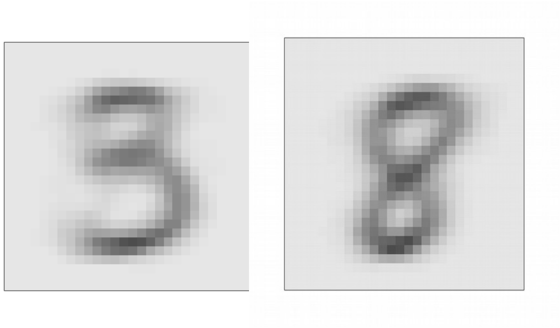
Data Mining Spring 2016

Project 4: Cluster Analysis

Assigned:	4/7/2016
Due:	5/2/2016 (via Canvas)
Points:	100

Please submit your report in PDF format.



Introduction

We will use the Numbers data set for this project. The data set contains images of handwritten digits. Recognizing handwritten digits is already a mature technology. The task of this project is to extract features and cluster the images into homogeneous groups. These groups do not necessarily have to be groups of the same digit, but should also group the data by the way a digit is written. For each image you have 28x28 pixels with 256 gray values (8 bit). The data and some code to get you started can be found on the course web site under data for projects (see http://michael.hahsler.net/SMU/EMIS7332/data/numbers/).

Follow the CRISP-DM Framework for your Report

Write a report covering in detail all the steps of the project. The results need to be reproducible using only this report. Describe all assumptions you make and programs/code that you have used.

3. Data Preparation [35 points]

- Describe several ways you could preprocess the data and extract features. Describe why these steps might be helpful. **Note:** Use the web to research this.
- Construct at least 3 features to the ones that are discussed and created in class. More is better!

4. Modeling [50 points]

- Perform cluster analysis using several methods (at least k-means and hierarchical clustering) for different features.
- How did you determine a suitable number of clusters for each method?
- Use internal validation measures to describe and compare the clusterings and the clusters (some visual methods would be good).
- Use external validation measures to describe the clusterings and the clusters. You can find the actual digits in the images in the file number_labels.csv.

5. Evaluation [5 points]

- Describe your results. What findings are the most interesting?

Exceptional Work [10 points]