K-means clustering

Alexandre Allauzen
Nicolas Pécheux

January 2014

1 Overview

The goal of this assignment is to implement the k-means algorithm on a set of images. We will use the MNIST dataset, which is a standard dataset for machine learning. This dataset of handwritten digits has a training set of 60,000 examples, and a test set of 10,000 examples. The digits have been size-normalized and centered in a fixed-size image.

Question 1. Recall the main idea and the principle of K-means algorithm.

2 Data handling

First of all take a look at the data and see what we can easily do with python: `http://perso.limsi.fr/Individu/allauzen/webpages/pmwiki.php?n=Cours.Main#toc7`

Write a program to load the dataset and take a look at some images.

Question 2. What is the structure of the dataset? How is represented each image? Are we going to use the labels in this assignment?

3 k-means clustering

To start, the number of clusters is set to 10.

Question 3. Why setting the number of clusters to 10? Do you think it is a good idea?

Question 4. What is the distance between two images? How can we compute it?

Implement that function.

Question 5. How can we chose the initial mean vectors?

The k-means implementation will follow these steps:

1. Write a function to initialize the mean vectors.
2. Write a function that given the set of mean vectors compute the affectations of each examples.
3. The next function will update the mean vectors knowing the affectations.
4. Finally write the main function that wrap the previous functions.

4 Experimentation

With 10 clusters, run the k-means algorithm on the training set for 20 iterations. At the end of each iteration print the distance between the new centers and their previous values, as well as the *squared error sum* quality criterium $\mathcal{J}$.

Question 6. What do you think about the number of iterations? Describe the evolution of the quality criterum.

Question 7. What happens with a different initialization of the mean vectors?

Question 8. Plot the images that correspond to the mean vectors at the end. You may also explore some images among one cluster. Comment.

Question 9. Restart the previous experience, but with 20 clusters. Comment.