Pixel Clustering

Alesis Novik and Guido Sanguinetti

Out: 5 March 2012
Due: 21 March 2012

1 Introduction

Image segmentation is an important part of image preprocessing that allows to find boundaries between different objects. The argument for that is that pixels representing different objects share different characteristics. Before expert annotation, we can't tell what classes are in the image, so clustering approaches work well in this scenario.
For this project you will use pixel information that represent 3 different clusters. The dimensionality of the data is reduced to 3 using the Principal Component Analysis (PCA) for simplicity.
This project will require you to use Matlab to implement your solution and it will have to be runnable on a DICE computer.
If you're interested in the full dataset description you can find it at
`http://archive.ics.uci.edu/ml/machine-learning-databases/image/segmentation.names`

2 Data

The data files for this assignment are available from:
`http://www.inf.ed.ac.uk/teaching/courses/inf2b/coursework/cwk2.html`
To unpack the files run this command on DICE:

```
tar -zxvf inf2b-pixelFiles.tar.gz
```

The *data_90.mat* file will contain 90 pixel data and the *data_900.mat* file will contain 900 pixel data. The *true_90.mat* will contain the true cluster labels

for the *data_90.mat* and the *true_900.mat* will contain the true cluster labels for the *data_900.mat*.

3 Tasks

The assignment consists of a programming and a report part.

3.1 Programming

The main method used in this task will be k-means. To avoid diverse results and different solutions this is the initialisation procedure you should use:

1. Compute a distance matrix between all points.
2. Take 2 points that are furthest apart and set them as $mean_1$ and $mean_2$.
3. For every additional mean required take the distances from a point to all the means and multiply them. Do this for each point. Make the point with the maximum resulting value a new mean.

The programming section will have 4 sub-tasks:

1. Cluster the data in *data_90.mat* into 2, 3, 4, 5 clusters using k-means and provide an sum-squared error plot.
2. Obtain the clusters for 3-means using *data_90.mat*.
3. Fit 3 full covariance Gaussians to the clusters.
4. Classify the *data_900.mat* using the Gaussians.
5. Cluster the *data_900.mat* using 3-means.
6. Using the true clusters, provide confusion matrices for results of steps 4 and 5.

Important! The code has to run on the DICE version of Matlab.

3.2 Report

The report has to contain the plots generated by the code as well as explanations of what the code does. Furthermore, the report has to have analysis of the results, specifically the comparisons of different methods and explanation why such results were received.
The report has to be called either *report.ps* (if it's PostScript) or *report.pdf* (if it's in PDF)

4 Submission

You will have to use *tar* to package the following files:

- Your report.
- All the code.

The archive has to be named *inf2b-cwk2.tar* and submitted via DICE using the

```
submit inf2b b2 inf2b-cwk2.tar
```

command.
If there are any questions or you find something unclear don't hesitate to ask via e-mail.