

Homework 5 – Unsupervised learning

Machine Learning / Greg Hamerly

April 16, 2007

This homework is due in class on Friday, April 27, 2007.

For this homework, you will use Matlab to implement k -means. Then you will apply the k -means algorithm to the problem of automatically segmenting an image.

Implement k -means Implement the k -means algorithm in Matlab. The algorithm should take as inputs the following:

- A matrix X of size n by d , with n examples in d dimensions. These are the data to cluster.
- A matrix U of size k by d , with k cluster centers in d dimensions.

And the output of the algorithm should be the final U , and the labels of each example (a vector of length n , with values in $[1, 2, \dots, k]$). The algorithm should also report the k -means metric – the sum of squared distances between each example and its closest cluster center.

Test k -means on a small 2-dimensional data set Download this dataset:

http://cs.baylor.edu/~hamerly/5v93_07s/kmeans_simple_dataset.txt

and perform k -means clustering with $k = 5$. Plot and evaluate your results. Try other values of k as well. Use the following code to visualize your results:

```
function plot_kmeans(X, U)
    [n, d] = size(X);
    if (d < 2)
        error('this function only works with >= 2 dimensions');
    end
    if (size(U,2) ~= d)
        error('dimension of X and U should be the same');
    end

    hold off;
    plot(X(:,1), X(:,2), 'b.');
```

```
    hold on;
    h = plot(U(:,1), U(:,2), 'kx');
```

```
    set(h, 'MarkerSize', 15, 'LineWidth', 4);
```

Image segmentation Use the k -means algorithm to do automatic segmentation of a color image. The most difficult part of this is to put the image into a representation that k -means can process, and then after segmenting it, putting it back into a visible format.

Use the Matlab function `imread()` to read an image in from the disk, and then convert it into the format described below. I recommend using a small, simple image, preferably less than 100 by 100 pixels.

Let's consider an image of size m by n pixels. There are $m \times n$ pixels, with each pixel being represented by 3 values (red, green, and blue), where each pixel may have a value in the range $[0, 255]$. We also want to describe each pixel by its location in the image – the x and y location. Therefore, each pixel will be represented by a 5-dimensional vector: $[\text{red}, \text{green}, \text{blue}, x, y]$. The five components allow k -means to identify that two pixels with similar colors that are close to one another should be part of the same group (cluster).

You should choose the number of clusters (k) that you think is appropriate for the image you choose.

After clustering, reconstruct the image in a way that shows the groups. For example, you might reconstruct a false-color image that shows one unique color for each group.

Discuss how this algorithm for image segmentation could be used for image compression, for computer vision, or other applications.

Your report should be written in \LaTeX . You should describe how you set up your experiments, make any relevant graphs to illustrate your findings, and analyze your results. Please also comment on what you learned. Finally, please provide the code you wrote as an appendix.