Dominic DiPasquale
CS1567 Introduction to Machine Learning
Professor Milos Hauskrecht

---

# Problem Assignment 9

## Problem 1

a) Starting with the means of (7,0) and (0,0) we will first calculate the distance from each data point to the given means:
The distance from (0, 0) to (0, 0) is 0.000000
The distance from (0, 0) to (7, 0) is 7.000000
The distance from (0, 5) to (0, 0) is 5.000000
The distance from (0, 5) to (7, 0) is 8.602325
The distance from (6, 7) to (0, 0) is 9.219544
The distance from (6, 7) to (7, 0) is 7.071068
The distance from (7, 0) to (0, 0) is 7.000000
The distance from (7, 0) to (7, 0) is 0.000000
With this information we can now start clustering the data:
Data point (0, 0), belongs to mean point (0, 0) group 1
Data point (0, 5), belongs to mean point (0, 0) group 1
Data point (6, 7), belongs to mean point (7, 0) group 2
Data point (7, 0), belongs to mean point (7, 0) group 2
And now we complete the process by re calculating the means:
New average point for group 1 is: (0.000000, 2.500000)
New average point for group 2 is: (6.500000, 3.500000)

b) We will repeat the process above, this time using starting means of (3,3) and (7,0)
The distance from (0, 0) to (3, 3) is 4.242641
The distance from (0, 0) to (7, 0) is 7.000000
The distance from (0, 5) to (3, 3) is 3.605551
The distance from (0, 5) to (7, 0) is 8.602325
The distance from (6, 7) to (3, 3) is 5.000000
The distance from (6, 7) to (7, 0) is 7.071068
The distance from (7, 0) to (3, 3) is 5.000000
The distance from (7, 0) to (7, 0) is 0.000000
Data point (0, 0), belongs to mean point (3, 3) group 1
Data point (0, 5), belongs to mean point (3, 3) group 1
Data point (6, 7), belongs to mean point (3, 3) group 1
Data point (7, 0), belongs to mean point (7, 0) group 2
New average point for group 1 is: (2.000000, 4.000000)
New average point for group 2 is: (7.000000, 0.000000)

c) To find the best clustering we should use the sum of all the distances from points to the mean of their cluster. This will punish poor groupings and reward better clustering of the data.

## Problem 2

a)

| Round | Cluster 1 | Cluster 2 |
|---|---|---|
| 1 | 2 | 178 |
| 2 | 1 | 179 |
| 3 | 1 | 179 |
| 4 | 174 | 6 |

| | | |
|---|---|---|
| 5 | 8 | 172 |
| 6 | 178 | 2 |
| 7 | 60 | 120 |
| 8 | 15 | 165 |
| 9 | 178 | 2 |
| 10 | 18 | 162 |
| 11 | 179 | 1 |
| 12 | 77 | 103 |
| 13 | 179 | 1 |
| 14 | 31 | 149 |
| 15 | 29 | 151 |
| 16 | 143 | 37 |
| 17 | 179 | 1 |
| 18 | 1 | 179 |
| 19 | 48 | 132 |
| 20 | 79 | 101 |
| 21 | 179 | 1 |
| 22 | 146 | 34 |
| 23 | 84 | 96 |
| 24 | 177 | 3 |
| 25 | 1 | 179 |
| 26 | 179 | 1 |
| 27 | 29 | 151 |
| 28 | 55 | 125 |
| 29 | 178 | 2 |
| 30 | 29 | 151 |

The best clustering is 101 and 79 with a sum of 8.8290e+03

b)

To generate the initial seeds I propose that we choose n different sample from the dataset for each cluster that we wish to make. The seeds will be a function of the average of the points. To begin I will choose a sample size of 5, for a total of 10 points. When using this strategy often the clusters conformed to a more even distribution (the number of points in both clusters was closer to equal than in the one where random seeds were chosen). For example a clustering of 98 and 82 was common with a sum around 8.8274e+03. The total sum was only marginally different.

c) To find out how well our class labels and clustering agrees we can simply check to see if the set of points with class label 1 matches with either the set of cluster 1 points or the set of cluster 2 points. Whichever has the higher amount of matches we assume is the correct cluster (since this can very). We can then sum and divide by the total number of points to see what percentage of points we place in the correct cluster. Using this method, and looking specifically at the clustering

for the minimum sum (i.e. what we were using to measure how well our clusters represented patterns in the data) I only got an agreement of ~50% over a number of trials. This leads me to believe that better clustering does not explain the class labels and thus the clustering and class labels do not agree.

d)  After running the kmeans on 1000 different random class assignments we can calculate that the chance that we got more correct with random variables than with the given class variables is around 70%. This is very high in infers that the variables we are given have no association with the clusters of data.