Clustering step by step

Roman Melnyk

Sunday, March 29, 2015

Dataset of 27 observations and 3 variables. There are 27 medicines, and it is known that they consist of two types. Our job, to figure out to which group belongs each medicine, using clustering.

```
##       x   y Species
## 1   1.0 1.0       A
## 2   2.0 1.0       B
## 3   1.5 2.0       C
## 4   4.5 2.5       D
## 5   3.5 2.5       F
## 6   4.5 2.7       G
## 7   3.5 3.5       K
## 8   4.0 2.7       P
## 9   2.8 1.2       M
## 10 1.0 2.5       N
## 11 2.5 2.1       L
## 12 1.5 2.5       K
## 13 2.5 1.6       V
## 14 2.0 1.8       J
## 15 1.0 1.5       H
## 16 1.5 1.5       S
## 17 1.5 1.5       I
## 18 2.0 1.3       R
## 19 4.0 3.0       C
## 20 4.5 3.5       W
## 21 4.5 3.5       Y
## 22 4.0 3.3       Q
## 23 4.0 4.0       O
## 24 4.0 2.5       I
## 25 4.5 3.1       V
## 26 5.0 4.0       Z
## 27 5.0 2.7       X
```

Step 1

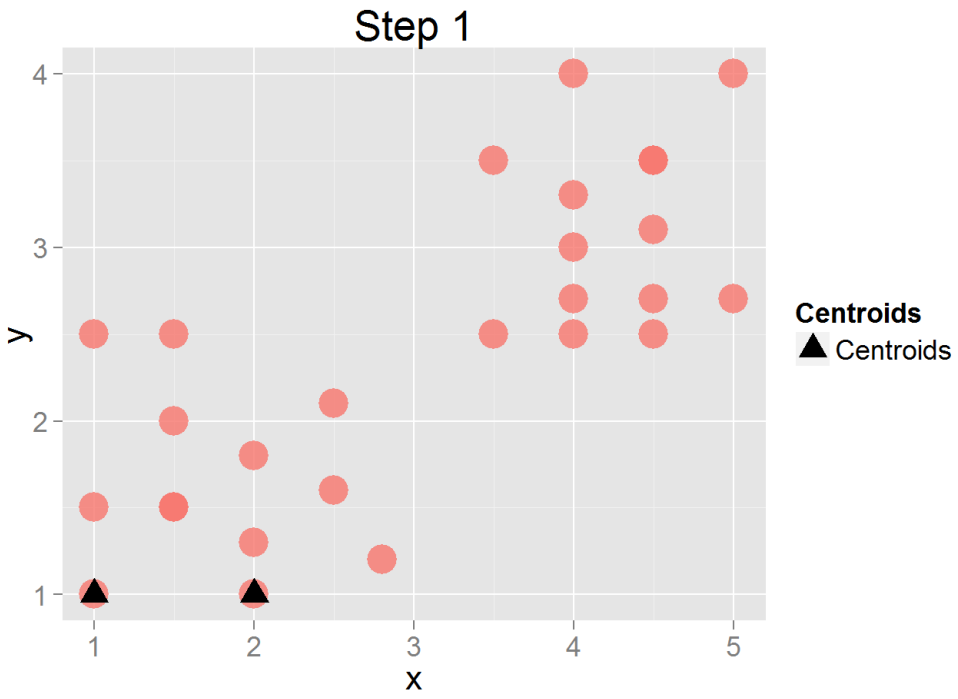
Slect centroids, and calculate distance from each point to each centroid

```
##       x   y Cent1 Cent2 species
## 1   1.0 1.0 0.000 1.000       A
## 2   2.0 1.0 1.000 0.000       B
## 3   1.5 2.0 1.118 1.118       C
## 4   4.5 2.5 3.808 2.915       D
## 5   3.5 2.5 2.915 2.121       F
## 6   4.5 2.7 3.891 3.023       G
## 7   3.5 3.5 3.536 2.915       K
## 8   4.0 2.7 3.448 2.625       P
## 9   2.8 1.2 1.811 0.825       M
## 10 1.0 2.5 1.500 1.803       N
## 11 2.5 2.1 1.860 1.208       L
## 12 1.5 2.5 1.581 1.581       K
## 13 2.5 1.6 1.616 0.781       V
## 14 2.0 1.8 1.281 0.800       J
## 15 1.0 1.5 0.500 1.118       H
## 16 1.5 1.5 0.707 0.707       S
## 17 1.5 1.5 0.707 0.707       I
## 18 2.0 1.3 1.044 0.300       R
## 19 4.0 3.0 3.606 2.828       C
## 20 4.5 3.5 4.301 3.536       W
## 21 4.5 3.5 4.301 3.536       Y
## 22 4.0 3.3 3.780 3.048       Q
## 23 4.0 4.0 4.243 3.606       O
```

```
## 24 4.0 2.5 3.354 2.500       I
## 25 4.5 3.1 4.082 3.265       V
## 26 5.0 4.0 5.000 4.243       Z
## 27 5.0 2.7 4.346 3.448       X
```

Assing each point to appropriate cluster.

```
##       x   y clust species
## 1   1.0 1.0     1       A
## 2   2.0 1.0     2       B
## 3   1.5 2.0     2       C
## 4   4.5 2.5     2       D
## 5   3.5 2.5     2       F
## 6   4.5 2.7     2       G
## 7   3.5 3.5     2       K
## 8   4.0 2.7     2       P
## 9   2.8 1.2     2       M
## 10 1.0 2.5     1       N
## 11 2.5 2.1     2       L
## 12 1.5 2.5     2       K
## 13 2.5 1.6     2       V
## 14 2.0 1.8     2       J
## 15 1.0 1.5     1       H
## 16 1.5 1.5     2       S
## 17 1.5 1.5     2       I
## 18 2.0 1.3     2       R
## 19 4.0 3.0     2       C
## 20 4.5 3.5     2       W
## 21 4.5 3.5     2       Y
## 22 4.0 3.3     2       Q
## 23 4.0 4.0     2       O
## 24 4.0 2.5     2       I
## 25 4.5 3.1     2       V
## 26 5.0 4.0     2       Z
## 27 5.0 2.7     2       X
```



Step 2

New Centroids

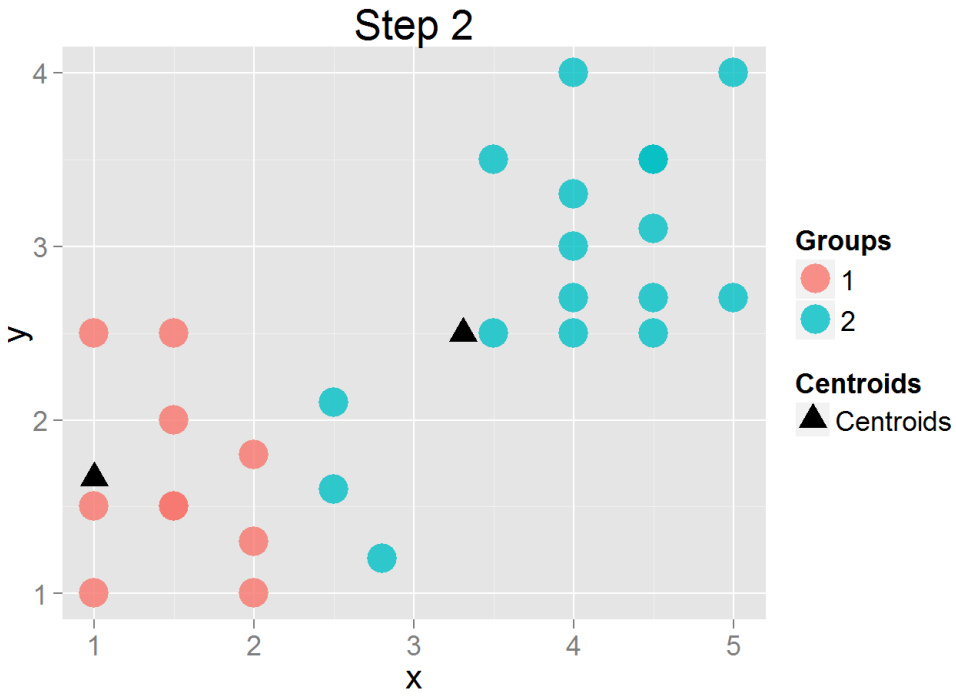
Distance for new centroids

```
##       x   y Cent21 Cent22 species
## 1   1.0 1.0  0.667  2.749       A
## 2   2.0 1.0  1.202  1.988       B
```

```
## 3  1.5 2.0  0.601  1.872       C
## 4  4.5 2.5  3.598  1.196       D
## 5  3.5 2.5  2.635  0.196       F
## 6  4.5 2.7  3.649  1.212       G
## 7  3.5 3.5  3.100  1.019       K
## 8  4.0 2.7  3.173  0.724       P
## 9  2.8 1.2  1.860  1.394       M
## 10 1.0 2.5  0.833  2.304       N
## 11 2.5 2.1  1.561  0.898       L
## 12 1.5 2.5  0.972  1.804       K
## 13 2.5 1.6  1.501  1.207       V
## 14 2.0 1.8  1.009  1.480       J
## 15 1.0 1.5  0.167  2.512       H
## 16 1.5 1.5  0.527  2.063       S
## 17 1.5 1.5  0.527  2.063       I
## 18 2.0 1.3  1.065  1.772       R
## 19 4.0 3.0  3.283  0.857       C
## 20 4.5 3.5  3.951  1.559       W
## 21 4.5 3.5  3.951  1.559       Y
## 22 4.0 3.3  3.416  1.060       Q
## 23 4.0 4.0  3.801  1.654       O
## 24 4.0 2.5  3.114  0.696       I
## 25 4.5 3.1  3.782  1.338       V
## 26 5.0 4.0  4.631  2.264       Z
## 27 5.0 2.7  4.131  1.708       X
```

Recalculate clusters. Points can change their previous cluster.

```
##      x   y clust species
## 1  1.0 1.0     1       A
## 2  2.0 1.0     1       B
## 3  1.5 2.0     1       C
## 4  4.5 2.5     2       D
## 5  3.5 2.5     2       F
## 6  4.5 2.7     2       G
## 7  3.5 3.5     2       K
## 8  4.0 2.7     2       P
## 9  2.8 1.2     2       M
## 10 1.0 2.5     1       N
## 11 2.5 2.1     2       L
## 12 1.5 2.5     1       K
## 13 2.5 1.6     2       V
## 14 2.0 1.8     1       J
## 15 1.0 1.5     1       H
## 16 1.5 1.5     1       S
## 17 1.5 1.5     1       I
## 18 2.0 1.3     1       R
## 19 4.0 3.0     2       C
## 20 4.5 3.5     2       W
## 21 4.5 3.5     2       Y
## 22 4.0 3.3     2       Q
## 23 4.0 4.0     2       O
## 24 4.0 2.5     2       I
## 25 4.5 3.1     2       V
## 26 5.0 4.0     2       Z
## 27 5.0 2.7     2       X
```



Step 3

New centroids

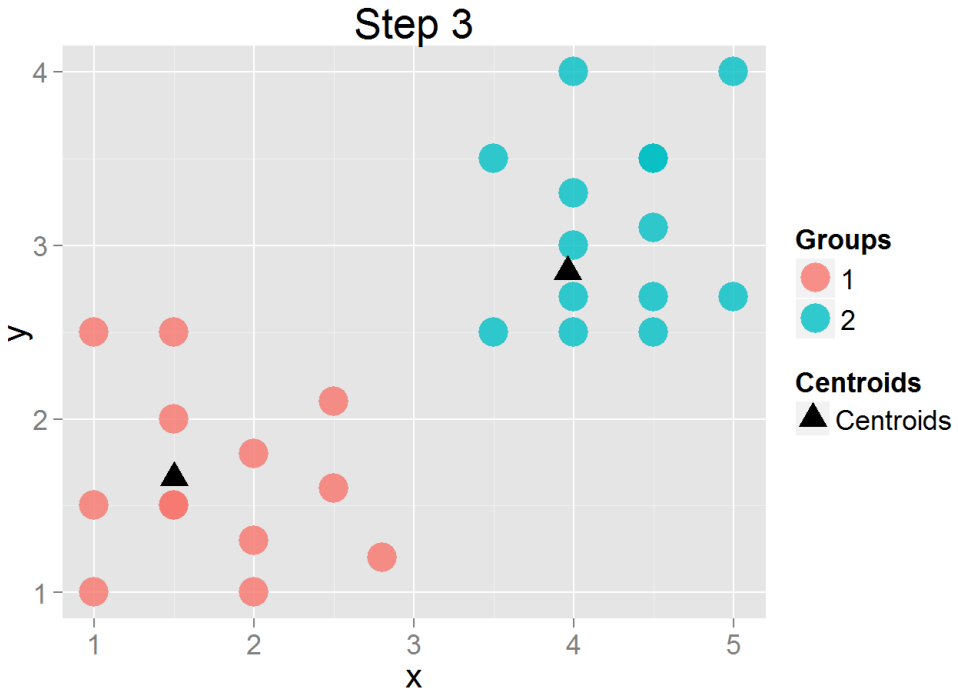
Distance for new centroids

```
##    Cent31 Cent32 species
## 1   0.828  3.488       A
## 2   0.828  2.692       B
## 3   0.340  2.601       C
## 4   3.115  0.643       D
## 5   2.169  0.575       F
## 6   3.175  0.561       G
## 7   2.718  0.798       K
## 8   2.708  0.153       P
## 9   1.379  2.014       M
## 10  0.978  2.979       N
## 11  1.093  1.639       L
## 12  0.840  2.483       K
## 13  1.002  1.919       V
## 14  0.519  2.221       J
## 15  0.525  3.251       H
## 16  0.160  2.804       S
## 17  0.160  2.804       I
## 18  0.616  2.496       R
## 19  2.836  0.158       C
## 20  3.519  0.848       W
## 21  3.519  0.848       Y
## 22  2.990  0.455       Q
## 23  3.424  1.154       O
## 24  2.637  0.349       I
## 25  3.328  0.597       V
## 26  4.210  1.553       Z
## 27  3.651  1.052       X
```

Repeate everything as in previous steps.

```
##      x   y clust species
## 1  1.0 1.0     1       A
## 2  2.0 1.0     1       B
## 3  1.5 2.0     1       C
## 4  4.5 2.5     2       D
## 5  3.5 2.5     2       F
## 6  4.5 2.7     2       G
## 7  3.5 3.5     2       K
```

```
## 8  4.0 2.7     2        P
## 9  2.8 1.2     1        M
## 10 1.0 2.5     1        N
## 11 2.5 2.1     1        L
## 12 1.5 2.5     1        K
## 13 2.5 1.6     1        V
## 14 2.0 1.8     1        J
## 15 1.0 1.5     1        H
## 16 1.5 1.5     1        S
## 17 1.5 1.5     1        I
## 18 2.0 1.3     1        R
## 19 4.0 3.0     2        C
## 20 4.5 3.5     2        W
## 21 4.5 3.5     2        Y
## 22 4.0 3.3     2        Q
## 23 4.0 4.0     2        O
## 24 4.0 2.5     2        I
## 25 4.5 3.1     2        V
## 26 5.0 4.0     2        Z
## 27 5.0 2.7     2        X
```



Step 4

New centroids

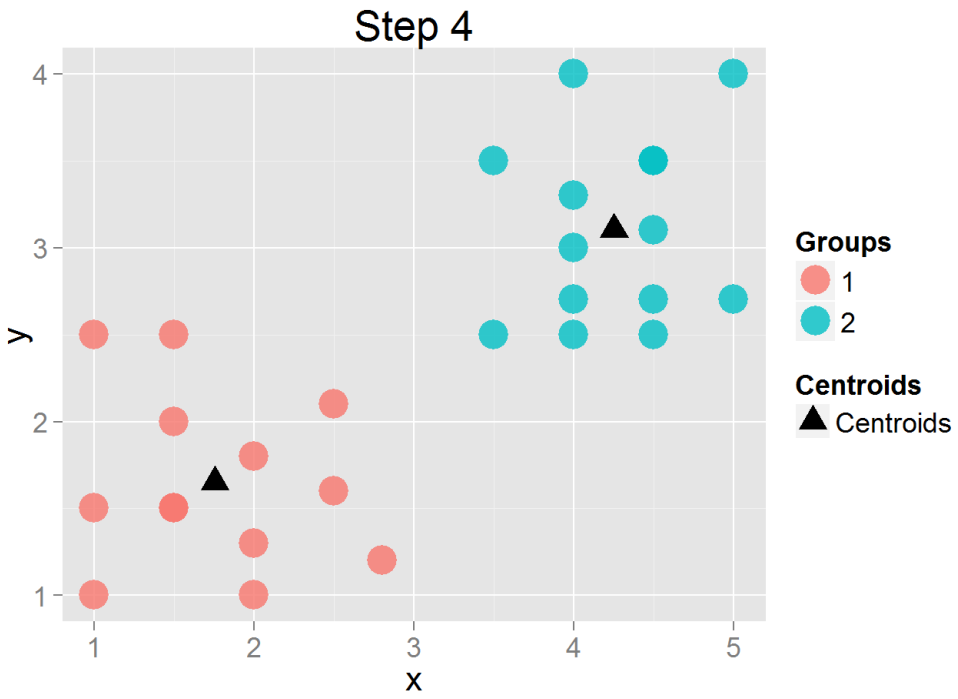
Distance for new centroids

```
##    Cent41 Cent42 species
## 1   0.998  3.873       A
## 2   0.699  3.083       B
## 3   0.429  2.965       C
## 4   2.874  0.657       D
## 5   1.940  0.965       F
## 6   2.939  0.478       G
## 7   2.541  0.847       K
## 8   2.478  0.478       P
## 9   1.140  2.396       M
## 10  1.133  3.306       N
## 11  0.869  2.019       L
## 12  0.883  2.816       K
## 13  0.748  2.310       V
## 14  0.286  2.602       J
## 15  0.769  3.626       H
## 16  0.297  3.185       S
## 17  0.297  3.185       I
## 18  0.431  2.886       R
```

```
## 19  2.619  0.272       C
## 20  3.309  0.466       W
## 21  3.309  0.466       Y
## 22  2.785  0.316       Q
## 23  3.248  0.927       O
## 24  2.400  0.657       I
## 25  3.104  0.250       V
## 26  4.005  1.166       Z
## 27  3.411  0.853       X
```

Fix points in new clusters.

```
##       x   y clust species
## 1   1.0 1.0     1       A
## 2   2.0 1.0     1       B
## 3   1.5 2.0     1       C
## 4   4.5 2.5     2       D
## 5   3.5 2.5     2       F
## 6   4.5 2.7     2       G
## 7   3.5 3.5     2       K
## 8   4.0 2.7     2       P
## 9   2.8 1.2     1       M
## 10 1.0 2.5     1       N
## 11 2.5 2.1     1       L
## 12 1.5 2.5     1       K
## 13 2.5 1.6     1       V
## 14 2.0 1.8     1       J
## 15 1.0 1.5     1       H
## 16 1.5 1.5     1       S
## 17 1.5 1.5     1       I
## 18 2.0 1.3     1       R
## 19 4.0 3.0     2       C
## 20 4.5 3.5     2       W
## 21 4.5 3.5     2       Y
## 22 4.0 3.3     2       Q
## 23 4.0 4.0     2       O
## 24 4.0 2.5     2       I
## 25 4.5 3.1     2       V
## 26 5.0 4.0     2       Z
## 27 5.0 2.7     2       X
```



There is no changes on a plot(centroids remain the same), that's why process stooped. Step 4 is a answer.