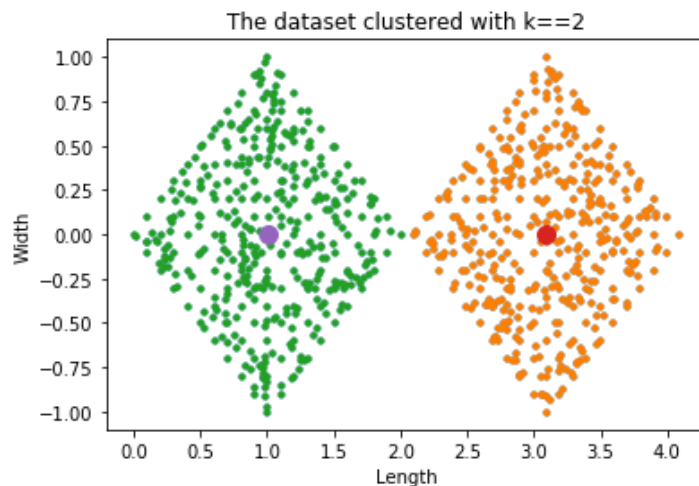


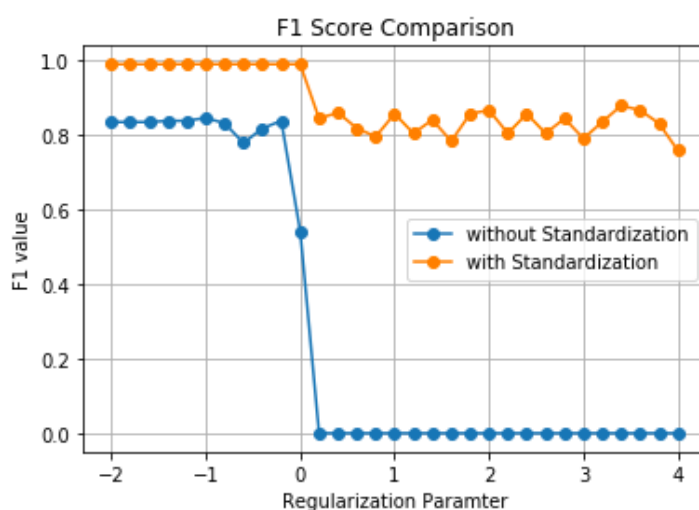
Problem 1

K is set to 2. At the 1st iteration, two random points are chosen as cluster centers. And using Euclidean distance, for every cluster, the closer points are assigned to that cluster. The cluster center is updated after assignment in each iteration. The result is as follows.



Problem 2

When standardization is not used. The scale of each column in the input data can be very huge. For x larger than 30, sigmoid function starts to gives 1.0 as output, which leads to gradient vanishing. We can see without standardization, logistic regression cannot continue for this dataset. When using standardization, logistic regression works better without regularization for this dataset.



Problem 3

confusion matrix of k_means is:

```
[[176  0  0  0  2  0  0  0  0  0]
 [  0 128  0  0  0  1  2  0 26 25]
 [  1 10 108  2  0  0  0  3  0 53]
 [  0  8  0 156  0  1  0  7  9  2]
 [  0  2  0  0 168  0  0 11  0  0]
 [  0  0  0  1  2 137  2  0 40  0]
 [  1  4  0  0  0  0 176  0  0  0]
 [  0  2  1  0  0  0  0 176  0  0]
 [  0 109  3  9  0  5  2  4 42  0]
 [  0  2  0  5  3  6  0  8 155 11]]
```

confusion matrix of Agglomerative Clustering is:

```
[[178  0  0  0  0  0  0  0  0  0]
 [  0 59 27  0  0  0  0  0  0 96]
 [  0  0 166  0  0  0  0  1 10  0]
 [  0  0  0 169  0  0  0  1 13  0]
 [  0  0  0  0 178  0  0  3  0  0]
 [  0  0  0  2  0 179  1  0  0  0]
 [  0  0  0  0  0  0 180  0  1  0]
 [  0  0  0  0  0  0  0 179  0  0]
 [  0  1  4  1  0  0  0  1 165  2]
 [  0 20  0 145  0  2  0 11  2  0]]
```

confusion matrix of Affinity Propagation Clustering is:

```
[[177  0  0  0  1  0  0  0  0  0]
 [  0 182  0  0  0  0  0  0  0  0]
 [  0  0 176  1  0  0  0  0  0  0]
 [  0  0  0 177  0  2  0  0  0  4]
 [  0  4  0  0 171  0  0  4  0  2]
 [  0  0  0  0  0 180  0  0  0  2]
 [  2  1  0  0  0  0 178  0  0  0]
 [  0  0  0  0  0  0  0 178  0  1]
 [  0 16  2  3  0  1  1  1 149  1]
 [  0  1  0  0  0  3  0  1  1 174]]
```

The Fowlkes Mallows Scores of these 3 methods are:

K-means: 0.683328021337

Agglomerative Clustering: 0.816751686074

Affinity Propagation: 0.94047760691

Problem 4

F measure on training set:

Linear SVC = 1.0

RBF SVC = 0.994974874372

RandomForest = 1.0

F measure on test set:

Linear SVC = 1.0

RBF SVC = 1.0

RandomForest = 1.0