Homework 8

Kaizhao Liang(kl2)

Part A. **SS_totals:**

k \ iteration	5	10	20
3	[82.4, 78.9, 78.9, 79.6]	[143.5, 78.9, 78.9, 78.9]	[78.9, 78.9, 142.9, 78.9]
6	[73.1, 42.9, 39.2, 42.4]	[48.9, 48.1, 38.9, 39.2]	[45.7, 41.8, 41.8, 47.6]
12	[29.4, 24.9, 24.5, 26.1]	[23.4, 35.1, 26.2, 28.8]	[26.1, 27.4, 31.2, 24.9]

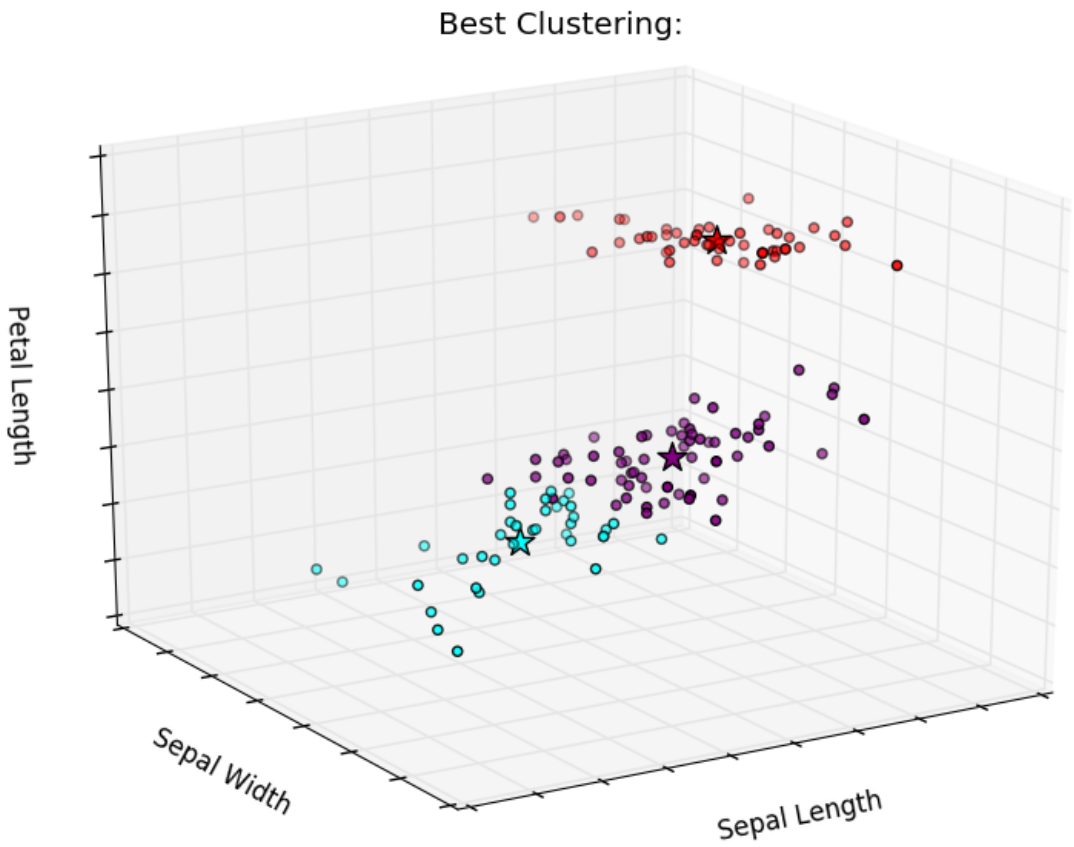
Explanation:
The SS_total decreases as the k increases, but only fluctuates a little as the number of iterations increases. The first trend is due to more clusters formed. It normally decrease the size of each cluster, the area the cluster encompasses, not the number of samples within the cluster, thus decreasing the overall SS_total. The second trend might be due to the fast convergence to the local optimum. Thus the increase of iterations makes little difference in the SS_total.
We should not choose K based on the SS_total. Because from the above experiment, we can see normally the larger the k is, the lower the SS_total. Yet it does not necessarily reveal the real clustering patterns and might give us unintended results.

Part B. **F1_scores:**for each clustering, three of the clusters that have highest number of one particular species are choosen. First it finds the clusters with the most number of each species and normally they are different. In the rare case, one cluster might have two labels and then it will be labeled by the majority. If two labels have same number of samples in the cluster, random tie breaking will be performed. To guarantee not to select the same cluster again, it pop the cluster out of the queue everytime one is choosen. After choosing the three distinct clusters, it computes the F1_scores.

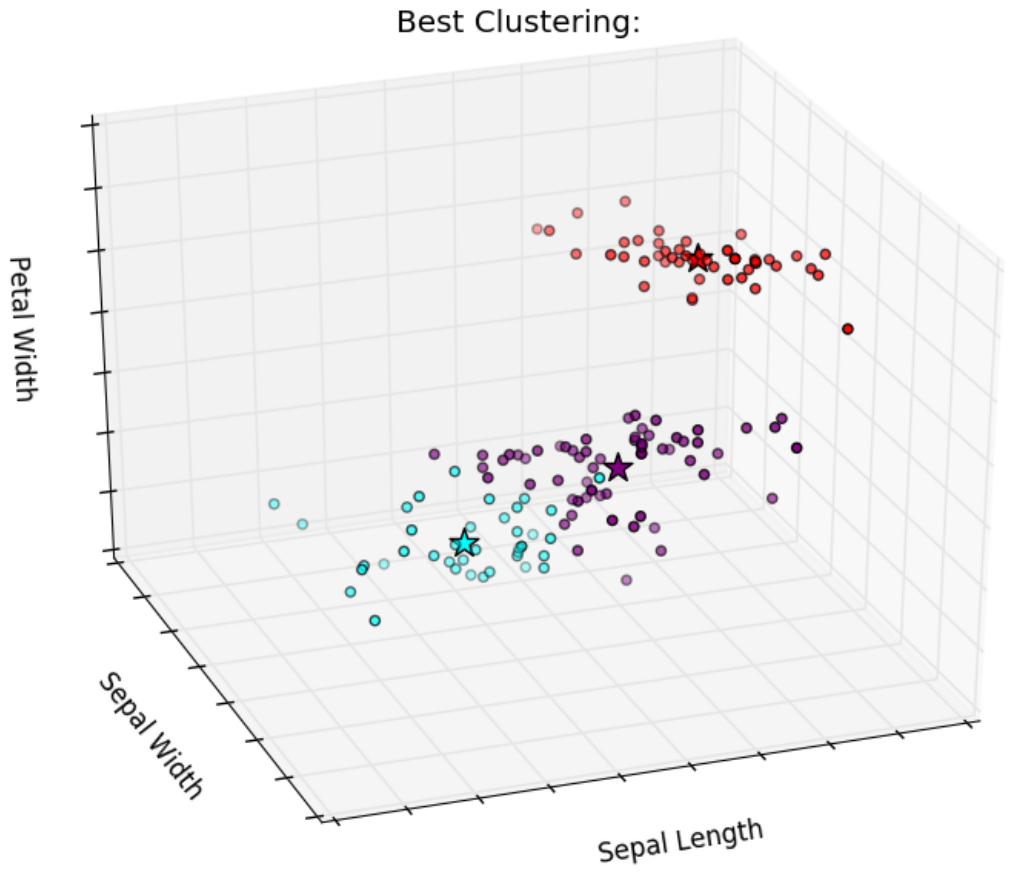
k \ iteration	5	10	20
3	Iris-setosa:1.00 Iris-versicolor:0.86 Iris-virginica:0.82 average:0.89	Iris-setosa:1.00 Iris-versicolor:0.85 Iris-virginica:0.81 average:0.89	Iris-setosa:1.00 Iris-versicolor:0.86 Iris-virginica:0.82 average:0.89
6	Iris-setosa:0.68 Iris-versicolor:0.58 Iris-virginica:0.67 average:0.64	Iris-setosa:0.70 Iris-versicolor:0.58 Iris-virginica:0.65 average:0.64	Iris-setosa:1.00 Iris-versicolor:0.68 Iris-virginica:0.61 average:0.76
12	Iris-setosa:0.68 Iris-versicolor:0.36 Iris-virginica:0.65 average:0.56	Iris-setosa:0.57 Iris-versicolor:0.53 Iris-virginica:0.39 average:0.50	Iris-setosa:0.59 Iris-versicolor:0.43 Iris-virginica:0.61 average:0.54

Explanation:
The average F1 score drops as the number of k increases. It is understandable given the expression of the F1 score. The larger the k is, the more possibly the samples spread in different clusters, thus lowering the recall. At the same time, the percision might drop, since more centroids than needed add to the confusion.
Increasing the iterations will generally increase the F1 score. It's obvious when k=6. More iterations will give chances for the unnecessary centroids to vanish. However, if the k is already optimized, increasing the iteration does not seem to make a difference. It's because it's trapped at local minimun. When k=12, the increase of iteration does not change the F1 score much, because number of iteration for it to converge may be way larger than given.
The best average F1 score is 0.89. When k=3 and iteration=5, it yeilds the best average F1 socre.
The centroids for the clustering with the highest F1 score:
Iris-versicolor:[5.9016129 , 2.7483871 , 4.39354839, 1.43387097]
Iris-setosa:[5.006, 3.418, 1.464, 0.244]
Iris-virginica:[6.85, 3.07368421, 5.74210526, 2.07105263]

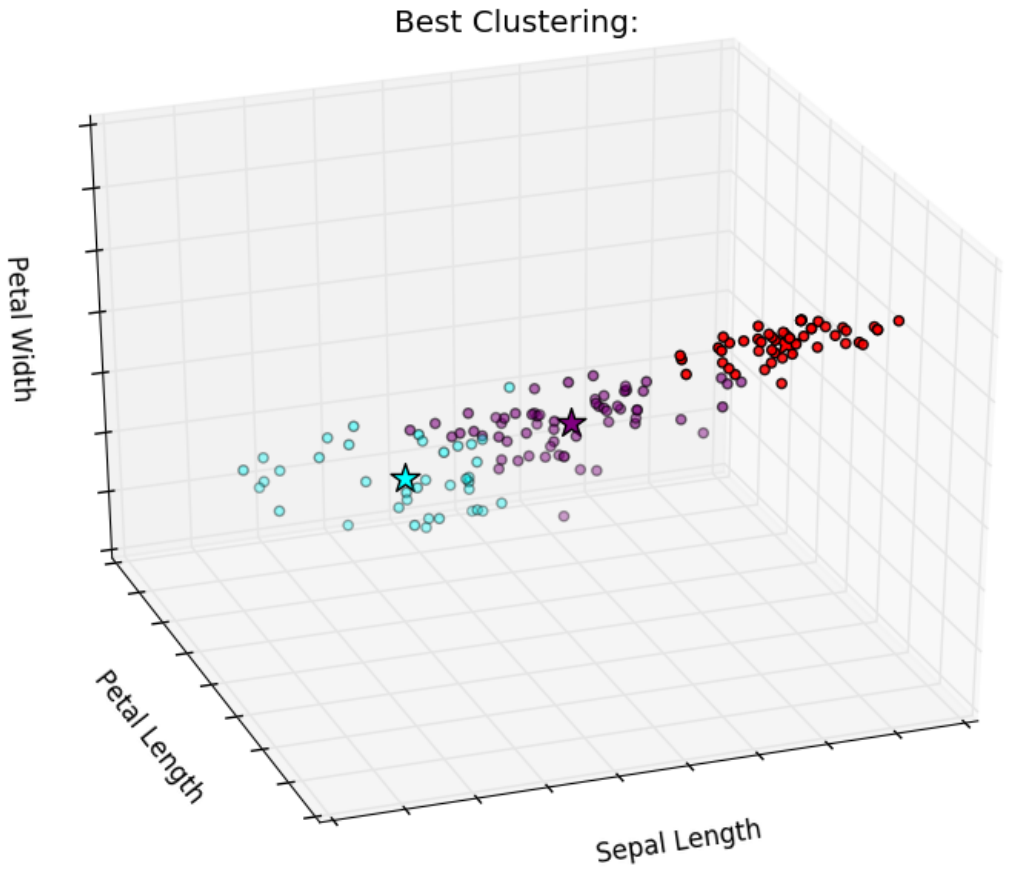
a. Sepal Length vs Sepal Width vs Petal Length



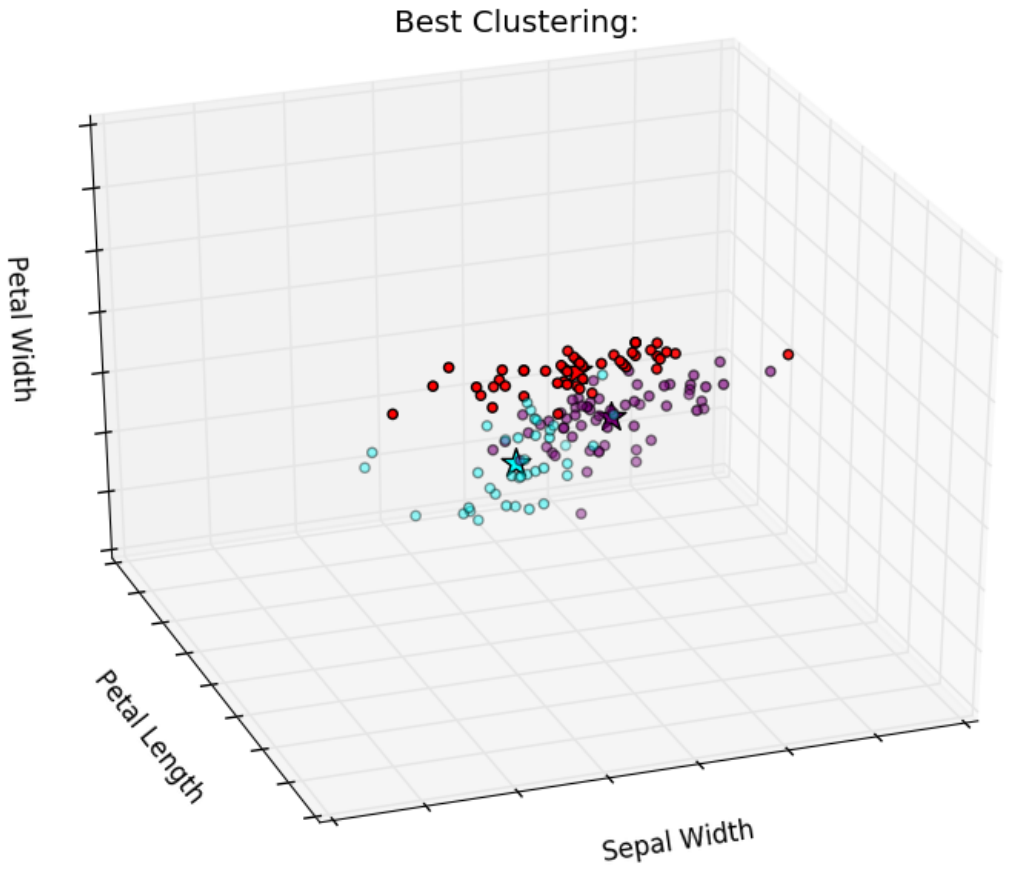
b. Sepal Length vs Sepal Width vs Petal Width



c. Sepal Length vs Petal Length vs Petal Width



d. Sepal Width vs Petal Length vs Petal Width


Best Clustering:
Petal Width
Petal Length
Sepal Width