Rutul Mehta - A20476293

In [23]:
```python
import math
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

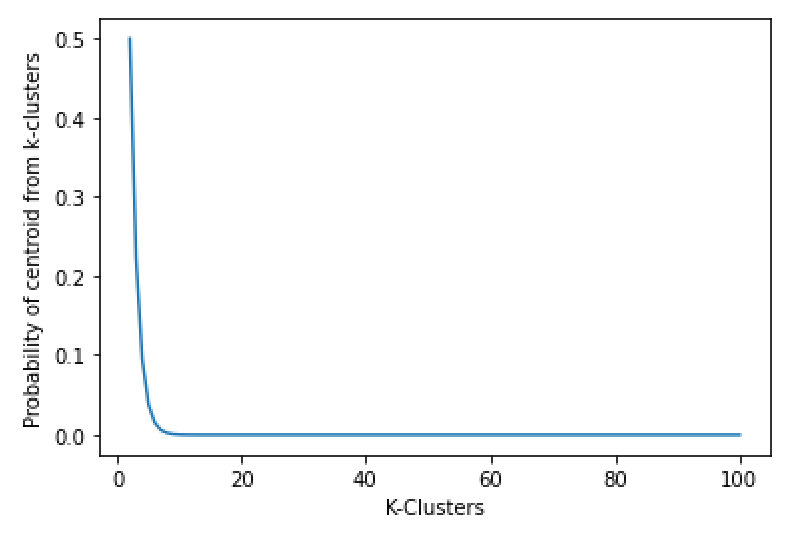
Exercise 4

a)

In [30]:
```python
def init_centroid(k_clusters=1, n=10):
    return math.factorial(k_clusters) * (n**k_clusters) / ((k_clusters*n) ** k_clusters)

results = pd.DataFrame([(k, init_centroid(k_clusters=k)) for k in range(2, 101)])
results.columns = ["K_Clusters", "Probability"]
plt.xlabel("K-Clusters")
plt.ylabel("Probability of centroid from k-clusters")
plt.plot(results["K_Clusters"], results["Probability"])
```

Out[30]:
```
[<matplotlib.lines.Line2D at 0x28e858a75e0>]
```



b)

In [31]:
```python
def init_centroid_mul_with_2(k_clusters=1, n=10):
    return 2*(math.factorial(k_clusters)) * (n**k_clusters) / ((k_clusters*n) ** k_clusters)

k_sizes = [10, 100, 1000]
probabilities = [(k, init_centroid_mul_with_2(k_clusters=k, n=2*k)) for k in k_sizes]
print(probabilities)
```

```
[(10, 0.00072576), (100, 1.866524308878883e-42), (1000, 0.0)]
```

Exercise 7

Answer is C - More centroids should be allocated to the denser region.

In the case of (c), a higher proportion of points will have lower squared errors in the dense regions and should thus minimize the SSE.

Exercise 11

What does it mean if the SSE for one variable is low for all clusters?

- If the SSE for any one variable is low for every cluster, than the variable is considered as a constant and contributes nothing in dividing the data into groups.

Low for just one cluster?

- If the SSE is low for only ONE cluster, than it would be helpful in defining the attribute of the cluster.

High for all clusters?

- High SSE for all clusters, than it means that it is noise or outliers, and has no affect on the resulting clusters.

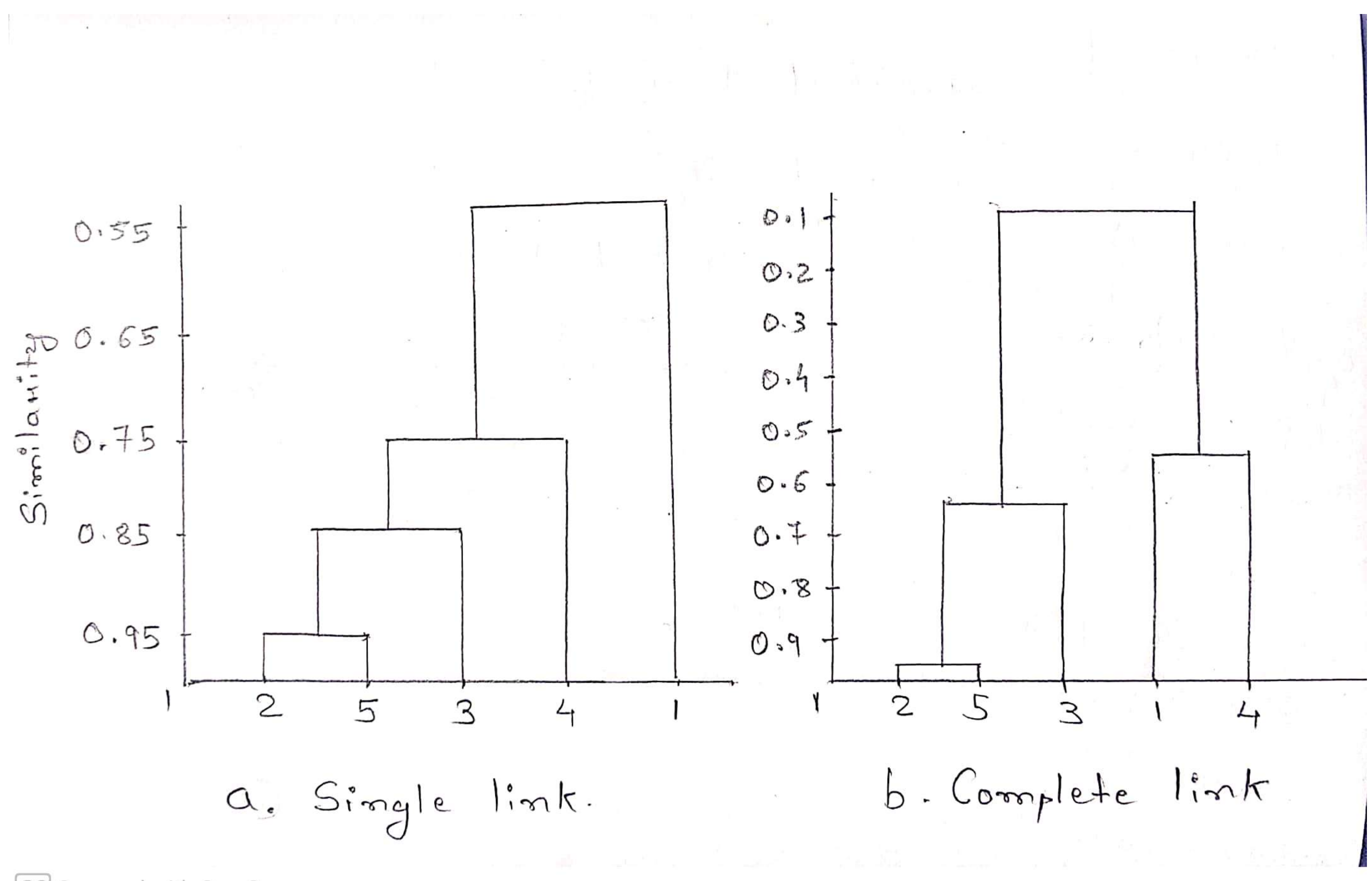
High for one cluster?

- High for one cluster means, than it would not be helpful in defining the cluster.

How could you use the per variable SSE info to improve your clustering?

- It can help in deciding which attributes to eliminate. As mentioned in the chapter 7, how sampling the data before clustering could be useful to eliminate the noise or outliers within the data, which would be useful to conserve time of the computation.

Exercise 16



Exercise 17

a)

i. {18, 45}

- First cluster is 6,12,18,24,30
- Error = 360
- Second Cluster = 42,48
- Error = 18
- Total = 360 + 18 = 378

ii. {15, 40}

- First cluster is 6,12,18,24
- Error = 180
- Second Cluster = 30,42,48
- Error = 168
- Total = 180 + 168 = 348

b)

Yes, they do represent stable solutions.

- If we run K-Means on either part, we can find the initial cluster with the new centroid and that is identical with the original centroid.

c)

The two clusters formed by a single link is {6, 12, 18, 24, 30} & {42, 48}.

d)

By "most natural clustering" im going to assume most seperated centroids. The two clusters are:

- K-Means -> [6,12,18,24] , [30,42,48] -> distnace between the centroids is 25
- Single-Link -> [6,12, 18, 24, 30], [42,48] -> distance betwween the centroid is 27

So single-link provides more natural clusters

e)

- MIN produces contiguous clusters.
- However, density is also possible.
- Center-based is also possible, since one set of centers gives the desired clusters.

f)

The K-means algorithm is weak towards finding clusters that have a variety in sizes, or when not well-separated. The objective of minimizing squared error leads it to breaking the larger cluster. Thus, producing the unnatural one in this case.

Exercise 21

Compute the entropy and purity for the confusion matrix.

Cluster #1:

Entropy = $-[(\frac{1}{693})\log(\frac{1}{693}) + (\frac{1}{693})\log(\frac{1}{693}) + (\frac{0}{693})\log(\frac{0}{693}) + (\frac{11}{693})\log(\frac{11}{693}) + (\frac{4}{693})\log(\frac{4}{693}) + (\frac{676}{693})\log(\frac{676}{693})] = 0.199 = 0.2$

Purity = $\frac{676}{693} = 0.975 = 0.98$

Cluster #2 :

Entropy = $-[(\frac{27}{1562})\log(\frac{27}{1562}) + (\frac{89}{1562})\log(\frac{89}{1562}) + (\frac{333}{1562})\log(\frac{333}{1562}) + (\frac{827}{1562})\log(\frac{827}{1562}) + (\frac{253}{1562})\log(\frac{253}{1562}) + (\frac{33}{1562})\log(\frac{33}{1562})] = 1.84$

Purity = $\frac{827}{1562} = 0.529 = 0.53$

Cluster #3:

Entropy = $-[(\frac{326}{949})\log(\frac{326}{949}) + (\frac{465}{949})\log(\frac{465}{949}) + (\frac{8}{949})\log(\frac{8}{949}) + (\frac{105}{949})\log(\frac{105}{949}) + (\frac{16}{949})\log(\frac{16}{949}) + (\frac{29}{949})\log(\frac{29}{949})] = 1.70$

Purity = $\frac{465}{949} = 0.49$

Total:

Entropy = $(0.2\frac{693}{3204}) + (1.84\frac{1562}{3204}) + (1.70\frac{949}{3204}) = 1.44$

Purity = $\frac{676+827+465}{3204} = 0.61$

Exercise 22

Given 2 sets of 100 points that fall within the unit square. One set of points is arranged so that the points are uniormly spaced. The other set of points is generated from a uniform distribution over the unit square.

a) Is there a difference between the 2 set of points?

- Definitely, the random points will have a region of less & more density, while the uniformly spaced will have uniform density throughout the unit square.

b) If so, which set of points will typically have a smaller SSE for K=10 clusters?

- The random set of points will have a lower SSE.

c) What will be the behavior of DBSCAN on the uniform data set? The random data set?

- Depending on the threshold, DBSCAN will either merge all points in the uniform data set into one cluster or classify them all as noise. In terms of the random data set, DBSCAN can often find clusters in random data due to the variety of density between regions.