# Homework #3

Assign date: 2015-06-21

**Due date: 2015-07-01, 6pm**

Submission:

1. **Please submit your results in email to the grader:** [**130301039@svuca.edu**](mailto:130301039@svuca.edu) **Chi Zhang**
2. **Please separate the written answers from the python code: you should submit 2 files in your email – cs596-29-hw2\_yourID#.doc & cs596-29-hw2\_yourID#.py**
3. **30 pts per day will be deducted for late submission**

**Problem 1)** K-means clustering (20 pts.)

You are given the following 10 data points of height & weight:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Height | 66 | 73 | 72 | 70 | 74 | 68 | 65 | 64 | 63 | 67 |
| Weight | 170 | 210 | 165 | 180 | 185 | 155 | 150 | 120 | 125 | 140 |

Manually apply k-means algorithms to get 2 clusters. Please produce the center and grouping step by step, using the following parameters:

1. Initialize with ID=1 and ID=2
2. Assume Euclidean distance

**Solution**

The distance matrix based on Euclidean distance between the 10 data points (ID=x), calculated base on this formula: **d(a,b)=sqrt((xb-xa)2+(yb-ya)2))**

* *d(a,b) denotes the Euclidean distance between data point* ***a*** *and* ***b****. It is obtained from the distance matrix*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ID=1 | ID=2 | ID=3 | ID=4 | ID=5 | ID=6 | ID=7 | ID=8 | ID=9 | ID=10 |
| ID=1 | 0.00 | 40.61 | 7.81 | 10.77 | 17.00 | 15.13 | 20.02 | 50.04 | 45.10 | 30.02 |
| ID=2 |  | 0.00 | 45.01 | 30.15 | 25.02 | 55.23 | 60.53 | 90.45 | 85.59 | 70.26 |
| ID=3 |  |  | 0.00 | 15.13 | 20.10 | 10.77 | 16.55 | 45.71 | 41.00 | 25.50 |
| ID=4 |  |  |  | 0.00 | 6.40 | 25.08 | 30.41 | 60.30 | 55.44 | 40.11 |
| ID=5 |  |  |  |  | 0.00 | 30.59 | 36.14 | 65.76 | 61.00 | 45.54 |
| ID=6 |  |  |  |  |  | 0.00 | 5.83 | 35.23 | 30.41 | 15.03 |
| ID=7 |  |  |  |  |  |  | 0.00 | 30.02 | 25.08 | 10.20 |
| ID=8 |  |  |  |  |  |  |  | 0.00 | 5.10 | 20.22 |
| ID=9 |  |  |  |  |  |  |  |  | 0.00 | 15.52 |
| ID=10 |  |  |  |  |  |  |  |  |  | 0.00 |

**Epoch 1**

We start with 2 seeds:

* seed C1 = ID1 = (66,170)
* seed C2 = ID2 = (73,210)

and traverse from column ID=3 to column ID=10 to review the distance between each node from seed C1 and seed C2 (row 1, 2 below). We will mark the the smaller distance yellow.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ID=1 | ID=2 | ID=3 | ID=4 | ID=5 | ID=6 | ID=7 | ID=8 | ID=9 | ID=10 |
| ID=1 | 0.00 | 40.61 | 7.81 | 10.77 | 17.00 | 15.13 | 20.02 | 50.04 | 45.10 | 30.02 |
| ID=2 |  | 0.00 | 45.01 | 30.15 | 25.02 | 55.23 | 60.53 | 90.45 | 85.59 | 70.26 |

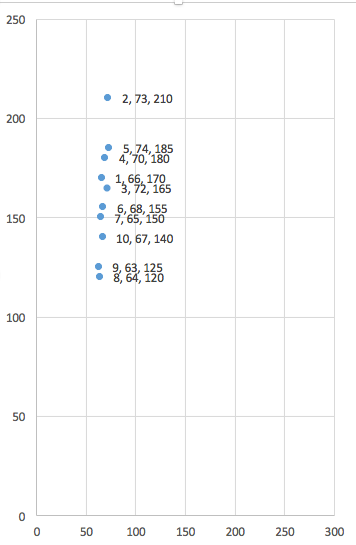


Figure -Epoch 1 Plot

Base on above result:

* Node {3, 4, 5, 6, 7, 8, 9, 10} ∈ Cluster 1 – seed C1
* Node {2} ∈ Cluster 2 – seed C2

Centers of the new clusters:

C1 = Average(ID1, ID3, ID4, ID5, ID6, ID7, ID8, ID9, ID10) = (67.67, 154.44)

C2 = (ID2) = (73, 210)

**Epoch 2**

Repeat the same steps with new seeds  
C1 = (66.67,154.44), C2 = (73, 210)

Cluster 1 = Node {1, 3,4,6,7,8,9,10}, new center = (66.875,150.625)

Cluster 2 = Node {2,5}, new center = (73.5,197.5)

**Epoch 3**

Repeat the same steps with new seeds  
C1 = (66.875,150.625), C2 = (73.5,197.5)

Cluster 1 = Node {1, 3, 6, 7, 8, 9, 10} new center = (66.428, 146.428)

Cluster 2 = Node {2,4,5} new center = (72.33,191.67)

**Epoch 4**

Repeat the same steps with new seeds  
C1 = (66.428,146.428), C2 = (72.33,191.67)

Cluster 1 = Node {3, 6, 7, 8, 9, 10} new center = (66.6, 142.5)

Cluster 2 = Node {1,2,4,5} new center = (70.75,186.25)

**Epoch 5 - final**

Repeat the same steps with new seeds  
C1 = (66.6, 142.5), C2 = (70.75,186.25)

Cluster 1 = Node {6, 7, 8, 9, 10} new center = (65.4, 138)

Cluster 2 = Node {1, 2, 3, 4, 5} new center = (71,182)

When we repeat these steps for **Epoch 6**, the Cluster members do not change anymore, so we know Epoch 5 has the final result with 2 clusters:

* **Cluster 1: Node {6, 7, 8, 9, 10}**
* **Cluster 2: Node {1, 2, 3, 4, 5}**

**Problem 2)** Agglomerative clustering (20 pts.)

Using the same 10 data points of height & weight as in Problem-1, apply agglomerative clustering manually. Produce the distance matrix and the resulting dendogram. Use the following assumptions:

1. Euclidean distance
2. Single linkage for cluster dissimilarity

**Solution**

**Agglomerative**: initially every point is a cluster of its own and we merge cluster until we end-up with one unique cluster containing all points.

**Single link:** distance between two clusters is the shortest distance between a pair of elements from the two clusters.

**For each k – we recalculate the distance matrix, start with k=10, i.e. each point is a cluster.**

k=10

+------+------+------+------+------+------+------+------+------+------+

| 0 | 40.6 | 7.81 | 10.7 | 17.0 | 15.1 | 20.0 | 50.0 | 45.0 | 30.0 |

| | 0788 | 0249 | 7032 | | 3274 | 2498 | 3998 | 9988 | 1666 |

| | 1008 | 6759 | 9614 | | 5950 | 4394 | 4012 | 9135 | 2039 |

| | 5 | 1 | 3 | | 4 | 5 | 8 | 1 | 6 |

+======+======+======+======+======+======+======+======+======+======+

| 40.6 | 0 | 45.0 | 30.1 | 25.0 | 55.2 | 60.5 | 90.4 | 85.5 | 70.2 |

| 08 | | 11 | 50 | 20 | 27 | 31 | 49 | 86 | 57 |

+------+------+------+------+------+------+------+------+------+------+

| 7.81 | 45.0 | 0 | 15.1 | 20.1 | 10.7 | 16.5 | 45.7 | 41 | 25.4 |

| 0 | 11 | | 33 | 00 | 70 | 53 | 06 | | 95 |

+------+------+------+------+------+------+------+------+------+------+

| 10.7 | 30.1 | 15.1 | 0 | 6.40 | 25.0 | 30.4 | 60.2 | 55.4 | 40.1 |

| 70 | 50 | 33 | | 3 | 80 | 14 | 99 | 44 | 12 |

+------+------+------+------+------+------+------+------+------+------+

| 17 | 25.0 | 20.1 | 6.40 | 0 | 30.5 | 36.1 | 65.7 | 61 | 45.5 |

| | 20 | 00 | 3 | | 94 | 39 | 65 | | 41 |

+------+------+------+------+------+------+------+------+------+------+

| 15.1 | 55.2 | 10.7 | 25.0 | 30.5 | 0 | 5.83 | 35.2 | 30.4 | 15.0 |

| 33 | 27 | 70 | 80 | 94 | | 1 | 28 | 14 | 33 |

+------+------+------+------+------+------+------+------+------+------+

| 20.0 | 60.5 | 16.5 | 30.4 | 36.1 | 5.83 | 0 | 30.0 | 25.0 | 10.1 |

| 25 | 31 | 53 | 14 | 39 | 1 | | 17 | 80 | 98 |

+------+------+------+------+------+------+------+------+------+------+

| 50.0 | 90.4 | 45.7 | 60.2 | 65.7 | 35.2 | 30.0 | 0 | 5.09 | 20.2 |

| 40 | 49 | 06 | 99 | 65 | 28 | 17 | | 9 | 24 |

+------+------+------+------+------+------+------+------+------+------+

| 45.1 | 85.5 | 41 | 55.4 | 61 | 30.4 | 25.0 | 5.09 | 0 | 15.5 |

| 00 | 86 | | 44 | | 14 | 80 | 9 | | 24 |

+------+------+------+------+------+------+------+------+------+------+

| 30.0 | 70.2 | 25.4 | 40.1 | 45.5 | 15.0 | 10.1 | 20.2 | 15.5 | 0 |

| 17 | 57 | 95 | 12 | 41 | 33 | 98 | 24 | 24 | |

+------+------+------+------+------+------+------+------+------+------+

k=9

+-------+-------+-------+-------+-------+-------+-------+-------+-------+

| 0 | 40.60 | 7.810 | 10.77 | 17.0 | 15.13 | 20.02 | 61.72 | 30.01 |

| | 78810 | 24967 | 03296 | | 27459 | 49843 | 69089 | 66620 |

| | 085 | 591 | 143 | | 504 | 945 | 274 | 396 |

+=======+=======+=======+=======+=======+=======+=======+=======+=======+

| 40.60 | 0 | 45.01 | 30.15 | 25.02 | 55.22 | 60.53 | 115.6 | 70.25 |

| 8 | | 1 | 0 | 0 | 7 | 1 | 57 | 7 |

+-------+-------+-------+-------+-------+-------+-------+-------+-------+

| 7.810 | 45.01 | 0 | 15.13 | 20.10 | 10.77 | 16.55 | 56.10 | 25.49 |

| | 1 | | 3 | 0 | 0 | 3 | 4 | 5 |

+-------+-------+-------+-------+-------+-------+-------+-------+-------+

| 10.77 | 30.15 | 15.13 | 0 | 6.403 | 25.08 | 30.41 | 75.46 | 40.11 |

| 0 | 0 | 3 | | | 0 | 4 | 2 | 2 |

+-------+-------+-------+-------+-------+-------+-------+-------+-------+

| 17 | 25.02 | 20.10 | 6.403 | 0 | 30.59 | 36.13 | 82.81 | 45.54 |

| | 0 | 0 | | | 4 | 9 | 0 | 1 |

+-------+-------+-------+-------+-------+-------+-------+-------+-------+

| 15.13 | 55.22 | 10.77 | 25.08 | 30.59 | 0 | 5.831 | 42.06 | 15.03 |

| 3 | 7 | 0 | 0 | 4 | | | 1 | 3 |

+-------+-------+-------+-------+-------+-------+-------+-------+-------+

| 20.02 | 60.53 | 16.55 | 30.41 | 36.13 | 5.831 | 0 | 35.03 | 10.19 |

| 5 | 1 | 3 | 4 | 9 | | | 1 | 8 |

+-------+-------+-------+-------+-------+-------+-------+-------+-------+

| 61.72 | 115.6 | 56.10 | 75.46 | 82.81 | 42.06 | 35.03 | 0 | 22.13 |

| 7 | 57 | 4 | 2 | 0 | 1 | 1 | | 2 |

+-------+-------+-------+-------+-------+-------+-------+-------+-------+

| 30.01 | 70.25 | 25.49 | 40.11 | 45.54 | 15.03 | 10.19 | 22.13 | 0 |

| 7 | 7 | 5 | 2 | 1 | 3 | 8 | 2 | |

+-------+-------+-------+-------+-------+-------+-------+-------+-------+

k=8

+--------+--------+--------+--------+--------+--------+--------+--------+

| 0 | 40.607 | 7.8102 | 10.770 | 17.0 | 21.494 | 61.726 | 30.016 |

| | 881008 | 496759 | 329614 | | 836265 | 908927 | 662039 |

| | 5 | 1 | 3 | | | 4 | 6 |

+========+========+========+========+========+========+========+========+

| 40.608 | 0 | 45.011 | 30.150 | 25.020 | 75.228 | 115.65 | 70.257 |

| | | | | | | 7 | |

+--------+--------+--------+--------+--------+--------+--------+--------+

| 7.810 | 45.011 | 0 | 15.133 | 20.100 | 16.272 | 56.104 | 25.495 |

+--------+--------+--------+--------+--------+--------+--------+--------+

| 10.770 | 30.150 | 15.133 | 0 | 6.403 | 35.052 | 75.462 | 40.112 |

+--------+--------+--------+--------+--------+--------+--------+--------+

| 17 | 25.020 | 20.100 | 6.403 | 0 | 42.545 | 82.810 | 45.541 |

+--------+--------+--------+--------+--------+--------+--------+--------+

| 21.495 | 75.228 | 16.272 | 35.052 | 42.545 | 0 | 54.904 | 14.877 |

+--------+--------+--------+--------+--------+--------+--------+--------+

| 61.727 | 115.65 | 56.104 | 75.462 | 82.810 | 54.904 | 0 | 22.132 |

| | 7 | | | | | | |

+--------+--------+--------+--------+--------+--------+--------+--------+

| 30.017 | 70.257 | 25.495 | 40.112 | 45.541 | 14.877 | 22.132 | 0 |

+--------+--------+--------+--------+--------+--------+--------+--------+

k=7

+--------+-----------+-----------+-----------+-----------+-----------+-----------+

| 0 | 40.607881 | 7.8102496 | 16.379178 | 21.494836 | 61.726908 | 30.016662 |

| | 0085 | 7591 | 3304 | 265 | 9274 | 0396 |

+========+===========+===========+===========+===========+===========+===========+

| 40.608 | 0 | 45.011 | 34.645 | 75.228 | 115.657 | 70.257 |

+--------+-----------+-----------+-----------+-----------+-----------+-----------+

| 7.810 | 45.011 | 0 | 21.354 | 16.272 | 56.104 | 25.495 |

+--------+-----------+-----------+-----------+-----------+-----------+-----------+

| 16.379 | 34.645 | 21.354 | 0 | 54.996 | 115.503 | 54.968 |

+--------+-----------+-----------+-----------+-----------+-----------+-----------+

| 21.495 | 75.228 | 16.272 | 54.996 | 0 | 54.904 | 14.877 |

+--------+-----------+-----------+-----------+-----------+-----------+-----------+

| 61.727 | 115.657 | 56.104 | 115.503 | 54.904 | 0 | 22.132 |

+--------+-----------+-----------+-----------+-----------+-----------+-----------+

| 30.017 | 70.257 | 25.495 | 54.968 | 14.877 | 22.132 | 0 |

+--------+-----------+-----------+-----------+-----------+-----------+-----------+

k=6

+--------+-------------+-------------+-------------+-------------+-------------+

| 0 | 54.47591060 | 24.39472644 | 24.41990187 | 84.46809174 | 34.40442317 |

| | 68 | 68 | 28 | 33 | 97 |

+========+=============+=============+=============+=============+=============+

| 54.476 | 0 | 34.645 | 75.228 | 115.657 | 70.257 |

+--------+-------------+-------------+-------------+-------------+-------------+

| 24.395 | 34.645 | 0 | 54.996 | 115.503 | 54.968 |

+--------+-------------+-------------+-------------+-------------+-------------+

| 24.420 | 75.228 | 54.996 | 0 | 54.904 | 14.877 |

+--------+-------------+-------------+-------------+-------------+-------------+

| 84.468 | 115.657 | 115.503 | 54.904 | 0 | 22.132 |

+--------+-------------+-------------+-------------+-------------+-------------+

| 34.404 | 70.257 | 54.968 | 14.877 | 22.132 | 0 |

+--------+-------------+-------------+-------------+-------------+-------------+

k=5

+--------+---------------+---------------+---------------+---------------+

| 0 | 54.4759106068 | 24.3947264468 | 34.2276795506 | 84.4680917433 |

+========+===============+===============+===============+===============+

| 54.476 | 0 | 34.645 | 87.830 | 115.657 |

+--------+---------------+---------------+---------------+---------------+

| 24.395 | 34.645 | 0 | 71.027 | 115.503 |

+--------+---------------+---------------+---------------+---------------+

| 34.228 | 87.830 | 71.027 | 0 | 51.252 |

+--------+---------------+---------------+---------------+---------------+

| 84.468 | 115.657 | 115.503 | 51.252 | 0 |

+--------+---------------+---------------+---------------+---------------+

k=4

+---------+---------------+---------------+---------------+

| 0 | 48.5938237752 | 64.7269108126 | 125.182321436 |

+=========+===============+===============+===============+

| 48.594 | 0 | 87.830 | 115.657 |

+---------+---------------+---------------+---------------+

| 64.727 | 87.830 | 0 | 51.252 |

+---------+---------------+---------------+---------------+

| 125.182 | 115.657 | 51.252 | 0 |

+---------+---------------+---------------+---------------+

k=3

+---------+---------------+---------------+

| 0 | 82.3284543958 | 142.982492917 |

+=========+===============+===============+

| 82.328 | 0 | 51.252 |

+---------+---------------+---------------+

| 142.982 | 51.252 | 0 |

+---------+---------------+---------------+

k=2

+---------+---------------+

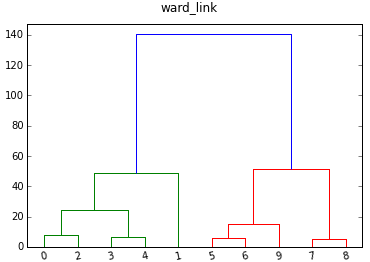
| 0 | 140.324632695 |

+=========+===============+

| 140.325 | 0 |

+---------+---------------+

**Dendogram**

****

**Problem 3)** Model Evaluation (20 pts.)  
We are given the following classification results for a cancer screening test. Please derive the evaluation parameters:

1. Accuracy
2. Specificity
3. Sensitivity
4. Precision
5. Recall

|  |  |  |  |
| --- | --- | --- | --- |
| Actual / Prediction | test = positive | test = negative | Total |
| cancer = yes | **90** | **210** | 300 |
| cancer = no | **140** | **9560** | 9700 |
| Total | 230 | 9770 | 10000 |

**Solution**

|  |  |  |  |
| --- | --- | --- | --- |
| Actual / Prediction | test = positive | test = negative | Total |
| cancer = yes | **TP = 90** | **FN = 210** | 300 |
| cancer = no | **FP = 140** | **TN = 9560** | 9700 |
| Total | 230 | 9770 | 10000 |

TP = 90

FN = 210

FP = 140

TN = 9560

**Accuracy = #correct / N**

**= (**TP + TN) / (TP + FP + TN + FN)

= (90 + 9560) / (90 + 140 + 9560 + 210)

= 0.965

**Specificity = TN / (TN + FP)**

= 9560 / 9700

= 0.986

**Sensitivity = TP / (TP + FN)**

= 90 / 300

= 0.3

**Precision = TP / (TP + FP)**

= 90 / 230

= 0.391

**Recall = Sensitivity**

= 0.3

**Problem 4)** Python program (40 pts)

Take the sk-learn sample code *cluster\_kmeans.py* discussed in the class that is used for clustering iris dataset. Make the following changes:

1. In addition to using PCA to reduce features from 4 to 2, also evaluate using the original feature pairs (1,2) and (3, 4).
2. For all the 3 clustering settings (original, (1,2), (3, 4)), calculate the clustering quality

CQ = IE / EV as defined in the class:

latex-image-1.pdf and latex-image-1.pdf

**Output:**

**3 plots of k-means clustering results for PCA, features (1,2), and features (3,4).**

**3 CQ values for PCA, features (1,2), and features (3,4).**