HackShitUp / **Data-Mining**

Branch: master ▾ **Data-Mining** / HW-4 /

Create new file | Find file | History

HackShitUp Completed hw 4. Latest commit 181b7e7 6 minutes ago

..

HW4.pdf	- Push	yesterday
README.md	Completed hw 4.	6 minutes ago
segment.arff	- Push	yesterday
test.swift	- Push	yesterday

README.md

HW-4

```
Josh Choi
Data Mining
Yijun Zhao
11/22/19
```

1.

(a) Report the SSE values of ach clustering run for each *k* value.

```
Given:
k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

∗ k = 1
• Seed = 1: 3804.744383487352
• Seed = 2: 3804.7443834873525
• Seed = 3: 3804.744383487353
• Seed = 4: 3804.744383487353
• Seed = 5: 3804.7443834873525

∗ k = 2
• Seed = 1: 2679.2831700304127
• Seed = 2: 2675.9791802778172
• Seed = 3: 2679.2831700304123
• Seed = 4: 2679.2831700304123
• Seed = 5: 2679.2831700304127

∗ k = 3
• Seed = 1: 2076.046974451557
• Seed = 2: 2067.3943009564287
• Seed = 3: 2097.2292114738843
• Seed = 4: 2163.2404801332495
• Seed = 5: 2070.162850875394
```

∗ k = 4
- Seed = 1: 2056.620348700981
- Seed = 2: 1551.9131911443642
- Seed = 3: 1642.9138686894187
- Seed = 4: 1715.9489102547257
- Seed = 5: 1665.3563431309058

∗ k = 5
- Seed = 1: 1634.2039979339227
- Seed = 2: 1633.2457462359876
- Seed = 3: 1170.4945259691226
- Seed = 4: 1194.3849991231245
- Seed = 5: 1645.8937444622636

∗ k = 6
- Seed = 1: 1165.1103592101838
- Seed = 2: 1162.0856789536233
- Seed = 3: 836.3536270105087
- Seed = 4: 1173.8415533841737
- Seed = 5: 1174.9295639714344

∗ k = 7
- Seed = 1: 809.290574529901
- Seed = 2: 1142.638143570465
- Seed = 3: 814.9944462630268
- Seed = 4: 818.6153664649133
- Seed = 5: 819.4695861690018

∗ k = 8
- Seed = 1: 474.9410917910018
- Seed = 2: 1128.4055223846658
- Seed = 3: 475.7815733735389
- Seed = 4: 474.99383053363255
- Seed = 5: 475.87600826995697

∗ k = 9
- Seed = 1: 456.0157786166358
- Seed = 2: 1106.0223417277023
- Seed = 3: 454.40848503934603
- Seed = 4: 466.9579060993053
- Seed = 5: 457.6651436775775

∗ k = 10
- Seed = 1: 448.0451849019205
- Seed = 2: 750.40252317349
- Seed = 3: 420.5315346241276
- Seed = 4: 464.009240158985
- Seed = 5: 430.1227625164528

(b) For each k = 1, 2, ..., 10 compute the mean SSE, which we denote $\boldsymbol{\mu_k}$ and the sample standard deviation of SSE, which we denote $\boldsymbol{\sigma_k}$ over all 5 clustering runs for that value of $_k$. Produce a table containing the 4 columns: **k, $\boldsymbol{\mu_k}$, $\boldsymbol{\mu_k - 2\sigma_k}$ and $\boldsymbol{\mu_k + 2\sigma_k}$** for each of the values of k = 1, 2, ..., 10.

Notes:
• μ_k = Mean
• σ_k = Sample Standard Deviation

1. Get mean

2. Per each number: subtract the mean and square the result
3. Get mean of those squared differences
4. Take square root

k	μ_k	σ_k	μ_k-2σ_k	μ_k + 2σ_k
1	3804.7443834874	7.1901869436451E-13	3804.7443834874	3804.7443834874
2	2678.6223720799	1.4775891367531	2675.6671938064	2681.5775503534
3	2094.8147635781	39.999177253144	2014.8164090718	2174.8131180844
4	1726.5505323841	193.84046230267	8245.0717373151	9020.4335865257
5	1455.6446027449	249.5934226265	6779.0361684714	7777.4098589774
6	1102.464156506	148.862398396	5214.5959857379	5810.0455793219
7	881.0016233995	146.3141545705	588.3733142585	1173.6299325405
8	605.9996052706	292.03410568512	21.9313939004	1190.0678166408
9	588.2139310321	289.50466685734	9.2045973174	1167.2232647468
10	502.622249075	139.51679609811	223.5886568788	781.6558412712

2. (20 points) Consider the following dataset:

```
{ 0, 4, 5, 20, 25, 39, 43, 44 }
```

(a) Build a dendrogram for this dataset using the single-link, bottom-up approach. Show your work.

```
-----        -----        -----
|   |        |   |        |   |
4   5       43  45       20  25

--------|          |--------------|
|    -----       -----       |   ----
|    |   |       |   |       |   |  |
0    4   5       20  25      39  43 45

                |
    -------------------------
    |                       |
--------|          |--------------|
|    -----       -----       |   ----
|    |   |       |   |       |   |  |
0    4   5       20  25      39  43 45
```

(b) Suppose we want the two top level clusters. List the data points in each cluster.

```
{0, 4, 5}
{20, 25, 39, 43, 45}
```

3. (20 points) Given two clusters

```
C1 = {(1,1),(2,2),(3,3)}
C2 = {(5,2),(6,2),(7,2),(8,2),(9,2)}
```

compute the values in (a) - (f). Use the definition for scattering criteria presented in class. Note that *tr* in the scattering criterion is referring to the trace of the matrix. 1
(a) The mean vectors m_1 and m_2

- x_1 = [1 + 2 + 3]/3 = 2

* y_1 = [1 + 2 + 3]/3 = 2
m_1 = (2, 2)

* x_2 = [5 + 6 + 7 + 8 + 9]/5 = 7
* y_2 = [2 + 2 + 2 + 2 + 2]/5 = 2
m_2 = (7, 2)
(b) The total mean vector *m*

```
_m_ = (4.38, 2)
```

(c) The scatter matrices S_1 and S_2
$S_i=\sum(\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T$

```
[1 2]       [1 2]T      [-1]
[ - ]   *   [ - ]   =   [  ]   * (-1, -1)   =   [1  1]
[1 2]       [1 2]       [-1]                    [1  1]

[2 2]       [2 2]T      [0]
[ - ]   *   [ - ]   =   [ ]   * (0, 0)     =   [0  0]
[2 2]       [2 2]       [0]                    [0  0]

[3 2]       [3 2]T      [1]
[ - ]   *   [ - ]   =   [ ]   * (1, 1)     =   [1  1]
[3 2]       [3 2]       [1]                    [1  1]

[1 1]   +   [0 0]   +   [1 1]              =   [2  2]
[1 1]       [0 0]       [1 1]                  [2  2]
-------------------------------------------------------
[5 7]       [5 7]       [-2]
[ - ]   *   [ - ]   =   [  ]   * (-2, 0)   =   [4  0]
[2 2]       [2 2]       [ 0]                   [0  0]

[6 7]       [6 7]       [-1]
[ - ]   *   [ - ]   =   [  ]   * (-1, 0)   =   [1  0]
[2 2]       [2 2]       [ 0]                   [0  0]

[7 7]       [7 7]       [-2]
[ - ]   *   [ - ]   =   [  ]   * (-2, 0)   =   [4  0]
[2 2]       [2 2]       [ 0]                   [0  0]

[8 7]       [8 7]       [1]
```

```
[ − ]   ∗   [ − ]   =   [ ]   ∗ (1, 0)    =    [1  0]
[2 2]       [2 2]       [0]                    [0  0]

[9 7]       [9 7]       [2]
[ − ]   ∗   [ − ]   =   [ ]   ∗ (2, 0)    =    [4  0]
[2 2]       [2 2]       [0]                    [0  0]

[4 0]   +   [1 0]   +   [0 0]   +   [1 0]   +   [4 0]   =   [10 0]
[0 0]       [0 0]       [0 0]       [0 0]       [0 0]       [0  0]
```

(d) The within-cluster scatter matrix S_W
S_W =

```
[2  2] ∗ [10 0]     =      [12  2]
[2  2]   [0  0]            [2   2]
```

(e) The between-cluster scatter matrix S_B

```
3   ∗   [2   −   4.38]    ∗   [2  −   4.38]T     =   3 ∗ [−2.38] ∗ (−2.38, 0)    =   [16.9   0]
        [2         2 ]        [2  −      2]              [  0  ]                     [0      0]
5   ∗   [7   −   4.38]    ∗   [7  −   4.38]T     =   5 ∗ [−2.62] ∗ (−2.62, 0)    =   [34.3   0]
        [2         2 ]        [2  −      2]              [  0  ]                     [0      0]
```

S_B =

```
[51.2   0]
[0      0]
```

(f) The scatter criterion $tr(S_B)/tr(S_W)$
$tr(S_B)/tr(S_W)$ = 14/51.2

4. (20 points) A Naive Bayes classifier gives the predicted probability of each data point belonging to the positive class, sorted in a descending order:

Instance #	True Class Label	Predicted Probability of Positive Class
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.43
8	N	0.42
9	N	0.41

10	P	0.4

Suppose we use 0.5 as the threshold to assign the predicted class label to each data point, i.e., if the predicted probability ≥ 0.5, the data point is assigned to positive class; otherwise, it is assigned to negative class. Calculate the Confusion Matrix, Accuracy, Precision, Recall, F1 Score and Specificity of the classifier.

- **Confusion Matrix**

```
 _________________________
 |       |   P   |   N   |
 |-----------------------|
 |   P   |   4   |   1   |
 |-----------------------|
 |   N   |   2   |   3   |
 _________________________
```

- **Accuracy**

```
[TP + TN]/[P + N] = [4 + 3]/[5 + 5] = 7/10
```

- **Precision**

```
[TP]/[TP + FP] = 4/[4+1] = 4/5
```

- **Recall**

```
[TP]/[TP + FN] = 4/[4+2] = 4/6 = 2/3
```

- **F1 Score**

```
[2TP]/[2TP + FP + FN] = [2*4]/[(2*4) + 1 + 2] = 8/11
```

- **Specifity**

```
[TP]/[FP + TN] = 3/[1+3] = 3/4
```