Problem 1

For instance

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | High School | Service | Less than 3 |

P(High School | High) = 1 + 1 / 4 + 2

P(Service | High) = 1 + 1 / 4 + 2

P(Less than 3 | High) = 1 + 1 / 4 + 3

P(High School | Low) = 4 + 1 / 6 + 2

P(Service | Low) = 4 + 1 / 6 + 2

P(Less than 3 | Low) = 2 + 1 / 6 + 3

P(High | X) = 0.0126984126984

P(Low | X) = 0.078125

Because P(High | X) < P(Low | X), this instance should be labeled as Low.

For instance

|  |  |  |  |
| --- | --- | --- | --- |
| 2 | College | Retail | Less than 3 |

P(College | High) = 3 + 1 / 4 + 2

P(Retail | High) = 0 + 1 / 4 + 3

P(Less than 3 | High) = 1 + 1 / 4 + 3

P(College | Low) = 2 + 1 / 6 + 2

P(Retail | Low) = 0 + 1 / 6 + 3

P(Less than 3 | Low) = 2 + 1 / 6 + 3

P(High | X) = 0.0108843537415

P(Low | X) = 0.00833333333333

Because P(High | X) > P(Low | X), this instance should be labeled as High.

For instance

|  |  |  |  |
| --- | --- | --- | --- |
| 3 | Graduate | Service | 3 to 10 |

P(Graduate | High) = 0 + 1 / 4 + 3

P(Service | High) = 1 + 1 / 4 + 2

P(3 to 10 | High) = 1 + 1 / 4 + 3

P(Graduate | Low) = 0 + 1 / 6 + 3

P(Service | Low) = 4 + 1 / 6 + 2

P(3 to 10 | Low) = 2 + 1 / 6 + 3

P(High | X) = 0.00544217687075

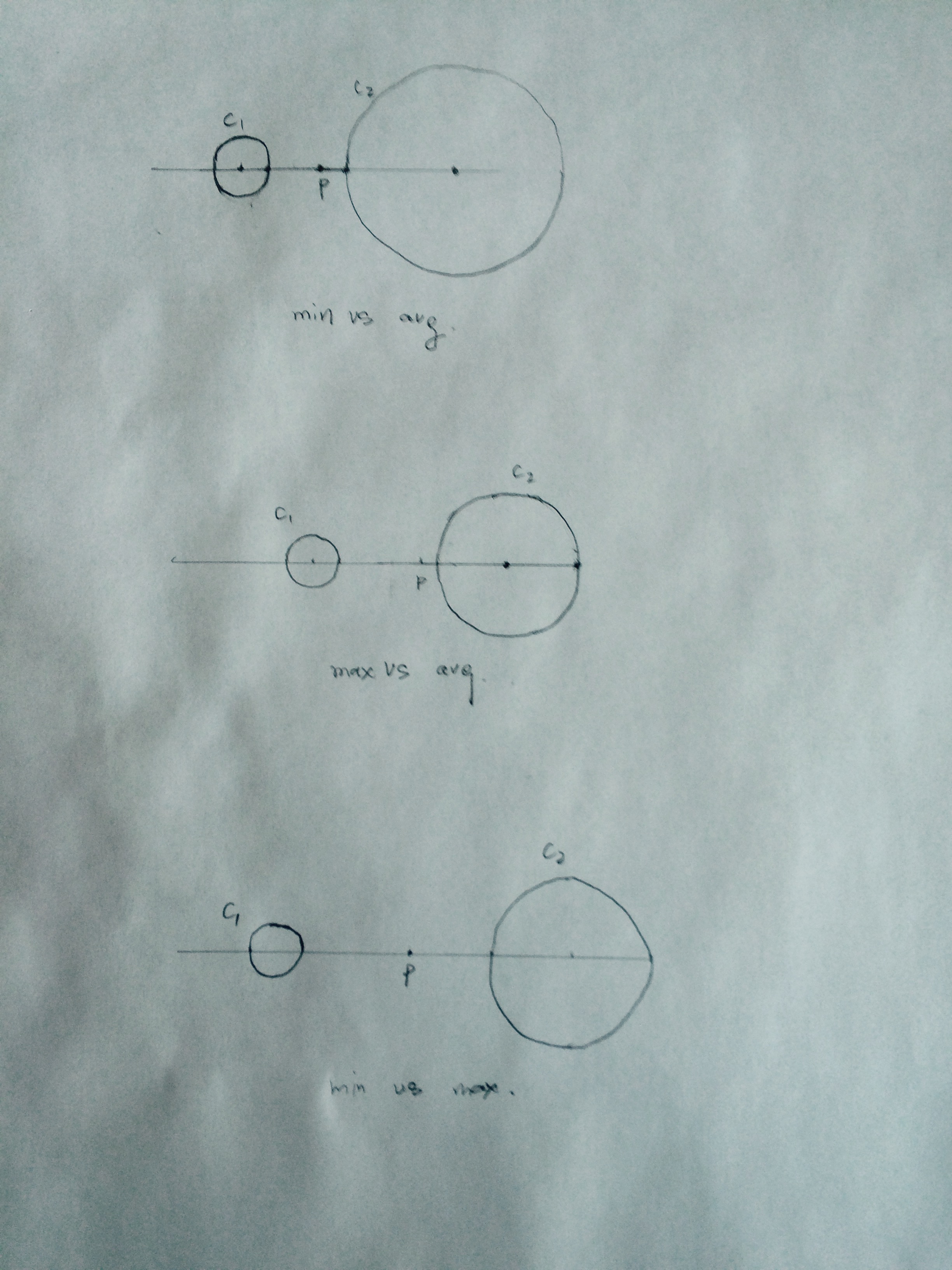
P(Low | X) = 0.0138888888889

Because P(High | X) < P(Low | X), this instance should be labeled as Low.

Problem 2

1. mean vector: m1 = [2,2]T, m2 = [7,2]T
2. Total mean: m = [5,125, 2] T
3. Scatter matrix : S1 = [[2,2],[2,2]], S2 = [[10,0], [0,0]]
4. Within cluster matrix: Sw = [[12,2],[2,2]]
5. Between cluster matrix: Sb = [[46.875,0],[0,0]]
6. Scatter criterion: 3.348

Problem3



In the first graph, if min distance between was choose, p will be merged into C2 at this time, but if avg distance was choose, p will be merged into C1

In the second graph, if max distance between cluster was choose, p will be merged into cluster C2 at this time, but it avg distance was choose p will be merged into C1

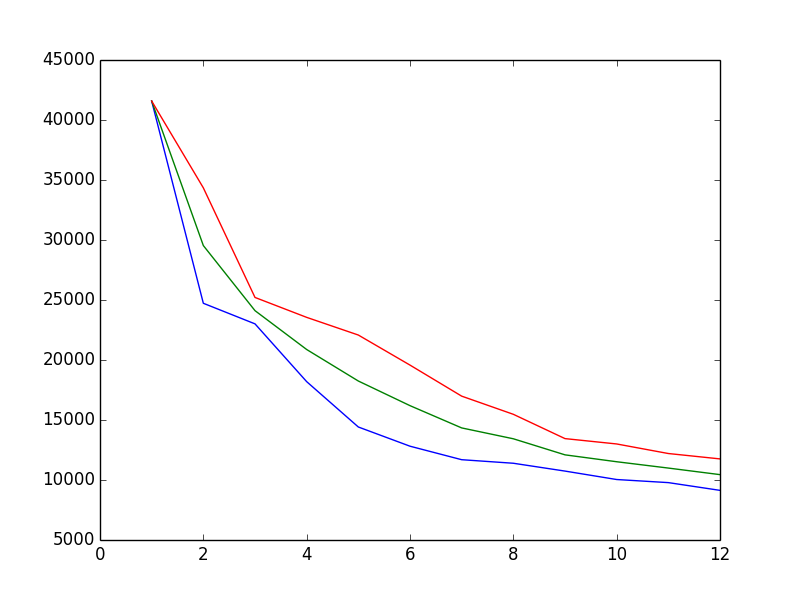
In the third graph, if min distance between cluster was choose, p will be merged into cluster at C2 at this time, but if max distance was choose p will be merged into C1.

Problem 4

1. C1 = [(1,2), (2,3), (3,4)], C2 = [(5,1),(4,2),(5,3),(6,2)]
2. Each pair of points within the same cluster is density-connected points. For this situation, (1,2) and (2,3), (1,2) and (3,4), (2,3) and (3,4) are density-connected points. (5,1) and (4,2), (5,1) and (5,3), (5,1) and (6,2), (4,2) and (5,3), (4,2) and (6,2), (5,3) and (6,2) are density-connected points.
3. In this situation, point (0,0), (1,6), (7,4) are considered as noisy.

Problem5

a)



b)

|  |  |  |  |
| --- | --- | --- | --- |
| k | u – 2 \* sigma | u | u+2\*sigma |
| 1 | 41580 | 41580 | 41580 |
| 2 | 24727.14693 | 29536.67079 | 34346.19465 |
| 3 | 23011.39774 | 24112.01358 | 25212.62942 |
| 4 | 18196.25124 | 20874.71562 | 23553.17999 |
| 5 | 14416.56462 | 18248.27137 | 22079.97812 |
| 6 | 12816.3799 | 16195.78241 | 19575.18492 |
| 7 | 11691.78374 | 14340.01024 | 16988.23674 |
| 8 | 11395.47541 | 13434.91712 | 15474.35884 |
| 9 | 10742.8972 | 12096.09477 | 13449.29235 |
| 10 | 10039.81162 | 11521.56762 | 13003.32361 |
| 11 | 9781.918432 | 10994.14173 | 12206.36502 |
| 12 | 9136.729933 | 10447.31399 | 11757.89805 |

1. By k increases and approaches the total number of N, SSE will decrease and when k=N SSE=0. In general, by increasing k SSE will decrease, so there is not a optimal SSE for k. If we simply think lower SSE means better cluster result, using SSE selected optimal k will equals to N.
2. We can use scatter criterion. A good partition should have high trace of between cluster matrix(Sb) and low trace of within cluster matrix(Sw), and have high scatter criterion trace(Sb) / trace(Sw).

Homework 2 Problem 1

node 0

Top: 6,4, 0.97

Education Level gain = 0.125

Career gain = 0.125

Years of Experience gain = 0.020

Selected Attribute: Education

node 1

High School 4,1, 0.72

Career gain = 0.171

Years of Experience gain = 0.322

Selected Attribute: Years of Experience

node 3

More than 10

Career gain = 1.0

Selected attribute Career

node 8

Management

Class High

node 9

Service

Class Low

node 4

Less than 3

Class Low

node 5

3 to 10

Class Low

node 2

College 3,2, 0.97

Career gain = 0.420

Years of Experience gain = 0.171

Selected Attribute: Career

node 6

Management

class High

node 7

Service 1,2, 0,91

Years of Experience gain = 0.918

selected attribute: Years of Experience

node 10

More than 10

Class Low

node 11

Less than 3

Class Low

node 12

3 to 10

Class High

Apply the pruning set,

|  |  |  |
| --- | --- | --- |
| node | prune error | keep error |
| 0 | 1 | 2 |
| 1 | 0 | 0 |
| 2 | 0 | 2 |
| 3 | 0 | 0 |
| 7 | 0 | 1 |

Node 2 will be pruned. Then the decision will be:

node 0

Top: 6,4, 0.97

Education Level gain = 0.125

Career gain = 0.125

Years of Experience gain = 0.020

Selected Attribute: Education

node 1

High School 4,1, 0.72

Career gain = 0.171

Years of Experience gain = 0.322

Selected Attribute: Years of Experience

node 3

More than 10

Career gain = 1.0

Selected attribute Career

node 8

Management

Class High

node 9

Service

Class Low

node 4

Less than 3

Class Low

node 5

3 to 10

Class Low

node 2

Class Low

Apply pruning set to decision tree:

|  |  |  |
| --- | --- | --- |
| node | prune error | keep error |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 3 | 0 | 0 |

No more node should be pruned.