INFS 4203 / 7203 Data Mining
Tutorial 4: Classification and Clustering

Seun Aremu, Doris He
o.aremu@uq.edu.au, d.he@uq.edu.au

T4-Q1

Construct a decision tree that will properly classify each observation using a GINI index based splitting criterion.

RID	AGE	INCOME	STUDENT	RATING	CLASS
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle-aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle-aged	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle-aged	Medium	No	Excellent	Yes
13	Middle-aged	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

T4-Q1

Decision Tree

- Objective: To construct a good decision tree from the training set.
- Problems:
  - How to evaluate the quality of a candidate split?
- GINI index:
  - GINI index indicate the impurity of the node t
    - GINI index = 0 = pure
  - GINI index:

$$GINI(t) = 1 - \Sigma_{j=1}^{n_c} p(j/t)^2$$

  - $p(j/t)$ is the relative frequency of class j at node t

T4-Q1

Decision Tree

- GINI index:

- Measuring the quality of split when node p is split into k partitions

$$GINI_{split} = \Sigma_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

  - n_i = number of records at child i
  - n = number of records at parent node p

- Minimize the GINI Index

T4-Q1

Decision Tree

- Compute GINI index for each attribute:
- Age:
  - Binary Split:
    - case1: $GINI_{split\ (age)}$ = *{youth, middle-aged} , {senior}*
    - case2: $GINI_{split\ (age)}$ = *{youth}, {middle-aged, senior}*
  - Multi-way Split:
    - case3: $GINI_{split\ (age)}$ = *{youth}, {middle-aged}, {senior}*
  - **Minimize GINI Index**

T4-Q1

Decision Tree

- Compute GINI index for each attribute:

	youth/ middle-age	senior
Yes	6	3
No	3	2

- Age:
  - Binary Split:
    - $GINI_{split\ (age)} = \left(\frac{9}{14} * \left(1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2\right)\right) + \left(\frac{5}{14} * \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right)\right) = 0.457$

T4-Q1

Decision Tree

- Compute GINI index for each attribute:

	youth/ middle-age	senior
Yes	6	3
No	3	2

	youth	senior/ middle-age
Yes	2	7
No	3	2

- Age:
  - Binary Split:
    - $GINI_{split\ (age)} = \left(\frac{9}{14} * \left(1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2\right)\right) + \left(\frac{5}{14} * \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right)\right) = 0.457$
    - $GINI_{split\ (age)} = \left(\frac{5}{14} * \left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right)\right) + \left(\frac{9}{14} * \left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right)\right) = 0.394$

T4-Q1

Decision Tree

- Compute GINI index for each attribute:

	youth/ middle-age	senior
Yes	6	3
No	3	2

	youth	senior/ middle-age
Yes	2	7
No	3	2

	youth	middle-age	senior
Yes	2	4	3
No	3	0	2

- Age:
  - Binary Split:
    - $GINI_{split\ (age)} = \left(\frac{9}{14} * \left(1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2\right)\right) + \left(\frac{5}{14} * \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right)\right) = 0.457$
    - $GINI_{split\ (age)} = \left(\frac{5}{14} * \left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right)\right) + \left(\frac{9}{14} * \left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right)\right) = 0.394$
  - Multi-way Split:
    - $GINI_{split\ (age)} = \left(\frac{5}{14} * \left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right)\right) + \left(\frac{4}{14} * \left(1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right)\right) + \left(\frac{5}{14} * \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right)\right) = 0.343$

T4-Q1

Decision Tree

- Compute GINI index for each attribute:

	youth/ middle-age	senior
Yes	6	3
No	3	2

	youth	senior/ middle-age
Yes	2	7
No	3	2

	youth	middle-age	senior
Yes	2	4	3
No	3	0	2

- Age:
  - Binary Split:
    - $GINI_{split\ (age)} = \left(\frac{9}{14} * \left(1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2\right)\right) + \left(\frac{5}{14} * \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right)\right) = 0.457$
    - $GINI_{split\ (age)} = \left(\frac{5}{14} * \left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right)\right) + \left(\frac{9}{14} * \left(1 - \left(\frac{7}{9}\right)^2 - \left(\frac{2}{9}\right)^2\right)\right) = 0.394$
  - Multi-way Split:
    - $GINI_{split\ (age)} = \left(\frac{5}{14} * \left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right)\right) + \left(\frac{4}{14} * \left(1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right)\right) + \left(\frac{5}{14} * \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right)\right) = 0.343$
  - **Minimize GINI Index**

T4-Q1

Decision Tree

- Compute GINI index for each attribute:
- Income:
  - Binary Split:
    - case1: $GINI_{split\ (income)}$ = *{low, medium} , {high}*
    - case2: $GINI_{split\ (income)}$ = *{low}, {medium, high}*
  - Multi-way Split:
    - case3: $GINI_{split\ (income)}$ = *{low}, {medium}, {high}*
  - **Minimize GINI Index**

T4-Q1

Decision Tree

- Compute GINI index for each attribute:

	low/ medium	high
Yes	7	2
No	3	2

	low	high/ medium
Yes	3	6
No	1	4

- Income:
  - Binary Split:
    - $GINI_{split\ (income)} = \left(\frac{10}{14} * \left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right)\right) + \left(\frac{4}{14} * \left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)\right) = 0.557$
    - $GINI_{split\ (income)} = \left(\frac{4}{14} * \left(1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right)\right) + \left(\frac{10}{14} * \left(1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2\right)\right) = 0.45$
  - Multi-way Split:
    - $GINI_{split\ (income)} = \left(\frac{4}{14} * \left(1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right)\right) + \left(\frac{6}{14} * \left(1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2\right)\right) + \left(\frac{4}{14} * \left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)\right) = 0.393$

	low	medium	high
Yes	3	4	2
No	1	2	2

T4-Q1

Decision Tree

- Compute GINI index for each attribute:

	yes	no
Yes	6	3
No	1	4

- Student:
  - Binary Split:
    - case1: $GINI_{split\ (student)}$ = {yes} , {no}
    - $GINI_{split\ (student)} = \left(\frac{7}{14} * \left(1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right)\right) + \left(\frac{7}{14} * \left(1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2\right)\right) = 0.367$

- Rating:
  - Binary Split:
    - case1: $GINI_{split\ (rating)}$ = {fair} , {excellent}
    - $GINI_{split\ (rating)} = \left(\frac{8}{14} * \left(1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2\right)\right) + \left(\frac{6}{14} * \left(1 - \left(\frac{3}{6}\right)^2 - \left(\frac{2}{6}\right)^2\right)\right) = 0.488$

	fair	*excellent*
Yes	6	3
No	2	3

T4-Q1

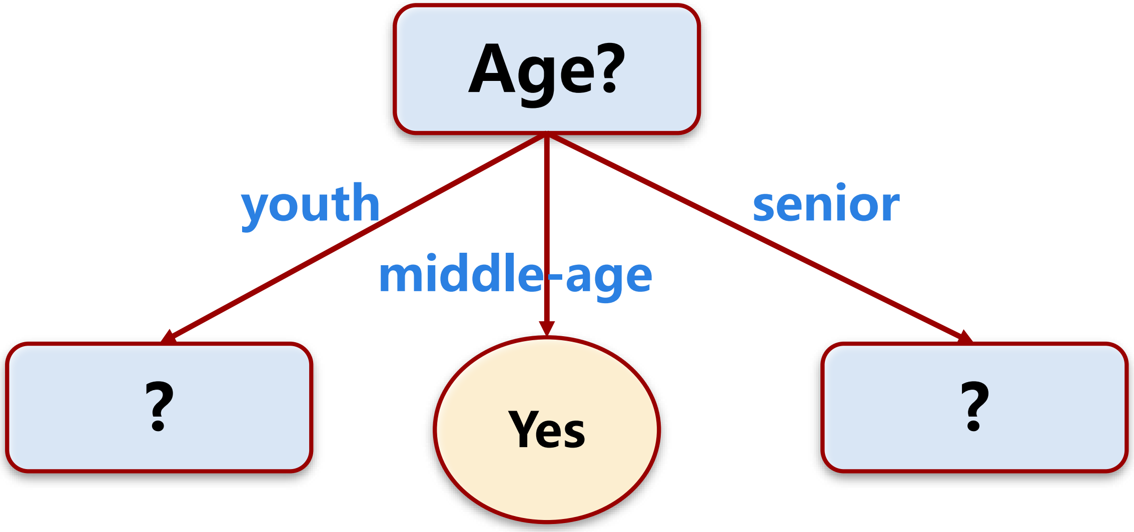
Decision Tree

- Compute GINI index for each attribute:
- Age: $GINI_{split\ (age)} = 0.343$
- Income: $GINI_{split\ (income)} = 0.393$
- Student: $GINI_{split\ (student)} = 0.367$
- Rating: $GINI_{split\ (rating)} = 0.488$

T4-Q1

Decision Tree

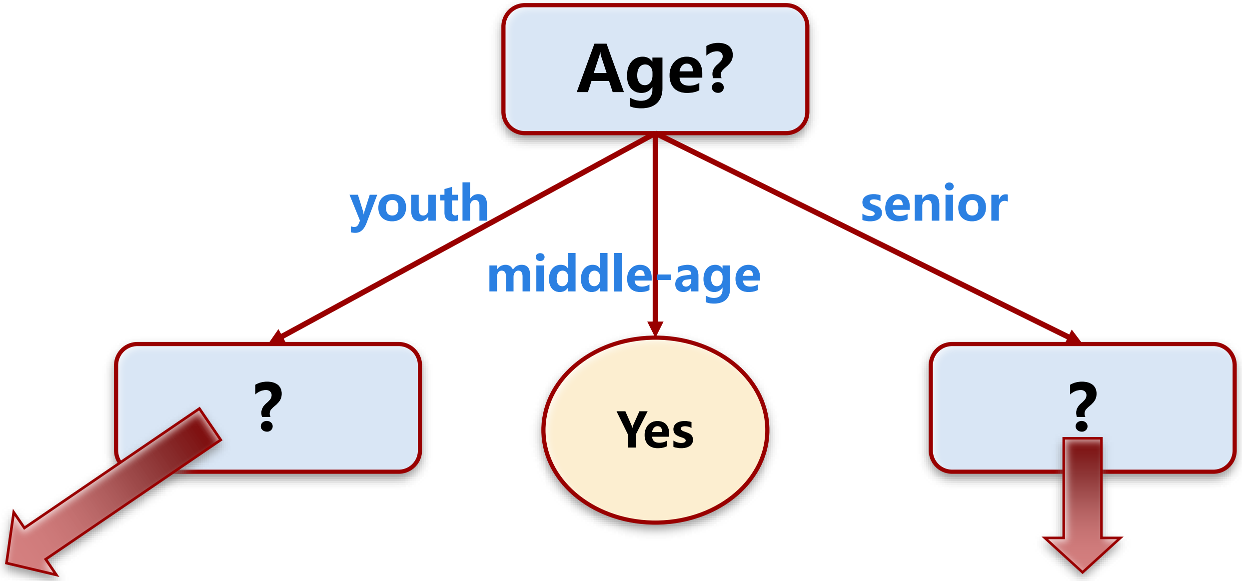
- Compute GINI index for each attribute:
- Age: $GINI_{split\ (age)} = 0.343$
- Income: $GINI_{split\ (income)} = 0.393$
- Student: $GINI_{split\ (student)} = 0.367$
- Rating: $GINI_{split\ (rating)} = 0.488$



T4-Q1

Decision Tree

- Compute GINI index for each attribute:
- Age: $GINI_{split\ (age)} = 0.343$
- Income: $GINI_{split\ (income)} = 0.393$
- Student: $GINI_{split\ (student)} = 0.367$
- Rating: $GINI_{split\ (rating)} = 0.488$



subset of data for the left branch

RID	AGE	INCOME	STUDENT	RATING	CLASS
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes

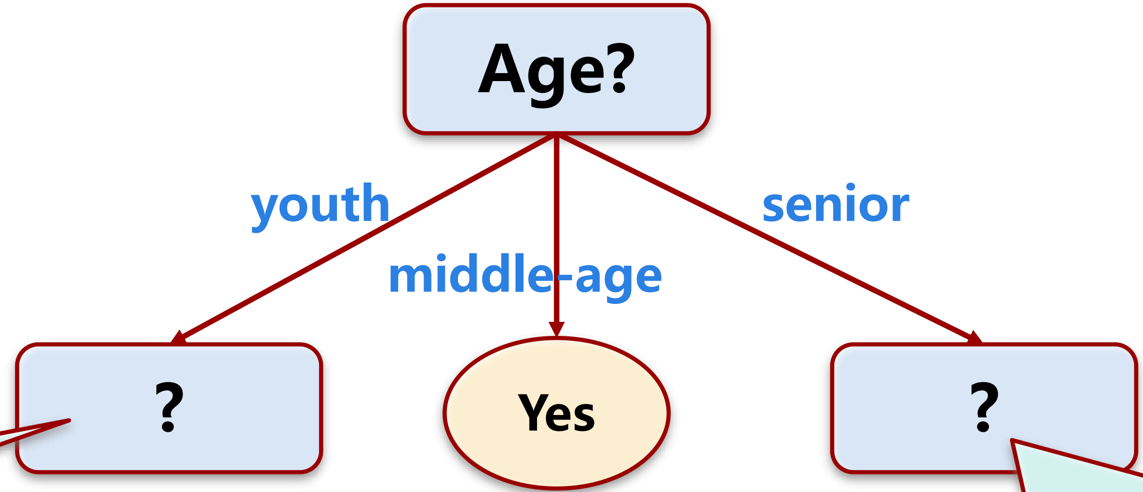
subset of data for the right branch

RID	AGE	INCOME	STUDENT	RATING	CLASS
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
10	Senior	Medium	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

T4-Q1

Decision Tree

- Compute GINI index for each attribute:
- Age: $GINI_{split\ (age)} = 0.343$
- Income: $GINI_{split\ (income)} = 0.393$
- Student: $GINI_{split\ (student)} = 0.367$
- Rating: $GINI_{split\ (rating)} = 0.488$



choose student for the left branch

choose rating for the right branch

RID	AGE	INCOME	STUDENT	RATING	CLASS
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes

RID	AGE	INCOME	STUDENT	RATING	CLASS
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
10	Senior	Medium	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

T4-Q1

Decision Tree

- Compute GINI index for each attribute:

- Student: $GINI_{split\ (student)} = \left(\frac{2}{5} * \left(1 - \left(\frac{2}{2}\right)^2 - (0)^2\right)\right) + \left(\frac{3}{5} * \left(1 - (0)^2 - \left(\frac{3}{3}\right)^2\right)\right) = 0$

	yes	no
Yes	2	0
No	0	3

RID	AGE	INCOME	STUDENT	RATING	CLASS
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes

T4-Q1

Decision Tree

- Compute GINI index for each attribute:

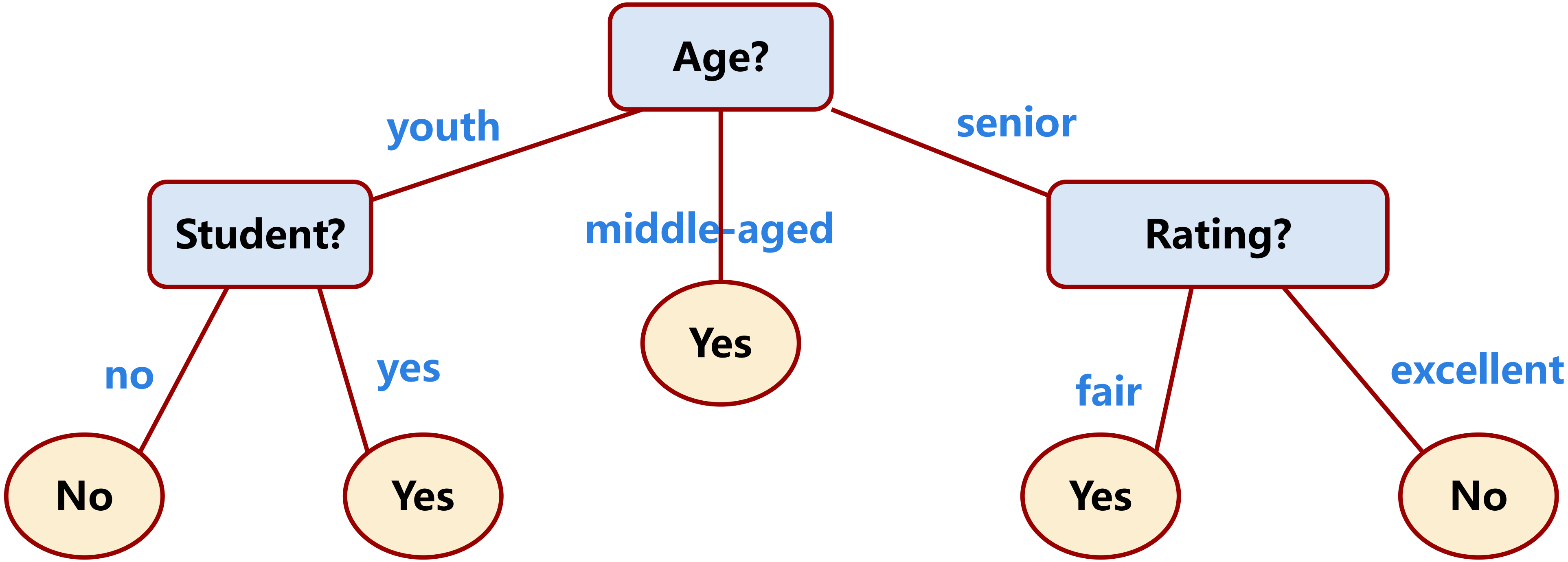
- Student: $GINI_{split\ (rating)} = \left(\frac{2}{5} * \left(1 - \left(\frac{3}{3}\right)^2 - (0)^2\right)\right) + \left(\frac{3}{5} * \left(1 - (0)^2 - \left(\frac{2}{2}\right)^2\right)\right) = 0$

	fair	excellent
Yes	3	0
No	0	2

RID	AGE	INCOME	STUDENT	RATING	CLASS
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
10	Senior	Medium	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

T4-Q1

Decision Tree



+

Decision tree

Summary

- Tree is constructed in a **top-down recursive divide and conquer manner**
  - At start, all the training examples are at the root.
  - Attributes are categorical
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.g., GINI Index)
- Stopping partitioning when
  - All samples for a given node belong to the same class
  - All the records have similar/the same attribute values

T4-Q2

Partitioning clustering method

- Suppose the data mining task is to cluster the following measurements of the variable *age* into **three** groups: {18, 22, 25, 42, 27, 43, 33, 35, 56, 28}
  - Use *k-means* algorithm to show the clustering procedures **step by step;** and
  - Calculate corresponding **SSE** values.

T4-Q2

Partitioning method

- K-partitioning method:
  - Partitioning a dataset D of into a set of K clusters so that an objective function is optimized.
- A typical objective function: Sum of Squared Errors (SSE)
  - $SSE(C) = \Sigma_{i=1}^{K} \Sigma_{x \in C_i} dist^2(C_i, x)$
- K-means

T4-Q2

K-means clustering

- Given K , the number of clusters
  - Select K points as initial centroids randomly
  - **Repeat**
    - Form K clusters by assigning each point to its closest centroid
    - Re-compute the centroids (mean point) of each cluster
  - **Until** convergence criterion is satisfied

T4-Q2

Initial centroids: 22, 35, 43

the old clusters are no use

Cluster#	Old Centroid	Cluster Elements	new Centroid
1	22	18,22,25,27,28	24
2	35	33,35	34
3	43	42,43,56	47

Cluster#	Old Centroid	Cluster Elements	new Centroid
1	24	18,22,25,27,28	24
2	34	33,35	34
3	47	42,43,56	47

SSE = 190

use the final clusters to calculate SSE

R1	22	35	43
18	4	7	25
22	0	13	21
25	3	10	18
42	20	7	1
27	5	8	16
43	21	8	0
33	11	2	10
35	13	0	8
56	34	21	13
28	6	7	15

R2	24	34	47
18	6	16	29
22	2	12	25
25	1	9	22
42	18	8	5
27	3	7	30
43	19	9	4
33	9	1	14
35	11	1	12
56	32	22	9
28	4	6	19

T4-Q2

Initial centroids: 18, 27, 35

Cluster	Old centroid	Cluster Elements	New Centroid	
1	18	18, 22	20	
2	27	25, 27, 28	26.7	ROUND1
3	35	33, 35, 42, 43, 56	41.8	
1	20	18, 22	20	
2	26.7	25, 27, 28, 33	28.25	ROUND2
3	41.8	35, 42, 43, 56	44	
1	20	18, 22	20	
2	28.25	25, 27, 28, 33, 35	29.6	ROUND3
3	44	42, 43, 56	47	
1	20	18, 22	20	
2	29.6	25, 27, 28, 33, 35	29.6	ROUND4
3	47	42, 43, 56	47	

SSE = 201.2

T4-Q2

Discussion of k-means

- When $k \ll n$, k-means is an efficient algorithm
- The clustering quality is sensitive to the **initial position**.
- Need to specify K
- Sensitive to noisy data and outliers
- Only valid to convex shapes

Thanks for your attention