DATA MINING (INFO-H-423)
Mahmoud SAKR
Intended Learning Outcomes

Introduction & Decision Tree
• Explain the steps of supervised learning. • Describe classification as one of the DM tasks. • Explain/Illustrate the concepts of Entropy, Gain, Gain-ratio. • Explain and Apply the ID3 Algorithm to a given dataset. • Illustrate and compare the different model validation methods: holdout, cross validation, bootstrap.
Data Preparation and Distance Measures
• Describe the different data preparation tasks, their goals, and few examples. The details of the methods are not required. It is enough to know the general idea. • Describe without detail the concepts and methods of data cleaning, scaling, normalization, and distance measures. • Explain and apply the Lp-norm distance, and the Edit distance.
Clustering
• Explain clustering as an unsupervised learning method. • Apply the K-Means, the K-Medians, and the DBSCAN Algorithms to given datasets. • Compare K-Means, K-Medians and K-Medoids Algorithms. • Compare between the different families of clustering Algorithms (representative-based, density based, probabilistic model-based). • Analyze the complexity of the studied clustering Algorithms. • Explain how to assess clustering quality Note: the details of the probabilistic model-based clustering are not required for this exam. Only the general concept of fuzzy clustering is required.
Classification
• Explain the naive Bayes classifier, and apply it to a given dataset. • Explain the concept of the confusion matrix, and use it to define the different classification quality measures. • Motivate ensemble learning, as a method to improve classification accuracy. • Illustrate the difference between bagging and boosting. • Describe random forest induction, and the tuning of parameters L, D.
Frequent Patterns & Association rules Mining
• Illustrate the uses of frequent pattern mining. • Explain and compare the measures: support, confidence, lift, correlation analysis. • Explain the Apriori property, and its applications in optimizing the search for frequent itemsets. • Apply the Apriori and the FP-Growth Algorithms to given datasets. • Compare the frequent pattern mining Algorithms: brute force enumeration of the itemset lattice, Apriori, and FP-Growth Algorithms. • Apply your understanding of the different frequent itemset mining methods to reason about further optimizations. • Extract association rules from frequent itemsets, and assess their quality.

Sequential Patterns Mining
• Explain sequential pattern mining and illustrate its uses • Apply the GSP algorithm on a given dataset • Illustrate the candidate generation in GSP • Compare the candidate pruning in GSP to candidate pruning in Apriori • Apply the SPADE algorithm on a given dataset • Apply the PrefixSpan algorithm on a given dataset • Compare GSP, SPADE and PrefixSpan algorithms for sequential pattern mining
Outlier Mining
• Illustrate with examples the types of outliers: global, contextual, and collective • Illustrate the use of grids in speeding up distance based outlier detection • Discuss in general terms the different methods of outlier mining: distance based-clustering based, classification based.
Stream Mining
• Describe and illustrate Bloom filter, Count-Min, Flajolet-Martin, and hyperloglog • Apply these methods to query real data streams
Spatial and Spatiotemporal Data Mining will not be included in the written exam

Note: No need to memorize equations and Algorithms. They will be given if needed.

With my best wishes.

Q) Consider the training examples shown in the Table below for a binary classification problem.

Maintenance	Persons	Luggage Boot	Safety	Buy
High	More	Medium	High	No
High	More	Medium	Medium	No
Medium	More	Medium	High	Yes
Low	5	Medium	High	Yes
Low	4	Big	High	Yes
Low	4	Big	Medium	No
Medium	4	Big	Medium	Yes
High	5	Medium	High	No
High	4	Big	High	Yes
Low	5	Big	High	Yes
High	5	Big	Medium	Yes
Medium	5	Medium	Medium	Yes
Medium	More	Big	High	Yes
Low	5	Medium	Medium	No

a Using the ID3 Algorithm, construct the full decision tree and motivate your answer by showing the steps.

b Evaluate the accuracy of your decision tree using the following testing examples.

Maintenance	Persons	Luggage Boot	safety	Buy
High	More	Big	High	Yes
High	More	Big	Medium	No
Medium	More	Big	High	Yes
Low	5	Big	High	No
Low	4	Medium	High	Yes
Low	4	Medium	Medium	No
Low	4	Medium	Medium	No
High	5	Big	High	Yes
High	4	Medium	High	Yes

Q) It is desired to partition customers into similar groups on the basis of their demographic profile. Which data mining problem is best suited to this task?

Q) Suppose in previous exercise, the merchant already knows for some of the customers whether or not they have bought widgets. Which data mining problem would be suited to the task of identifying groups among the remaining customers, who might buy widgets in the future?

Q) Suppose in previous exercise, the merchant also has information for other items bought by the customers (beyond widgets). Which data mining problem would be best suited to finding sets of items that are often bought together with widgets?

Q) Consider the training examples shown in the Table below for a binary classification problem.

Maintenance (M)	Persons (P)	Luggage Boot (L)	Safety (S)	Buy (B)
High	More	Medium	High	No
High	More	Medium	Medium	No
Medium	More	Medium	High	Yes
Low	5	Medium	High	Yes
Low	4	Big	High	Yes
Low	4	Big	Medium	No
Medium	4	Big	Medium	Yes
High	5	Medium	High	No
High	4	Big	High	Yes
Low	5	Big	High	Yes
High	5	Big	Medium	Yes
Medium	5	Medium	Medium	Yes
Medium	More	Big	High	Yes
Low	5	Medium	Medium	No

a Using the ID3 Algorithm, construct the full decision tree and motivate your answer by showing the steps.

M, P, L, S

$$E(S) = -\sum_{j=1}^{n} p_j \log(p_j)$$

$$E(B) = -\frac{5}{14}\log\left(\frac{5}{14}\right) - \frac{9}{14}\log\left(\frac{9}{14}\right) = 0.53 + 0.41 = 0.94$$

$$E\text{-}S(B \Rightarrow M) = \frac{5}{14}E(B|High) + \frac{4}{14}E(B|Medium) + \frac{5}{14}E(B|Low)$$

$$= \frac{-5}{14}\left[\frac{3}{5}\cdot\log\left(\frac{3}{5}\right) + \frac{2}{5}\log\left(\frac{2}{5}\right)\right] - \frac{4}{14}\left[1.0\right] - \frac{5}{14}\left[\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right]$$

$$= \frac{10}{14}\left(0.44 + 0.53\right) = 0.69 \leftarrow$$

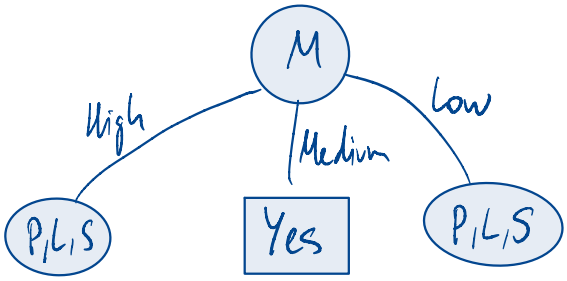
$$E\text{-}S(B \Rightarrow P) = \frac{4}{14}E(B|More) + \frac{6}{14}E(B|5) + \frac{4}{14}E(B|4) =$$

$$= -\frac{4}{14}\cdot\left[\log\frac{1}{2}\right] - \frac{6}{14}\left[\frac{4}{6}\log\frac{4}{6} + \frac{2}{6}\log\frac{2}{6}\right] - \frac{4}{14}\left[\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right]$$

$$= \frac{4}{14} + \frac{6}{14}\cdot\left[0.39 + 0.53\right] + \frac{4}{14}\left[0.31 + 0.5\right] = 0.91$$

$E\text{-}S(B \Rightarrow L) = \frac{7}{14}\left[\frac{4}{7}\log\frac{4}{7} + \frac{3}{7}\log\frac{3}{7}\right] + \frac{7}{14}\left[\frac{6}{7}\log\frac{6}{7} + \frac{1}{7}\log\frac{1}{7}\right] = 0'49 + 0'296 = 0'79$

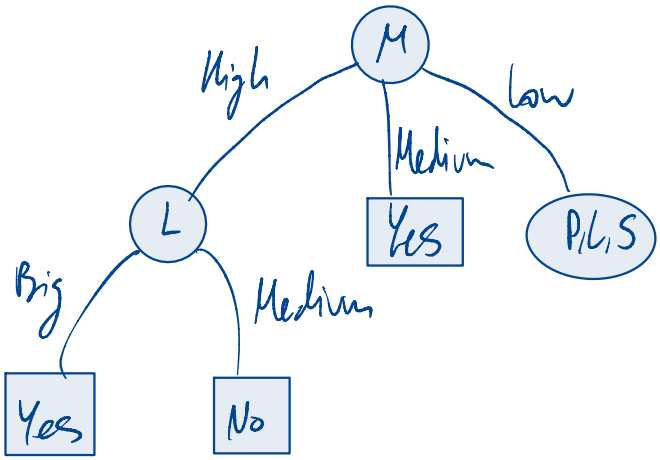
$E\text{-}S(B \Rightarrow S) = \frac{8}{14}\left[\frac{2}{8}\log\frac{2}{8} + \frac{6}{8}\log\frac{6}{8}\right] + \frac{6}{14}\left[\log\frac{1}{2}\right] = 0'46 + \frac{6}{14} = 0'84$



$E(B|High) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0'97$

$E\text{-}S(B \Rightarrow P|High) = \frac{2}{5}\cdot 0 - \frac{2}{5}\cdot\log\frac{1}{2} - \frac{1}{5}\cdot 0 = 0'4$

$E\text{-}S(B \Rightarrow L|High) = \frac{3}{5}\cdot 0 + \frac{2}{5}\cdot 0 = 0 \leftarrow$

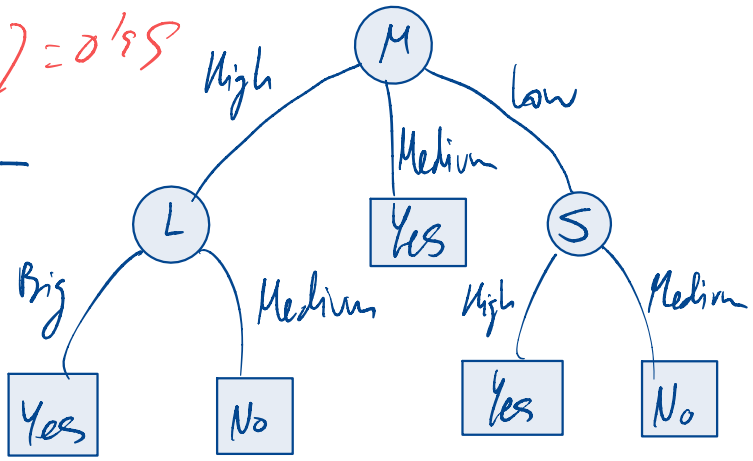


$E(B|low) = \frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5} = 0'97$

$E\text{-}S(B \Rightarrow P|low) = \frac{3}{5}\left[\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right] + \frac{2}{5}\left[-\log\frac{1}{2}\right] = 0'95$

$E\text{-}S(B \Rightarrow L|low) = \frac{2}{5} + \frac{3}{5}\left[-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right] = 0'95$

$E\text{-}S(B \Rightarrow S|low) = \frac{3}{5}\cdot 0 + \frac{2}{5}\cdot 0 = 0 \leftarrow$



b Evaluate the accuracy of your decision tree using the following testing examples.

Maintenance (M)	Persons (P)	Luggage Boot (L)	safety (S)	Buy	Pred
High	More	Big	High	Yes	Yes
High	More	Big	Medium	No	Yes ✗
Medium	More	Big	High	Yes	Yes
Low	5	Big	High	No	Yes ✗
Low	4	Medium	High	Yes	Yes
Low	4	Medium	Medium	No	No
Low	4	Medium	Medium	No	No
High	5	Big	High	Yes	Yes
High	4	Medium	High	Yes	No ✗

accuracy = $\frac{\text{correct}}{\text{total}} = \frac{6}{9} = 0.67$

Q) It is desired to partition customers into similar groups on the basis of their demographic profile. Which data mining problem is best suited to this task?

Clustering seems like a good option because the idea here is to group people that might have similar interests considering their age, location, family status, ...

Q) Suppose in previous exercise, the merchant already knows for some of the customers whether or not they have bought widgets. Which data mining problem would be suited to the task of identifying groups among the remaining customers, who might buy widgets in the future?

We could train a classification model using the already labeled data, and then use it to classify the remaining customers based on their attributes

Q) Suppose in previous exercise, the merchant also has information for other items bought by the customers (beyond widgets). Which data mining problem would be best suited to finding sets of items that are often bought together with widgets?

In this case, we are looking for association rules between items and widgets, so we should tackle this task using association pattern mining.

Q) If a source send n messages, with equal probability, the receiver needs to ask $\log_2 n$ yes/no questions to know the message. Explain the concept of Entropy in light of this sentence

Q) Explain with examples the concept of Entropy given its formula

$$E(S) = -\sum_{j=1}^{k} p_j \log_2(p_j)$$

Q) Describe a situation (or give an example) that motivates the use of GainRatio instead of Gain while inducing decision trees.

Q) Do you agree or disagree, and why:
a) holdout, as a method for classification model validation, is a special case of k-fold cross validation
b) leave-one-out, as a method for classification model validation, is a special case of k-fold cross validation
c) k-fold cross validation helps improving the classification accuracy of a model

Q) Describe how to assess the data skew using mean, median, mode summaries.

Q) Given two strings of lengths n and m, what are the time, space, and backtrace complexities of the recursive Edit distance algorithm ? Sketch (e.g., as pseduo code) a way to improve.

Q) Assume that Edit(X, Y) represents the cost of transforming the string X to Y . Show that Edit(X, Y) and Edit(Y , X) are the same, as long as the insertion and deletion costs are the same. (Edit(.,.) is the string edit distance function)

Q) Compute the edit distance between:
(a) ababcabc and babcbc and
(b) cbacbacba and acbacbacb.
For the edit distance, assume equal cost of insertion, deletion, or replacement.

Q) One of the following terms does not fit with the others. Choose it, and concisely explain why

1. Euclidean.
2. Binning
3. Manhattan.
4. L_∞-norm.

Q) If a source send n messages, with equal probability, the receiver needs to ask $\log_2 n$ yes/no questions to know the message. Explain the concept of Entropy in light of this sentence

Entropy is a measure of how many information a message carries. In this way of thinking, we can measure it by the amount of yes/no questions that need to answered to decipher the message. Basically, each message is either 0 or 1, so assuming they have each the same probability, $p(0)=p(1)=\frac{1}{2}$, then each new message possess maximum uncertainty, and the information of answering "is it 1?" is needed totally to decipher the whole message. We need thus n questions (one per bit), so we can measure it as $\log_2(n)$ — number of messages; 2 = # of answers to questions

If $p(0)=0$ and $p(1)$, no question is ever needed, and thus the entropy is 0.

In any case in-between the two extremes it is

$$E(S) = -p_1 \cdot \log_2(p_1) - p_0 \cdot \log_2(p_0)$$

p_1: probability of value 1 appearing

$\log_2(p_1)$: information provided by answering yes to "Is the message 1?"

Q) Explain with examples the concept of Entropy given its formula

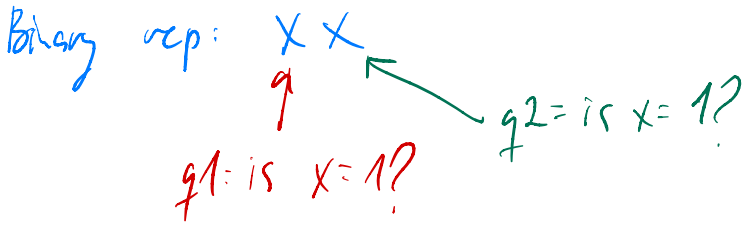
$$E(S) = -\sum_{j=1}^{k} p_j \log_2(p_j)$$

S carries no information: $S = \{1, 1, 1, 1, 1, \ldots, 1\}_n \rightarrow p_1 = 1$

$E(S) = -\sum 1 \cdot \log(1) = -\sum 1 \cdot 0 = -\sum 0 = 0 \rightarrow$ no info

S carries maximum information: $S = \{1, 2, 3, 4\} \rightarrow p_i = \frac{1}{4}$ $i = 1, \ldots, 4$

$E(S) = -4 \cdot \frac{1}{4} \log_2\left(\frac{1}{4}\right) = 2 \rightarrow$ we need to ask 2 questions



Q) Describe a situation (or give an example) that motivates the use of GainRatio instead of Gain while inducing decision trees.

If an attribute has very few duplicate values, the entropy restricted to one specific value gets very reduced, but it is unlikely that the value will appear in unseen records:

Name	Gender	Plays LOL
"John"	M	Yes
"Mike"	M	No
"James"	M	Yes
"Will"	M	Yes
"Sarah"	F	No
"Karen"	F	Yes

Unique values → E(S/Name) = 0 but not useful for prediction

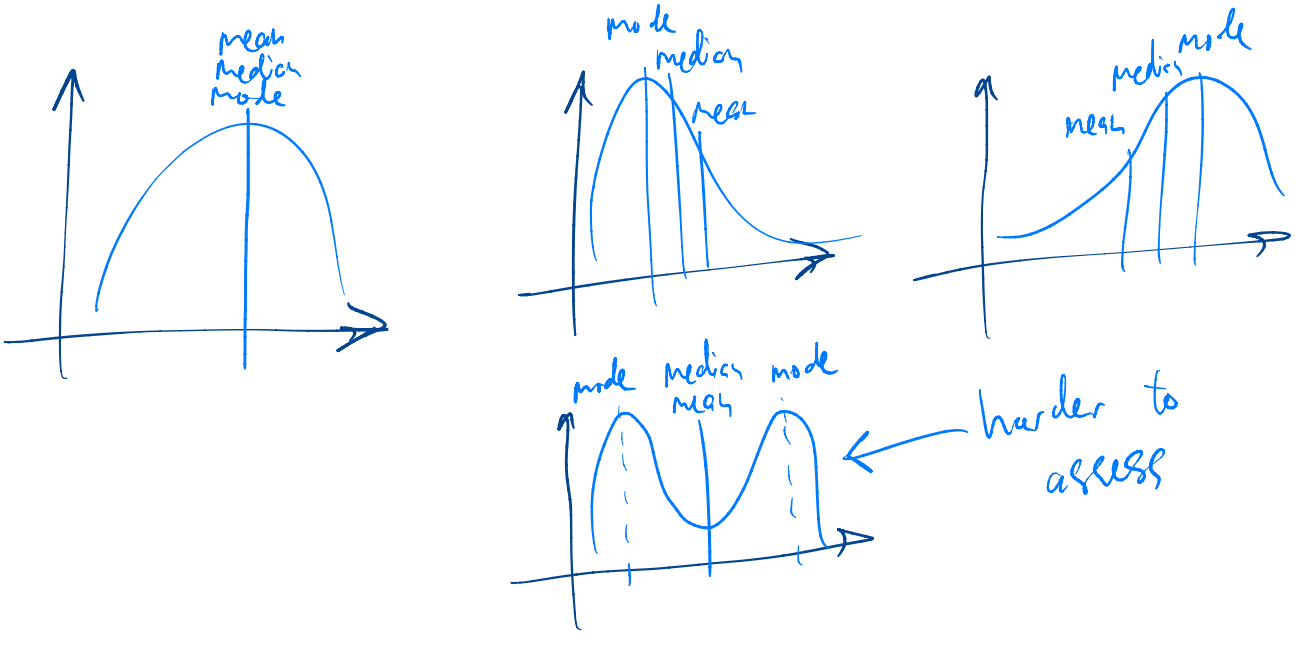
Q) Do you agree or disagree, and why:
a) holdout, as a method for classification model validation, is a special case of k-fold cross validation
b) leave-one-out, as a method for classification model validation, is a special case of k-fold cross validation
c) k-fold cross validation helps improving the classification accuracy of a model

A) False, in hold-out there is a training set and a validation set, which do not exchange roles later, as in K-fold CV.

B) True, if there are N triples, it is N-CV.

C) the goal of K-CV is to minimize the overfitting to the training set, by confronting the model with several validation sets. This way, k-CV is useful to better assess how the model will perform with unseen data, but this does not necessarily mean that the accuracy of the predictions will improve.

Q) Describe how to assess the data skew using mean, median, mode summaries.



Q) Given two strings of lengths n and m, what are the time, space, and backtrace complexities of the recursive Edit distance algorithm ? Sketch (e.g., as pseduo code) a way to improve.

We compute the max possibilities and the backtrace is at most as long as both strings, so

space: $O(n \cdot m)$

time: $O(n \cdot m)$

backtrace: $O(n+m)$

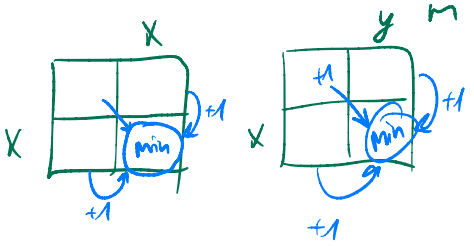
Example LOCK PICK with dynamic programming

M	P	I	C	K
L	1	1	1	1
O	1	2	2	2
C	1	2	2	3
K	1	2	3	(2)

mod(L→P)

mod(O→I)

$$M[i,j] = \min\begin{pmatrix} m[i-1,j-1] + s1[i] == s2[j], \\ m[i-1,j]+1, \\ m[i,j-1]+1 \end{pmatrix}$$



solution in $M[n,m]$

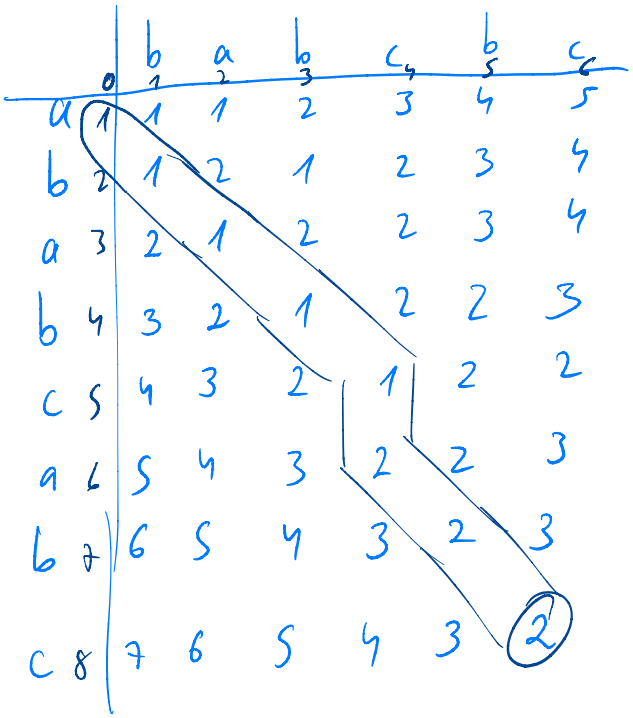
Q) Assume that Edit(X, Y) represents the cost of transforming the string X to Y . Show that Edit(X, Y) and Edit(Y , X) are the same, as long as the insertion and deletion costs are the same. (Edit(.,.) is the string edit distance function)

the transformation from X to Y is invertible:

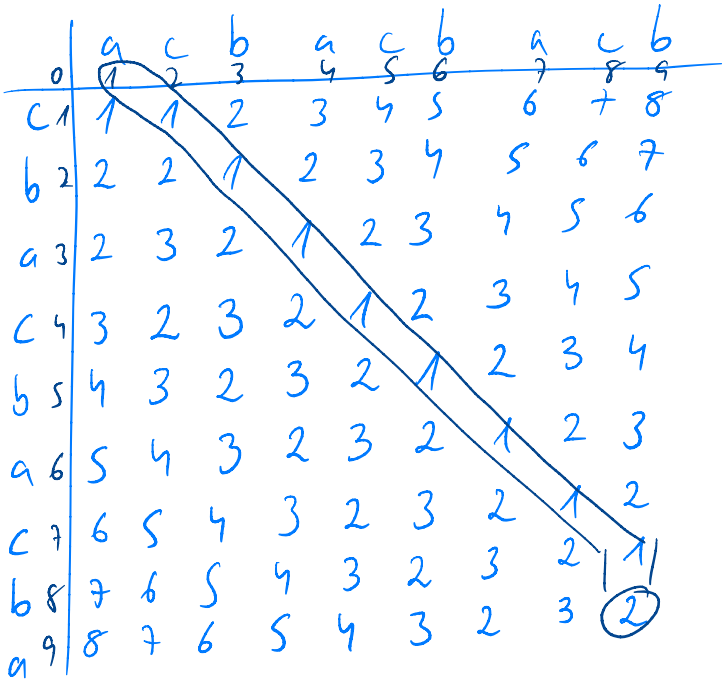
- modify $(\alpha \rightarrow \beta)$ is reverted by modify $(\beta \rightarrow \alpha)$
- remove(α) is reverted by insert(α)
- insert(α) is reverted by remove(α)

Thus the series of changes to go from X to Y can be reverted using the same amount of steps.

Q) Compute the edit distance between:
(a) ababcabc and babcbc and
(b) cbacbacba and acbacbacb.
For the edit distance, assume equal cost of insertion, deletion, or replacement.



Rem(a) → babcabc
Rem(a) → babcbc ✓



Insert(a) → acbacbacba
Rem(a) → acbacbacb ✓

Q) One of the following terms does not fit with the others. Choose it, and concisely explain why

1. Euclidean.
2. Binning
3. Manhattan.
4. L_∞-norm.

The intruder is "Binning" because while the rest of the terms are distance function, Binning is referring to a way of summarizing data by grouping it and aggregating the values.

Q) Consider the following database of student grades.

Student	Grade
Student 1	7
Student 2	6
Student 3	10
Student 4	14
Student 5	10
Student 6	19
Student 7	2
Student 8	21
Student 9	0
Student 10	17
Student 11	1
Student 12	6
Student 13	16
Student 14	5

a Use K-Medians to cluster them according to their grades into three cluster. Start with this set as initial cluster representatives {4,6,10}. Show the steps of your answer until convergence.

For K-medians, Dist(X_i, Y_j) is the Manhattan distance:

$$Dist(\overline{X_i}, \overline{Y_j}) = ||X_i - Y_j||_1.$$

Q) Explain the complexity of DBSCAN (you may look at the pseduo code)

Q) Give an example that motivates the need for probabilistic clustering.

Q) Consider the following database of student grades.

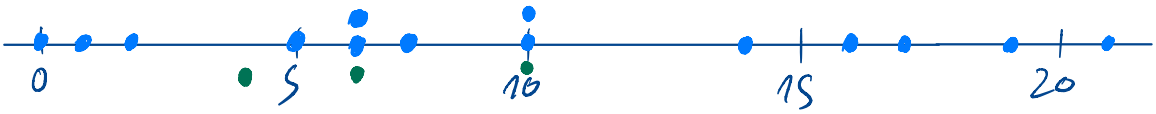
Student	Grade
Student 1	7
Student 2	6
Student 3	10
Student 4	14
Student 5	10
Student 6	19
Student 7	2
Student 8	21
Student 9	0
Student 10	17
Student 11	1
Student 12	6
Student 13	16
Student 14	5

a Use K-Medians to cluster them according to their grades into three cluster. Start with this set as initial cluster representatives {4,6,10}. Show the steps of your answer until convergence.

For K-medians, Dist(X_i, Y_j) is the Manhattan distance:

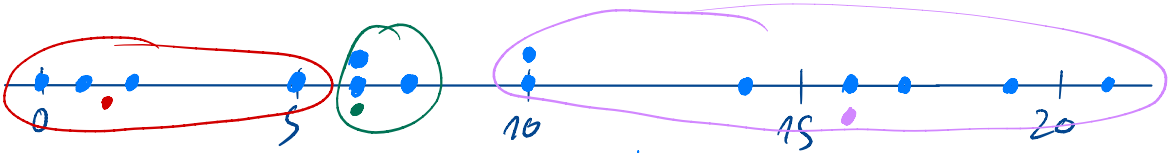
$$Dist(\overline{X_i}, \overline{Y_j}) = ||X_i - Y_j||_1.$$

• data • reps.



① Assign: 4 → 0,1,2,5
6 → 6,6,7
10 → 10,10,14,16,17,19,21

Recompute centers: 4 → 15
6 → 6
10 → 16



② Assign: 15 → 0,1,2
6 → 5,6,6,7,10,10
16 → 14,16,17,19,21

Rec. centers: 15 → 1
6 → 6'S
16 → 17

③ Assign: 1 → 0,1,2
6'S → 5,6,6,7,10,10
17 → 14,16,17,19,21

Rec. centers: 1 → 1
6'S → 6'S
17 → 17

→ converged

Q) Explain the complexity of DBSCAN (you may look at the pseduo code)

Each point is visited once.

Each time we visit a new point:

- we compute its ε-neighbourhood

Thus the complexity is $O(n \cdot h(n))$

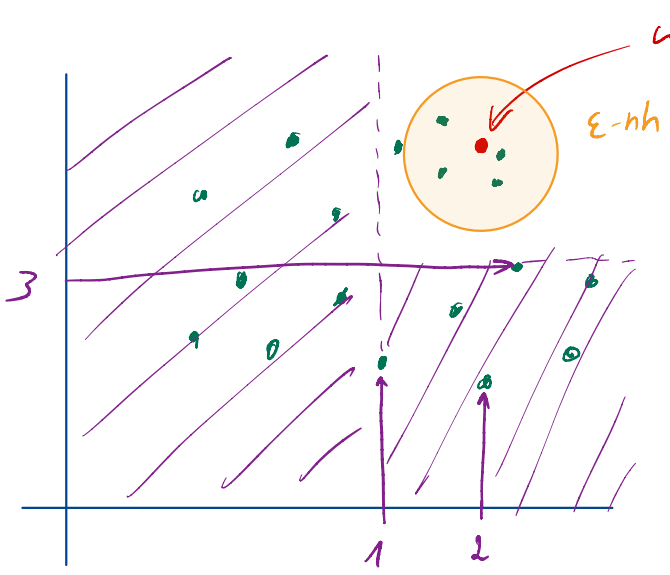
where $h(n)$ is the cost of computing the ε-neighbourhoods

The most naive approach is $h(n) = n$

So $O(n^2)$

If an spatial index is defined, dividing the space into a grid, it is possible to reduce $h(n) = \log n \rightarrow O(n \log n)$

Idea



- using 3 comparisons we are able to discar most of the data points
- In general, it is similar to a binary search $\rightarrow \log n$

Q) Give an example that motivates the need for probabilistic clustering.

Probabilistic clustering is specially interesting when we don't want to categorize records into exclusive categories, but to allow some kind of intertwinement between clusters.

Example

We want to automatically add tags to videos.
If we have a probabilistic clustering, we can assess the level of membership of a new video to each fuzzy cluster using some characteristics of the video, as the tags defined by the user, the duration, location, ...
Then, we can take the n top clusters and add those labels that the video doesn't have but should because of its similarity.

Q) Given the following database of car theft incidents:

Example No	Color	Type	Origin	Stolen ?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	White	Sports	Domestic	No
5	White	Sports	Imported	Yes
6	White	SUV	Imported	No
7	White	SUV	Imported	Yes
8	White	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

a Use the naive Bayes classifier to predict whether or not the following car will be stolen (color= Red, type=SUV, Origin= Domestic).
b Use the naive Bayes classifier to predict whether or not the following car will be stolen (color= Yellow, type=SUV, Origin= Domestic). (10 pts)

Naive Bayes classifier

$$P(X \vee C_i) = \prod_{k=1}^{n} P(x_k \vee C_i)$$

Q) Explain the conditions for an ensemble of classifiers to yields more accurate predictions than its individuals.

Q) Which of the following is true in an itemset lattice?

1. Every subset of a frequent itemset is also frequent.
2. Every superset of an infrequent itemset is also infrequent.
3. The lattice size is $2^{|U|}$, where U is the itemset.
4. The support of an itemset is always greater than or equal to the support of any of its supersets.

Q) Given the following database of car theft incidents:

Example No	Color	Type	Origin	Stolen ?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	White	Sports	Domestic	No
5	White	Sports	Imported	Yes
6	White	SUV	Imported	No
7	White	SUV	Imported	Yes
8	White	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

a Use the naive Bayes classifier to predict whether or not the following car will be stolen (color= Red, type=SUV, Origin= Domestic).
b Use the naive Bayes classifier to predict whether or not the following car will be stolen (color= Yellow, type=SUV, Origin= Domestic). (10 pts)

Naive Bayes classifier

$$P(X \vee C_i) = \prod_{k=1}^{n} P(x_k \vee C_i)$$

(a)

$P(\text{Yes} \mid \text{Red} \wedge \text{SUV} \wedge \text{Domestic}) = P(\text{Yes}) \cdot P(\text{Red} \mid \text{Yes}) \cdot P(\text{SUV} \mid \text{Yes}) \cdot P(\text{Dom} \mid \text{Yes})$

$= \frac{1}{2} \cdot \frac{3}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} = \frac{6}{250} = \frac{3}{125}$

$P(\text{No} \mid \text{Red} \wedge \text{SUV} \wedge \text{Dom}) = P(\text{No}) \cdot P(\text{Red} \mid \text{No}) \cdot P(\text{SUV} \mid \text{No}) \cdot P(\text{Dom} \mid \text{No})$

$= \frac{1}{2} \cdot \frac{2}{5} \cdot \frac{3}{5} \cdot \frac{3}{5} = \frac{18}{250} = \frac{9}{125}$

Prediction: No

(b) $P(\text{Yes} \mid \text{Yellow} \wedge \text{SUV} \wedge \text{Dom}) = \frac{1}{2} \cdot \frac{1}{6} \cdot \frac{1}{5} \cdot \frac{2}{5} = \frac{2}{300}$

$\frac{0+1}{5+1} \rightarrow \frac{1}{6}$

$P(\text{No} \mid \text{Yellow} \wedge \text{SUV} \wedge \text{Dom}) = \frac{1}{2} \cdot \frac{1}{6} \cdot \frac{3}{5} \cdot \frac{3}{5} = \frac{9}{300}$

Prediction: No

Q) Explain the conditions for an ensemble of classifiers to yields more accurate predictions than its individuals.

Q) Which of the following is true in an itemset lattice?

1. Every subset of a frequent itemset is also frequent.
2. Every superset of an infrequent itemset is also infrequent.
3. The lattice size is $2^{|U|}$, where U is the itemset.
4. The support of an itemset is always greater than or equal to the support of any of its supersets.

1- True
2- True
> Apriori property

3- True if we count the empty set, $2^{|U|}-1$ if not.

4- True, equivalent to 2-

Q) Consider the transaction database in the table below:

tid	items
1	a, c, d, e
2	a, d, e, f
3	b, c, d, e, f
4	b, d, e, f
5	b, e, f
6	c, d, e
7	c, e, f
8	d, e, f

a) Determine all frequent patterns and maximal patterns at support counts of 3, 4, and 5.

b) Determine the confidence of the rules {a} ⇒ {f }, and {a, e} ⇒ {f } in this transaction database

c) Show the candidate itemsets and the frequent itemsets in each level-wise pass of the Apriori algorithm. Assume a minimum support count of 2.

Q) Consider the 1-dimensional data set with 10 data points {1, 2, 3, . . . 10}.

a) Show three iterations of the k-means algorithms when k = 2, and the random seeds are initialized to {1, 2}.

b) Repeat previous exercise with an initial seed set of {2, 9}. How did the different choice of the seed set affect the quality of the results?

Q) Consider a 1-dimensional data set with three natural clusters. The first cluster contains the consecutive integers {1 . . . 5}. The second cluster contains the consecutive integers {8 . . . 12}. The third cluster contains the data points {24, 28, 32, 36, 40}. Apply a k-means algorithm with initial centers of 1, 11, and 28. Does the algorithm determine the correct clusters?

Q) If the initial centers are changed to 1, 2, and 3, does the algorithm discover the correct clusters? What does this tell you?

Q) The Apriori algorithm builds C_{k+1} by combining pairs in F_k that have k-1 items in common. Prove that with this strategy $F_{k+1} \subseteq C_{k+1}$

Credits

Some questions are copied from

- Charu C.Aggarwal. Data Mining The Textbook, Springer.
- Jiawei Han, Micheline Kamber and Jian Pe, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012.

Q) Consider the transaction database in the table below:

tid	items
1	a, c, d, e
2	a, d, e, f
3	b, c, d, e, f
4	b, d, e, f
5	b, e, f
6	c, d, e
7	c, e, f
8	d, e, f

a) Determine all frequent patterns and maximal patterns at support counts of 3, 4, and 5.

b) Determine the confidence of the rules {a} ⇒ {f }, and {a, e} ⇒ {f } in this transaction database

c) Show the candidate itemsets and the frequent itemsets in each level-wise pass of the Apriori algorithm. Assume a minimum support count of 2.

(a)

Item	#
a	2
b	3
c	4
d	5
e	8
f	6

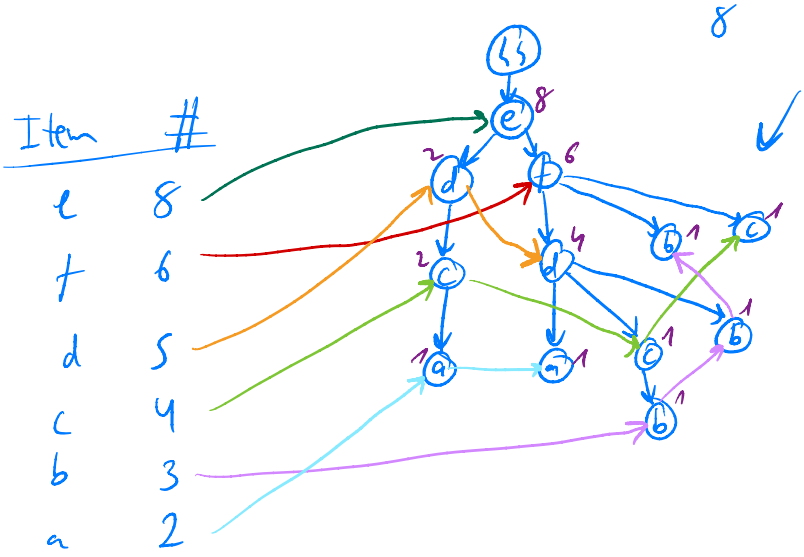
reorder →

e	8
f	6
d	5
c	4
b	3
a	2

reorder DB →

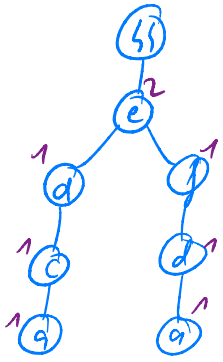
TID	Items
1	e, d, c, a
2	e, f, d, a
3	e, f, d, c, b
4	e, f, d, b
5	e, f, b
6	e, d, c
7	e, f, c
8	e, f, d

Item	#
e	8
f	6
d	5
c	4
b	3
a	2



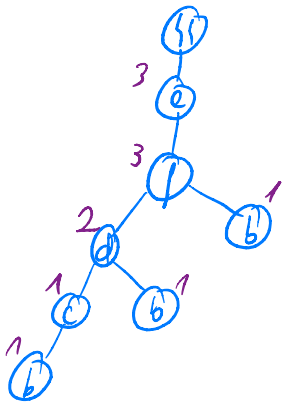
✓ FP-Tree

a] Cond. Patt. Base : { [{e,d,c}:1], [{e,f,d}:1] }



No freq. patterns

b] Cond. Patt. Base : {[{e,f,d,c}:1], [{e,f,d}:1], [{e,f}:1]}



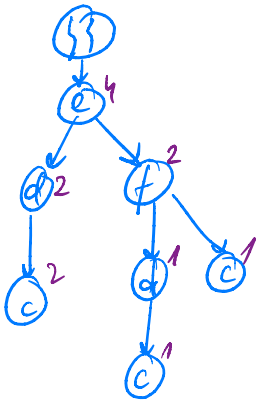
No frequent patterns for minsup = 4,5

For 3: {e,f,b}, {e,b}, {f,b}

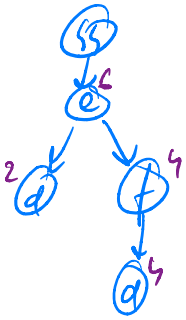
c] C.P.B: {[{e,d}:2], [{e,f,d}:1], [{e,f}:1]}

For 3: {e,c}

For 4: {e,c}



d] CPB: {[{e}:2], [{e,f}:4]}



For 3: {e,f,d}, {e,d}, {f,d}

For 4: {e,f,d}, {e,d}, {f,d}

For 5: {e,d}

f] CPB = {[{e}:6]}

For 3: {e,f}

For 4: {e,f}

For 5: {e,f}

For 3: {e}, {f}, {d}, {c}, {b}, {e,b}, {f,b}, {e,d}, {f,d}, {e,c}, {e,f}, {e,f,d}, {e,f,b}

For 4: {e}, {f}, {d}, {c}, {e,c}, {e,f}, {e,d}, {f,d}, {e,f,d}

For 5: {e}, {f}, {d}, {e,d}, {e,f}

(b) $conf(\{a\} \rightarrow \{f\}) = \frac{sup(\{a,f\})}{sup(\{a\})} = \frac{1}{2}$

$conf(\{a,e\} \rightarrow \{f\}) = \frac{sup\{a,e,f\}}{sup\{a,e\}} = \frac{1}{2}$

(c)

tid	items
1	a, c, d, e
2	a, d, e, f
3	b, c, d, e, f
4	b, d, e, f
5	b, e, f
6	c, d, e
7	c, e, f
8	d, e, f

C1

Itemset	#
a	2
b	3
c	4
d	5
e	8
f	6

F1

IS	#
a	2
b	3
c	4
d	5
e	8
f	6

C2

IS	#
~~a,b~~	~~0~~
~~a,c~~	~~1~~
a,d	2
a,e	2
~~a,f~~	~~1~~
~~b,c~~	~~1~~
b,d	2
b,e	3
b,f	3
c,d	3
c,e	4
c,f	2
d,e	5
d,f	3
e,f	6

F2

IS	#
a,d	2
a,e	2
b,d	2
b,e	3
b,f	3
c,d	3
c,e	4
c,f	2
d,e	5
d,f	3
e,f	6

C3

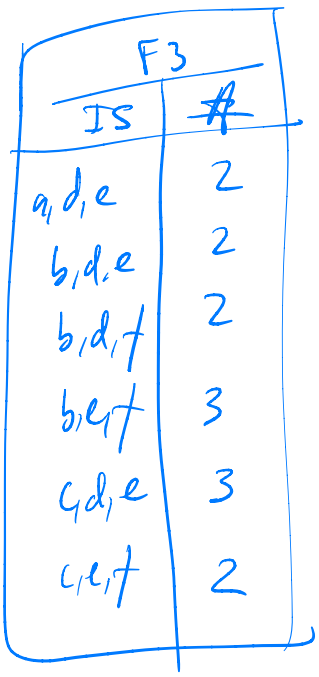
IS	#
a,d,e	2
~~a,d,b~~	
~~a,d,c~~	
~~a,d,f~~	
~~a,e,b~~	
~~a,e,c~~	
~~a,e,f~~	
b,d,e	2
b,d,f	2
~~b,d,c~~	
~~b,e,c~~	
b,e,f	3
~~b,f,c~~	
c,d,e	3
~~c,d,f~~	~~1~~
c,e,f	2

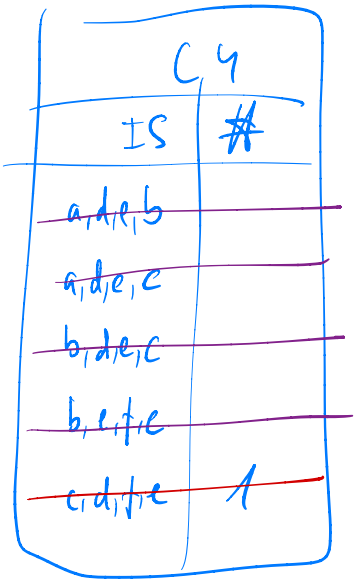
F3

IS	#
a,d,e	2
b,d,e	2
b,d,f	2
b,e,f	3
c,d,e	3
c,e,f	2

— not freq.

— prune


F3
IS
#
a, d, e
2
b, d, e
2
b, d, f
2
b, e, f
3
c, d, e
3
c, e, f
2


C4
IS
#
a, d, e, b
a, d, e, c
b, d, e, c
b, e, f, e
c, d, f, e
1

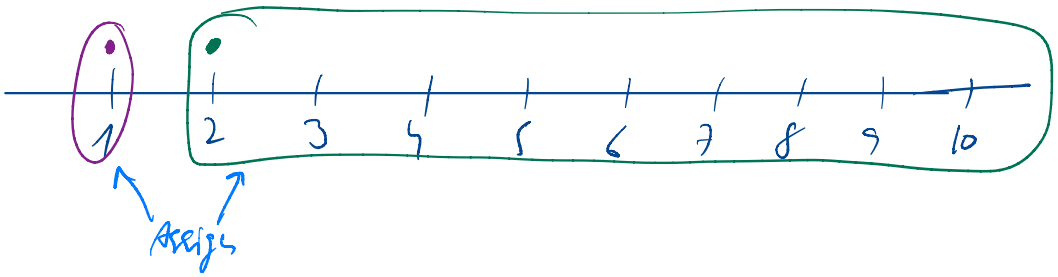
Q) Consider the 1-dimensional data set with 10 data points {1, 2, 3, . . . 10}.

a) Show three iterations of the k-means algorithms when k = 2, and the random seeds are initialized to {1, 2}.

b) Repeat previous exercise with an initial seed set of {2, 9}. How did the different choice of the seed set affect the quality of the results?

a)

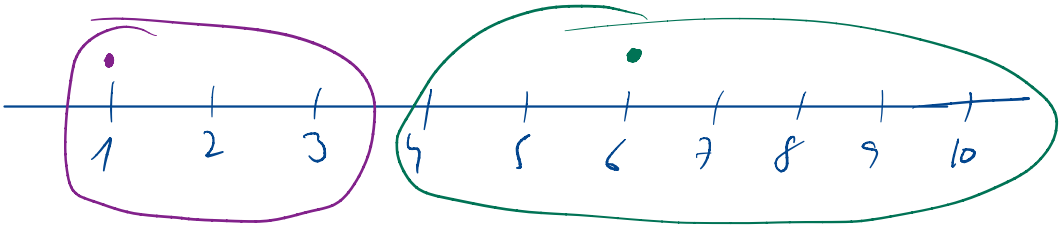
1



Recompute $c_1 = 1$

$$c_2 = \frac{2+\dots+10}{9} = \frac{12+12+12+12+6}{9} = \frac{54}{9} = 6$$

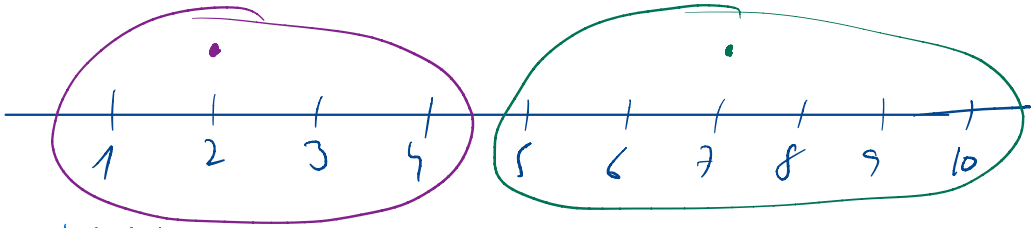
2



$$c_1 = \frac{1+2+3}{3} = 2$$

$$c_2 = \frac{4+\dots+10}{7} = \frac{14+14+14+7}{7} = \frac{3\cdot 2\cdot 7+7}{7} = 7$$

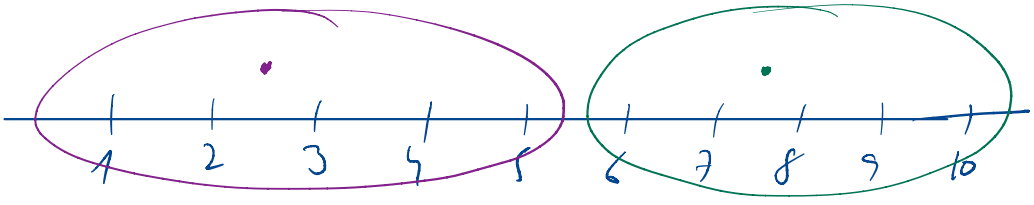
3



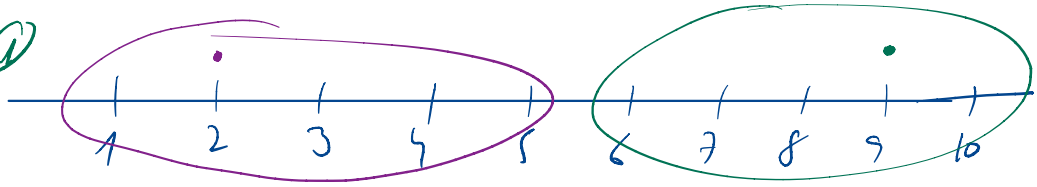
$$c_1 = \frac{1+2+3+4}{4} = \frac{10}{4} = 2{,}5$$

$$c_2 = \frac{5+\dots+10}{6} = \frac{15+15+15}{6} = \frac{3\cdot 3\cdot 5}{3\cdot 2} = \frac{15}{2} = \frac{7}{5}$$

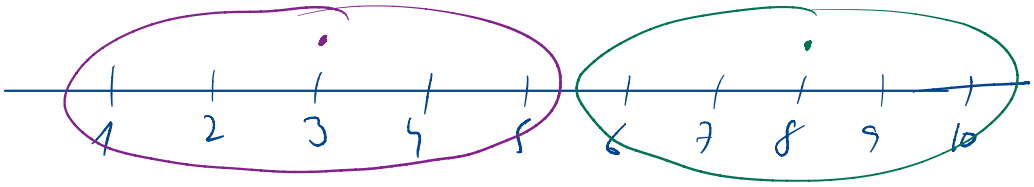
Final result



(b) (1)



$c_1 = 3$ $c_2 = 8$



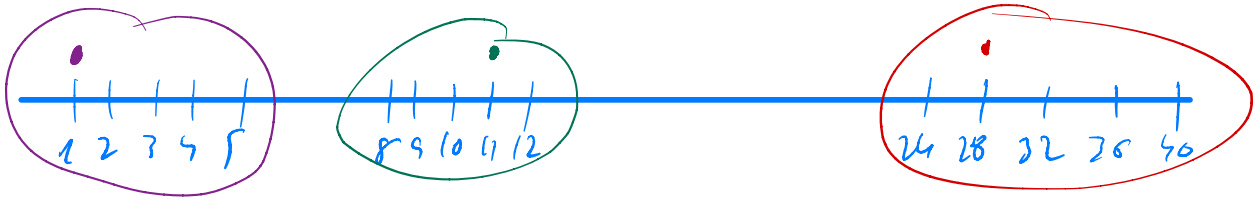
$c_1 = 3$ $c_2 = 8$

⋮

stable

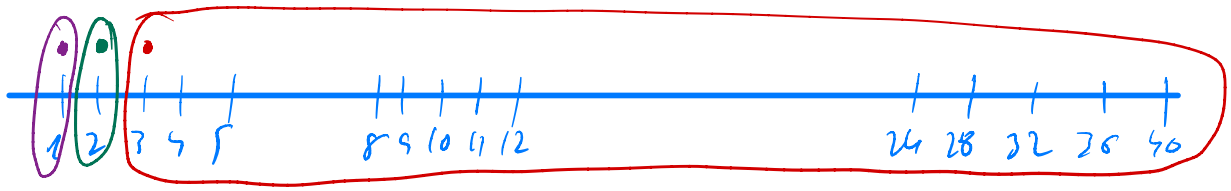
The resulting clusters are the same, but convergence is faster with the second initial seed.

Q) Consider a 1-dimensional data set with three natural clusters. The first cluster contains the consecutive integers {1 . . . 5}. The second cluster contains the consecutive integers {8 . . . 12}. The third cluster contains the data points {24, 28, 32, 36, 40}. Apply a k-means algorithm with initial centers of 1, 11, and 28. Does the algorithm determine the correct clusters?

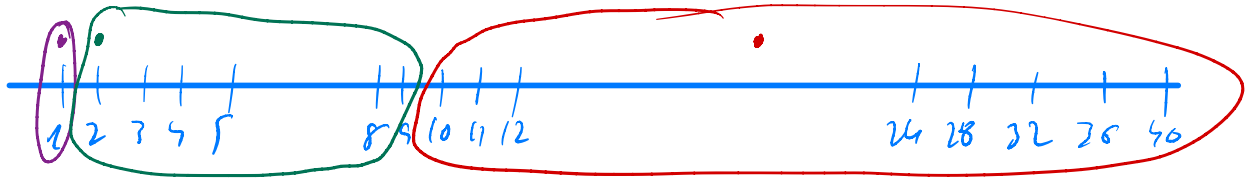


$c_1 = 3$ $c_2 = 10$ $c_3 = 32$ OK

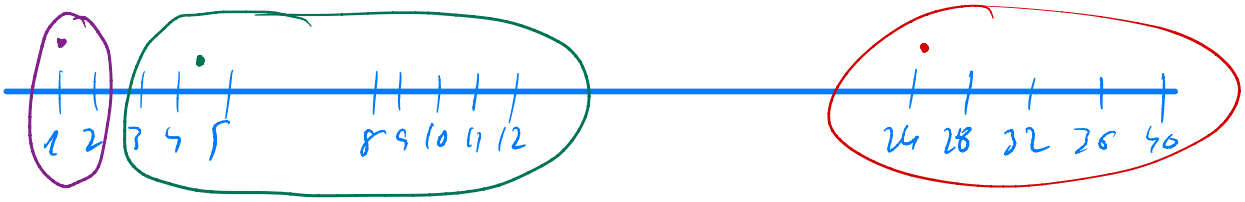
Q) If the initial centers are changed to 1, 2, and 3, does the algorithm discover the correct clusters? What does this tell you?



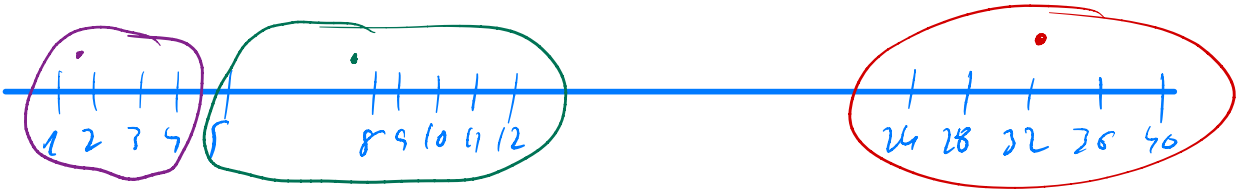
$c_1 = 1$ $c_2 = 2$ $c_3 = \frac{3+4+5+8+\ldots+12+24+\ldots+40}{13} = 17{,}1$



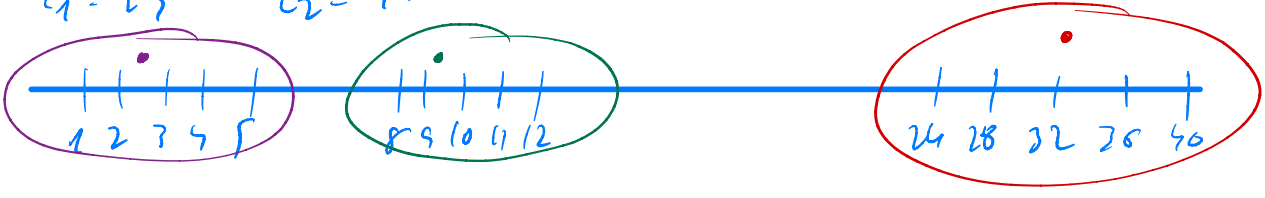
$c_1 = 1$ $c_2 = \frac{2+3+4+5+8}{5} = 4{,}4$ $c_3 = \frac{10+11+12+24\ldots}{8} = 24{,}1$



$c_1 = 1{,}5$ $c_2 = 7{,}75$ $c_3 = 32$



$c_1 = 2{,}5$ $c_2 = 9{,}16$



It found the correct clusters, but much slower.
The initial seed is important for performance, but usually the convergence will arrive eventually.

Q) The Apriori algorithm builds C_{k+1} by combining pairs in F_k that have k-1 items in common. Prove that with this strategy $F_{k+1} \subseteq C_{k+1}$

Suppose the contrary: $F_{k+1} \not\subseteq C_{k+1} \Rightarrow \exists$ item $\in F_{k+1}$ such that item $\notin C_{k+1}$

But, if item $\in F_{k+1}$, it is frequent, and by the apriori property, every subset is also frequent.

Say item $= \{a_1, \dots, a_{k+1}\}$, then

$x = \{a_1, \dots, a_k\}$ is frequent $\rightarrow x \in F_k$

and

$y = \{a_2, \dots, a_{k+1}\}$ is frequent $\rightarrow y \in F_k$

But x and y share k-1 elements, so they can be merge to form item. Thus item $\in C_{k+1}$ # but this is a contradiction.

Thus, it follows that $F_{k+1} \subseteq C_{k+1}$