CPSC 100 2016W2: Practice Midterm

Data Representation

Convert the binary number 011111111111111 to hexadecimal and decimal

Hexadecimal: 3FFF
Decimal: 16383

Convert the hexadecimal number 0x12CD3 to binary.

0001 0010 1100 1101 0011

Convert the hexadecimal number 0x45 to decimal.

69

Convert the decimal number 1034 to binary.

0100 0000 1010

Convert decimal number 132 to hexadecimal.

0x84

DNA has four nucleotides (A, C, T, and G) (i.e., a base-4 system). Codons are “words” of three nucleotides that code for amino acids. For example AAA codes for an amino acid called lysine. What is the maximum number of amino acids that could be represented if we use four nucleotides and each codon can only be made up of three nucleotides?

64 amino acids ($4^3 = 64$)

There are just 20 amino acids. In a base-4 system (i.e., a system where each “digit” can be A, C, T, or G), what is the smallest number of digits that I need per position, in order to be able to represent all 20 amino acids using three-nucleotide codons?

3 digits ($4^3 = 64$)

CPSC 100 2016W2: Practice Midterm

What is rasterization?

Rasterization is when we take something that is in vector form and convert it to pixels so that we can save it in bitmap form.

Henry's friend has given him the dimensions (the number of rows and columns) of an image, and the number of bytes used to represent each pixel. How can Henry use that information to calculate the approximate size of the file?

Note: This calculation would not get you the exact file size since files also include metadata (data about the file (e.g., header information in a bitmap image file).

# pixels width x # pixels height x # bytes used to represent a pixel = file size

How does the blur filter work?

Each pixel value is replaced with the average of the pixel and the pixels surrounding it.

CPSC 100 2016W2: Practice Midterm

Apriori

Show all the steps to finding the frequent itemsets with >50% support.

Transaction	Items
T1	Coffee, Tea, Juice, Water
T2	Tea, Juice,
T3	Coffee, Juice

Frequent itemset of size 1

Number of times each item appears:

- Coffee: 2
- Tea: 2
- Water: 1
- Juice: 3

Eliminate itemsets with less than 50% support which leaves:

- Coffee
- Tea
- Juice

Frequent itemset of size 2

Number of times each item appears:

- {Tea, Juice}: 2
- {Coffee, Tea}: 1
- {Coffee, Juice}: 2

Eliminate itemsets with less than 50% support which leaves:

- {Tea, Juice}
- {Coffee, Juice}

Frequent itemset of size 3

Number of times each item appears:

- {Coffee, Tea, Juice}: 1

Eliminate itemsets with less than 50% support which leaves:

- Nothing

Frequent itemsets with > 50% support

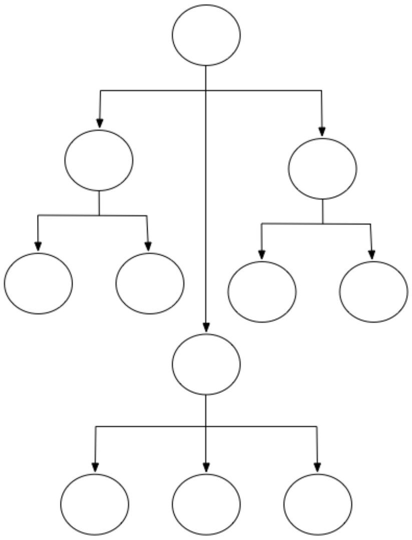
- Coffee
- Tea
- Juice
- {Tea, Juice}
- {Coffee, Juice}

Explain how the itemsets with ≥50% support would change if we added T4: Lemonade in the original table.

No effect as lemonade would have been eliminated within the first round.

Decision Trees

Based on the following diagram, determine how many of the following there are:



- Nodes: 11
- Edges: 10
- Leaves: 7
- Depth: 2
- Parent: 4

CPSC 100 2016W2: Practice Midterm

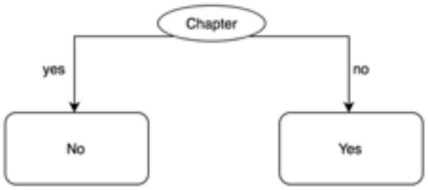
You are procrastinating from studying for your CPSC 100 midterm and you are trying to decide whether you should study, so you make a decision tree to help you determine if you should study for each of the five chapters of the textbook.

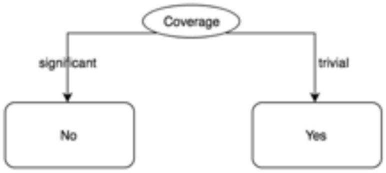
Did I read the chapter?	How well do I understand the text?	How long will it take me to review?	What impact will the chapter have on the exam?	How difficult are the questions?	Should I study?
Yes	Confident	Long	Significant	Hard	No
No	Fairly well	Short	Trivial	Medium	Yes
Yes	Confused	Long	Significant	Medium	No
Yes	Confused	Medium	Significant	Hard	No
No	Confident	Medium	Trivial	Easy	Yes

For each attribute (i.e., Read Chapter or Not, Undersanding of Text, ..., Question Difficulty), what is the overall entropy if we split on that attribute?

Read Chapter or Not	Understanding of Text	Review Time	Chapter Impact	Question Difficulty
0	2	2	0	2

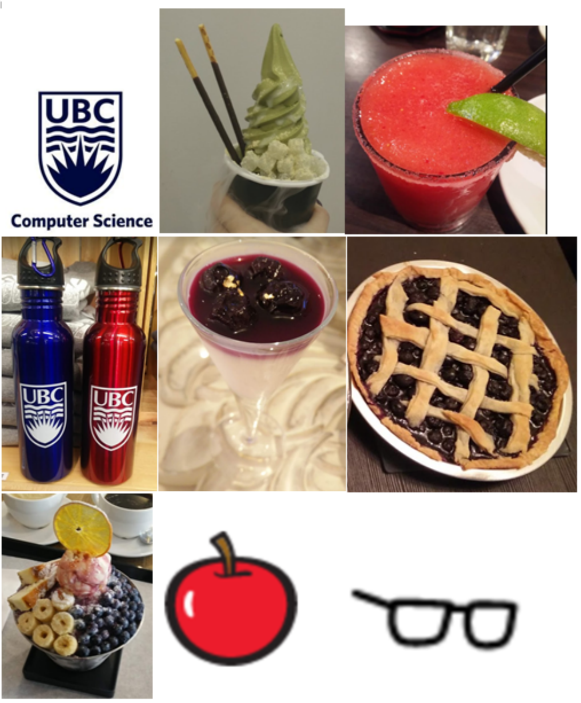
Draw the tree(s) that split on attributes with the greatest reduction in overall entropy.





Clustering

Given the following 9 items, how would you group these items (what measure of quality would you use)? How many data points and what are the data dimensions based on what how you clustered these items?



Possible answers:

- Group by food and drinks vs. UBC material (9 data points, 3 clusters)
- Group by colours (pictures with some red, if not red group with black (9 data points, 2 clusters)
- Group by items which are round or contains circles vs. no circles (9 data points, 2 clusters)

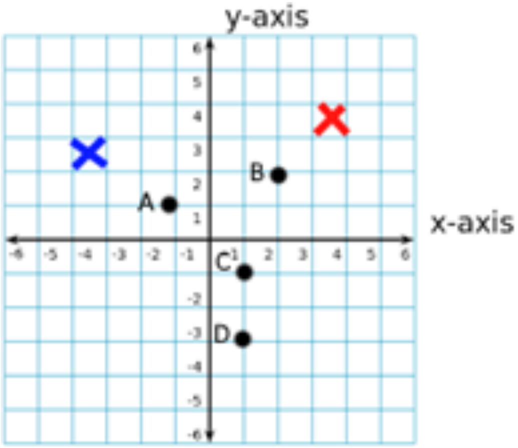
What are the benefits of clustering and how does it help in data mining?

Possible answers:

- Explore data for any hidden patterns or correlations which can guide in decision making (e.g., Netflix, cancer tumours)
- Organize data, useful classification or targeted messages (e.g., online ads, species, types of energy consumers, shopping)
- Reduce data complexity (e.g., heat maps, use few points in a cluster to represent the whole cluster, filters for categories and folders)

CPSC 100 2016W2: Practice Midterm

Below is a graph with data points A, B, C, and D and two centroids (blue and red X's).



What will happen in the following 2 steps (cluster assignment and move centroid) in the k-means clustering algorithm, given the following table?

Data point	Distance to Red Centroid (4,4)	Distance to Blue Centroid (-4,3)
A (-1,1)	5.8	3.6
B (2,2)	2.8	6.1
C (1,-1)	5.8	6.4
D (1,-3)	7.6	7.8

1. **Cluster assignment:** Which clusters will these data points be assigned to?
   Assign each point to the closest centroid and put it in the cluster of that centroid

   The blue cluster will have point A.

   The red cluster will have points B, C, and D.

2. **Move centroid**: Given the following data, what coordinates will the red and blue centroid move to? State which of the following calculations should be performed (average or median and for which points. Also, state which of the following will determine where the red centroid, and blue centroid will move to.

Calculation	New Centroid

1) Average of points A & D	**(0,-1)**
	x-coord = (-1 +1)/2 = **0**
	y-coord = (1+-3)/2 = **-1**

2) Median of points A & D	**(0,-1)** X-coord = -1, 1, even number so take the average (-1+1)/2 = **0** y-coord = -3,1. (-3+1)/2 = **-1**

3) Average of points B & C & D	**(1.3,-0.7)**
	x-coord = (2+2+1)/3 = **1.3**
	y-coord = (2+-1+-3)/3 =**-0.67**

4) Median of points B & C & D	**(1,-1)**
	x-coord = 1,1,2 median is **1**
	y-coord = -3,-1,2, median is **-1**

5) Average of points B & C	**(1.5,0.5)**
	x-coord = (2+1)/2 = **1.5**
	y-coord = (2+-1)/2 = **0.5**

6) Median of points B & C	**(1.5,0.5)** X-coord = 1,2 even number so take the average (1+2)/2 = **1.5** y-coord = -1,2. (-1+2)/2 = **0.5**
7) Average of point A	**(-1,1)**
8) Median of point A	**(-1,1)**

The two calculations used should be numbers ___________ and _____________.
Red centroid will move to ___________.
Blue centroid will move to ___________.

The two calculations used should be numbers ____3)_______ and ______7)______.
Red centroid will move to ______3)_____.
Blue centroid will move to ______7)_____.