COMP9318 – Unofficial Practice Exam (By Daniel Hocking)

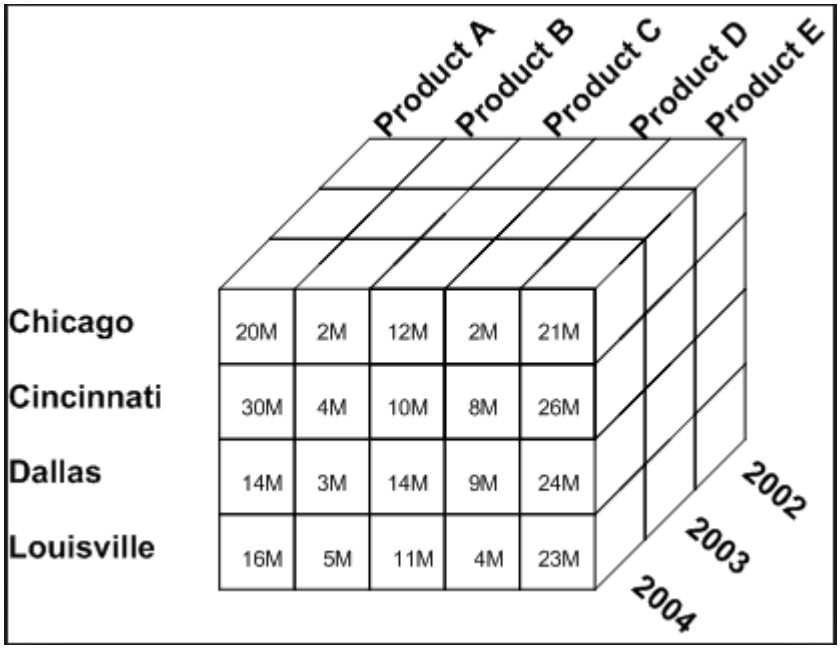
Part 1

1. Briefly compare and contrast **data mining** and **KDD**, what are they, and how to they relate to each other?
2. What are some of the key challenges that **data mining** tries to solve (that is, what do you use data mining to achieve)?

Part 2

3. What is a **data mart** and how does it compare to a **data warehouse**?
4. There are **four key characteristics** of a data warehouse, list them, and provide some further detail about one of them?
5. What are the four most common **OLAP operations**?
6. You are given the following dimensions:
    a. Product: ALL, product
    b. Location: ALL, country, store
    c. Time: ALL, quarter, month, day

    How many cuboids would the complete cube contain? Draw a lattice of the cuboids formed using the lowest levels of the given dimensions?
7. Given the following data cube:



    What would be the result of **pivot on (city, year)**?
8. Using the BUC-SR algorithm compute the data cube for the following input:

	A	B	M
r1	1	1	25
r2	2	3	40
r3	2	1	50
r4	3	1	60
r5	3	3	80

Part 3

9. What is **singular vector decomposition** used for?
10. Determine whether the vector **w** lies in the span of the vectors $\mathbf{v}_i$?

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 14 \\ 3 \\ 15 \end{bmatrix}$$

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 4 \\ 7 \\ 3 \end{bmatrix}$$

Part 4

11. Compare and contrast some methods that can be used to deal with **noisy data**?
12. Use the MinMaxScaler to normalise the following input data:
    5, 2, 7, 4, 4, 3, 12, 9
13. Find the **optimal binning** of the following data (show the table generated by the DP optimal binning algorithm):
    4, 5, 7, 12, 14, 19 and B = 3
14. What solution would the **MaxDiff** algorithm find?

Part 5

15. You have been informed that a model you developed has a training error of 0.1 and a testing error of 0.5, the model is predicting between 4 classes that have approx. equal frequency in the sample data.
    - Is this a good result?
    - What might this result indicate?
    - Briefly discuss ways to help resolve and problems you have identified
16. Construct a decision tree using the **ID3 algorithm** on the following data:

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

17. Calculate the **Gini index** of the outlook feature in the above data table to find the optimal way to split the data into two groups.

18. Apply **Naïve Bayes algorithm** to the following data:

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

What is the probability of playing golf given the following is true of today: Sunny, Hot, Normal, False

19. **SVM** attempts to find a hyperplane that separates the input data with a maximal margin, explain what can be done in cases where the data can't be separated in this way?
20. **Logistic Regression** is a discriminative model, give an example of a generative model and explain briefly what differentiates them.

Part 6

21. Given the following input data calculate the **Jaccard coefficient** for each group of two people:

Name	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
Bob	Y	Y	Y	N	Y	N
Jack	Y	Y	N	N	N	Y
Sam	N	Y	N	Y	Y	N

22. Given the following vector:
    [1, 2, 2, 4]
    - What is the L_1 norm?
    - What is the L_2 norm?
23. Discuss the advantages and disadvantages of the **K-means algorithm**, is there an alternative that works better?
24. Show steps as you merge clusters using the **group average** similarity metric on the following similarity matrix:

	P1	P2	P3	P4	P5
P1	1	0.6	0.9	0.35	0.2
P2		1	0.55	0.4	0.8
P3			1	0.5	0.1
P4				1	0.45
P5					1

Draw a dendrogram of the resulting clusters.

25. Describe a situation in which **spectral clustering** would produce a better clustering than a partition-based clustering method like **k-means**?

Part 7

26. What is the difference between **support** and **confidence**? Can you describe a situation in which a rule has high confidence but this doesn't represent a positive correlation?
27. Use the **Apriori algorithm** to generate frequent itemset and rules based on the following input:

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

minimum support count is 2
minimum confidence is 60%

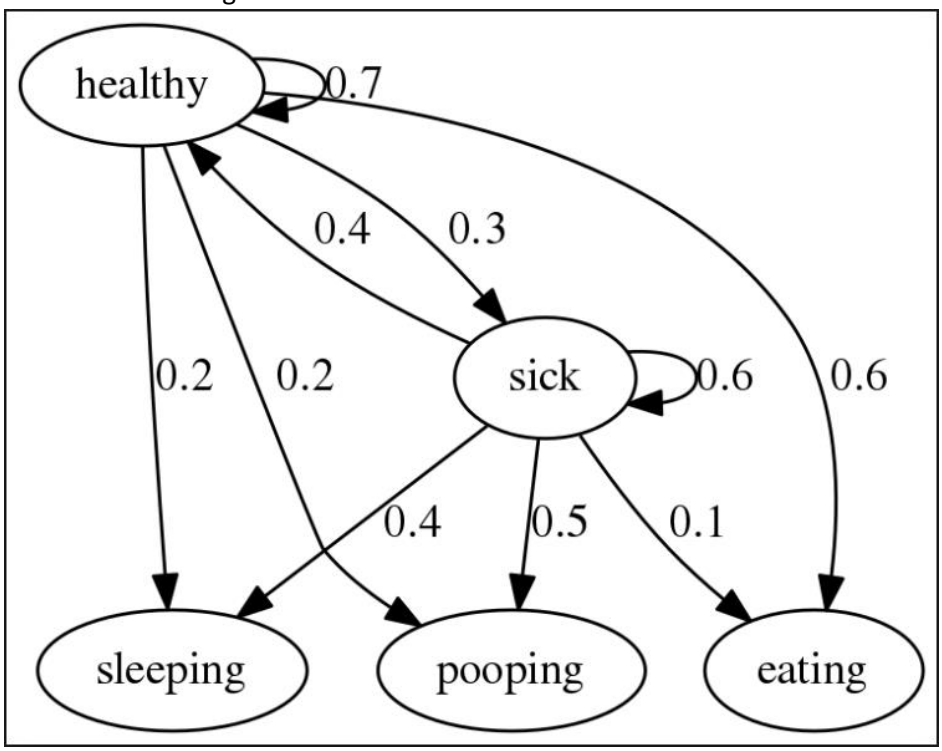
28. Find the frequent itemset for the following data using the FP-growth algorithm:

TID	Itemset
100	{M,O,N,K,E,Y}
200	{D,O,N,K,E,Y}
300	{M,A,K,E}
400	{M,U,C,K,Y}
500	{C,O,O,K,I,E}

Min_support = 3

Part 8

29. Given the following HMM:



What is the most likely sequence of states that would have the following observed emissions: sleeping, sleeping, eating, pooping, sleeping

Assume that initial probabilities for each state are equal

Hints

1. Page 11+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/1intro.pdf
2. Page 16+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/1intro.pdf
3. Page 10+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/2dw.pdf
4. Page 5+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/2dw.pdf
5. Page 27+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/2dw.pdf
6. Page 25+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/2dw.pdf
7. Page 31+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/2dw.pdf
8. Page 69+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/2dw.pdf
9. http://www.ams.org/publicoutreach/feature-column/fcarc-svd
10. http://faculty.bard.edu/~belk/math213s14/LinearCombinationsAndSpanRevised.pdf
11. Page 12+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/3prep.pdf
12. Page 19+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/3prep.pdf
13. Page 55+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/3prep.pdf
14. Page 56+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/3prep.pdf
15. Page 8+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/7class-a.pdf
16. Page 19+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/7class-a.pdf the first branch should be based around outlook
17. Page 26+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/7class-a.pdf
18. https://www.geeksforgeeks.org/naive-bayes-classifiers/
19. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
20. https://stackoverflow.com/questions/879432/what-is-the-difference-between-a-generative-and-a-discriminative-algorithm
21. Page 16+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/8clst.pdf
22. Page 13+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/8clst.pdf or https://machinelearningmastery.com/vector-norms-machine-learning/
23. Page 34+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/8clst.pdf
24. Page 42+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/8clst.pdf
25. http://www.cis.hut.fi/Opinnot/T-61.6020/2008/spectral_kmeans.pdf
26. Page 5+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/6asso.pdf and https://towardsdatascience.com/association-rules-2-aa9a77241654
27. https://www.geeksforgeeks.org/apriori-algorithm/
28. Page 22+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/6asso.pdf and https://www.youtube.com/watch?v=yCbankIouUU
29. Page 14+ from http://www.cse.unsw.edu.au/~cs9318/19t1/lect/Lx_HMM.pdf