

COMP37332 tutorial 3 (Data Mining)

1. Summarise the differences between OLAP and data mining.
2. Give examples of data mining applications in main application areas (business, science, government). Discuss the differences between predictive and descriptive data mining.
3. Explain the support and confidence measures and their role in mining association rules.
4. Describe the main steps in the Apriori algorithm. Explain how the anti-monotonicity property of itemsets can be used to reduce the search space in generating frequent itemsets.
5. Extract all rules with confidence above 75% and support above 25% from the following data:

<u>Customer</u>	<u>Items</u>
1	Orange Juice, Soda
2	Milk, Orange Juice, Window Cleaner
3	Orange Juice, Detergent
4	Orange Juice, Detergent, Soda
5	Window Cleaner, Soda

6. Explain how the decision trees are used for classification.
7. Describe the precision and recall evaluation measures. Give examples of applications that would prefer higher precision over or higher recall, and vice-versa.
8. Explain the main differences between classification and clustering.
9. An example dataset consists of five products whose amounts of sale in two regions are shown bellow. Cluster these products into two groups using the k-means algorithm, the Euclidean distance, and products A and E as initial cluster members.

data point	product	region 1	region 2
1	A	22	21
2	B	19	20
3	C	18	22
4	D	1	3
5	E	4	2

10. Cluster the data from the previous example using the k-means algorithm, the Manhattan distance and product A and E as initial cluster members.
11. Briefly describe the idea of agglomerative clustering. What is the difference between single and complete linkage methods for measuring inter-cluster distances?
12. Cluster the data from question 10 using the agglomerative clustering with single linkage method. The distance between points (i.e. products) should be calculated using the Euclidean distance. Compare the results.