**Past exam questions. Question 4, clustering:**

**2013 January paper:**

1. Calculate the **Euclidean** distance between each of the three rows of data below. Note: attributes are already scaled to the range [0, 10]. Which two rows are the most similar?

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Age** | **Level of education** | **Income** |
| Row 1: | 5 | 2 | 3 |
| Row 2: | 2 | 2 | 7 |
| Row 3: | 6 | 8 | 7 |

**(7 marks)**

1. Explain the difference between a **partition based** clustering algorithm and a **density based** clustering algorithm.

**(6 marks)**

1. One of the disadvantages of **k-means** clustering is that the number of clusters must be specified in advance. Assuming a dataset has 3 clusters, explain in detail how k-means clustering identifies the three cluster.

**(10 marks)**

1. Explain how **DBScan** identifies **core points**, **border points** and **noise points** in a dataset. How are these labels used to define clusters in the dataset?

**(7 marks)**

***Total: 30 marks***

**2013 Repeat paper:**

Below is a screen shot showing 12 rows of data from a dataset recording characteristics of chocolate bars. There are 8 attributes and no class label. Answer the questions below based on this dataset.



1. What is **unsupervised lea**rning, and why is it appropriate for this dataset?

**5 marks**

1. Explain in detail one preprocessing technique that should be applied to the dataset above prior to using a clustering algorithm. Justify your choice.

**6 marks**

1. Calculate the **Manhattan** distance between the first two rows in the chocolate dataset above. How would you include categorical attributes in a distance calculation?

**7 marks**

1. Recommend one algorithm you could use to identify groups of chocolate bars that have similar characteristics. Explain in detail how your chosen algorithm identifies clusters in the dataset.

**12 marks**

***Total: 30 marks***

**2012, Jan paper:**

1. Using the information provided in Figure 1 and Figure 2 below, calculate the *z-score* normalisation for the first two values of the attribute sepalwidth*.*

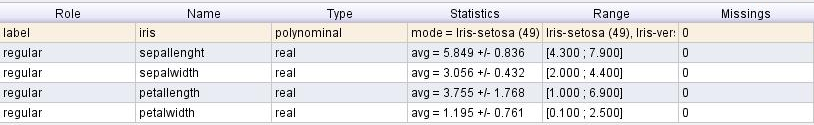
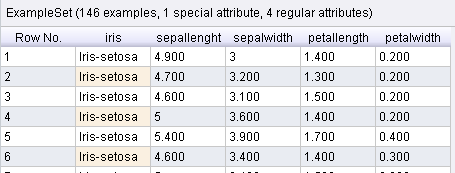


Figure 1 Meta data for the Iris dataset



**Figure 2 Excerpt of Iris dataset**

*(5 marks)*

1. What is *Clustering analysis*? In your answer contrast the following approaches to clustering: hierarchical vs. partitional vs. density; exclusive vs. overlapping; and complete vs. partial.

*(7 marks)*

1. Explain the *K-means* clustering algorithm. Include in your answer the advantages and disadvantages of k-means clustering.
2. Describe one method that can be used for cluster evaluation.

*(8 marks)*

**2012, repeat paper:**

1. Calculate the Euclidean distance for the data shown in the table below. What needs to be done if the scales of the attributes are different?

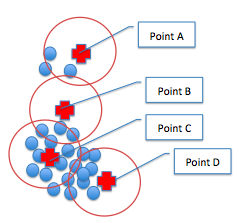
|  |  |  |  |
| --- | --- | --- | --- |
| row1 | 5 | 2 | 8 |
| row2 | 7 | 5 | 3 |

(6 marks)

1. Differentiate between *hierarchical*, *partitional* and *density* based clustering.

*(6 marks)*

1. Describe the DBScan clustering algorithm. For what type of dataset would DBScan be a better choice than K-Means? Using the diagram below, how would a DBScan algorithm label the four points (core, border or noise) assuming MinPts = 5, and Eps is illustrated by the four circles.



*(9 marks)*

**2009, summer paper:**

1. Explain why the notion of a **cluster** can be ambiguous.

**5 marks**

1. Calculate the Euclidean distance between the following three data points:



Note:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Attribute 1** | **Attribute 2** | **Attribute 4** |
| Point a | 10 | 2 | 100 |
| Point b | 15 | 3 | 70 |
| Point c | 12 | 7 | 90 |

**12 marks**

1. Explain the benefit of **normalising** attributes before calculating distances. Make reference to the data in part (b) above to illustrate you answer.

**4 marks**

1. Explain in detail how the **k-means clustering** algorithm works.

**9 marks**

**2009, repeat paper:**

1. Explain the difference between **partitioning** and **hierarchical** clustering.

**5 marks**



1. Calculate the relative distance between the following three data points:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Attribute 1** | **Attribute 2** | **Attribute 4** |
| Point a | 8 | 2 | 50 |
| Point b | 10 | 4 | 200 |
| Point c | 5 | 1 | 250 |

**12 marks**

1. With reference to the data above, explain how a variables range can be used to weight the variable in terms of their influence on distance calculations.

**4 marks**