UNIVERSITY OF BIRMINGHAM

School of Computer Science

Machine Learning and Intelligent Data Analysis

Resit Examinations 2021

Machine Learning and Intelligent Data Analysis

Question 1 Clustering

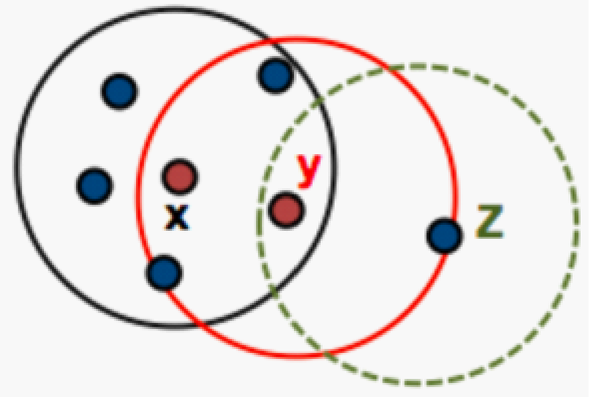
(a) Explain the purpose of the k-means algorithm and how it works. **[4 marks]**

(b) Give two examples of distance (also known as similarity) metrics commonly used in clustering algorithms and explain how they affect the result obtained. **[2 marks]**

(c) Explain when you would use k-means clustering and when you would use hierarchical clustering. **[3 marks]**

(d) A dataset $\mathbf{X} = \{0, 2, 4, 6, 24, 26\}$ consists of six one-dimensional data points. The k-means clustering algorithm is initialized with 2 cluster centres at $c_1 = 3$ and $c_2 = 4$. What are the values of c_1 and c_2 after one iteration of k-means? What are the values of c_1 and c_2 after the second iteration of k-means? **You must show your working for full marks.** **[4 marks]**

(e) In density based clustering, each data point is categorised as being a 'core' point, a 'border' point or a 'noise' point. The figure below shows multiple data points, three of which are labelled as x, y, and z. The circles represent the Eps-Neighbourhoods of the three labelled points and the parameter MinPts $= 6$. Identify whether each of the points (x, y, z) is a 'core' point, a 'border' point or a 'noise' point. **Explain your reasoning.** **[7 marks]**



Question 2 Classification

(a) Consider the following optimisation problem corresponding to Soft Margin Support Vector Machines:

$$\text{argmin}_{\mathbf{w},b,\xi} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^{N} \xi^{(n)} \right\}$$

subject to

$$y^{(n)} f(\mathbf{x}^{(n)}) \geq 1 - \xi^{(n)}, \ \forall n \in \{1, 2, \cdots, N\},$$

where $\mathbf{w}$ are the hyperplane parameters, b is the bias, ξ are the slack variables, $(\mathbf{x}^{(n)}, y^{(n)})$ is the training example n, and N is the number of training examples.

Should the constant C be positive or negative? **Explain why.** **[10 marks]**

(b) Consider the k-Nearest Neighbour algorithm learnt in Lecture 3b, applied to classification problems. In this algorithm, all k nearest neighbours contribute equally to the prediction of a given example. One may wish that examples closer to the example being predicted contribute more towards such prediction. Propose an alteration to the k-Nearest Neighbour algorithm that satisfies this requirement. **Explain how this alteration works.** **[10 marks]**

Question 3 Document Analysis

(a) You are given the following three documents.

- d_1: The cat sat on the dog's mat
- d_2: The dog chased the cat
- d_3: The dog ate its dinner

Stop words (the, on, its) are removed and the documents are stemmed.

Construct the document index for these documents following stop-word removal and stemming. **Explain why this data structure is useful.** **[12 marks]**

(b) Compare and contrast the LSA and word2vec methods for semantic embedding. **[8 marks]**