Machine Learning and Intelligent Data Analysis
Solutions

Resit Examinations 2021

Machine Learning and Intelligent Data Analysis

Learning Outcomes

(a) Demonstrate knowledge and understanding of core ideas and foundations of unsupervised and supervised learning on vectorial data

(b) Explain principles and techniques for mining textual data

(c) Demonstrate understanding of the principles of efficient web-mining algorithms

(d) Demonstrate understanding of broader issues of learning and generalisation in machine learning and data analysis systems

Question 1 Clustering

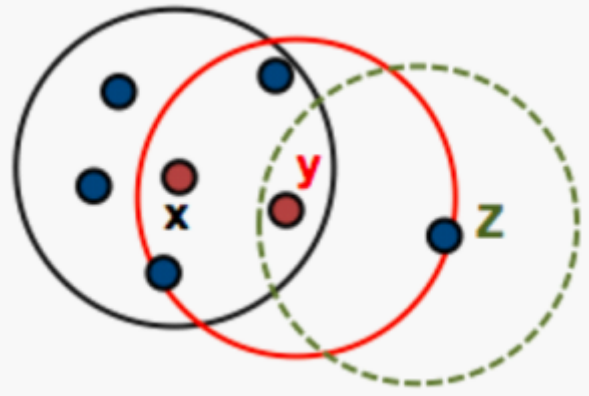
(a) Explain the purpose of the k-means algorithm and how it works. **[4 marks]**

(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]**



Model answer / LOs / Creativity:
Learning outcomes 1 and 4. Part d is creative.

(a) Find groups of points in a dataset. Choose the number of clusters to find. Randomly allocate points to the clusters. Compute the centroid (mean) of the clusters. Re-allocate points to the cluster with the closest centroid. Repeat until clusters are stable. **[4]**

(b) Euclidean distance groups points that are close in terms of straight-line distance. Cosine distance groups points together that are in the same "direction" from the origin. **2**

(c) In hierarchical clustering, no assumption about the number of clusters is made whereas in k-means clustering, the number of clusters to be made are specified before-hand. What is useful is that if unaware about the number of clusters to be formed, use hierarchical clustering to determine the number and then use k-means clustering to make more stable clusters as hierarchical clustering is a single-pass exercise whereas k-means is an iterative process.**3**

(d) Iteration 1: $c_1 = 1$, $c_2 = 15$; Iteration 2: $c_1 = 3$, $c_2 = 25$.**[4]**

(e) A point is a core point if it has more than a specified number of points (MinPts) within Eps These are points that are at the interior of a cluster. A border point has fewer than MinPts within Eps, but is in the neighbourhood of a core point.A noise point is any point that is not a core point nor a border point. Therefore: x = core. Y = border, z = noise. **[7]**

Question 2 Classification

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

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

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]**

Model answer / LOs / Creativity:
Learning outcomes 1 and 4. Part b is creative

(a) It should be positive [1 mark], as this is a minimisation problem and we want to reduce the amount of slack $\xi^{(n)}$ [3 marks] so that we don't have too many examples too far away from the correct side of the margin [3 marks]. Had it been negative, we would be encouraging to increase the amount of slack instead of reducing it [3 marks].

(b) One can use the weighted majority vote instead of a simple majority vote [5 marks]. The weight can be set to the inverse of the Euclidean distance [5 marks].

Learning outcomes:

- Demonstrate knowledge and understanding of core ideas and foundations of unsupervised and supervised learning on vectorial data.
- Demonstrate understanding of broader issues of learning and generalisation in machine learning and data analysis systems.

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]**

Model answer / LOs / Creativity:

Learning outcomes 2, 3. Parts a, b are creative.

(a) The document index is a look-up table that is indexed by the document vocabulary. For each term in the vocabulary, the index contains the id of the document, and the number of occurences of that term in that document. This is useful because it allows the inverse document frequency and the term frequencies to be computed easily by counting the elements in the table.**[5]**

Using the example in part a):

cat	sat	dog	mat	chase	ate	dinner
(1,1) (2,1)	(1,1)	(1,1) (2,1) (3,1)	(1,1)	(2,1)	(3,1)	(3,1)

[7]

(b) LSA: based on term frequency only. Topics (embedding dimensions) derived from correlations across a corpus. Models documents as linear combinations of topics. Can work well with small corpora but does not scale well to large datasets. Does not include any sub-document information. **[2 marks for correct description, 2 marks for analysis]**

word2vec: predictive model based on either BoW (predict missing word) or skip-gram (predict word context) approach. Takes both global and local context into account. Generally requires large corpus. **[2 marks for correct description, 2 marks for analysis]**