MSc Data Science/MSc Computer Science

Data Mining and Machine Learning

Final Examination, II Semester, 2020–2021

Date : 6 July, 2021 Marks : 40
Duration : 3 hours + 0.5 hours upload time Weightage : 40%

1. We have a dataset $X = \{x_1, x_2, \ldots, x_N\}$ equipped with a symmetric distance function: $d(x_i, x_j) = d(x_j, x_i)$ is the distance between x_i and x_j. We construct an $N \times N$ matrix D such that $D[i, j] = d(x_i, x_j)$. We can cluster the N columns of D using the usual Euclidean distance, since each column is a vector of length N. Explain whether the clusters formed by the columns of D have any meaningful interpretation with respect to the original set X. *(5 marks)*

2. Explain how to detect outliers using (a) density based clustering (DBScan) and (b) clustering using a mixture of Gaussians. *(5 marks)*

3. There are three biased coins c_1, c_2, and c_3. You are given a sequence of 1000 coin tosses, where each outcome corresponds to tossing one of $\{c_1, c_2, c_3\}$, chosen uniformly at random. Let $\{p_1, p_2, p_3\}$ be the probabilities of heads for the coins $\{c_1, c_2, c_3\}$, respectively. You have prior information that p_1 is less than 0.5 and p_2 and p_3 are greater than 0.5. Describe, in algorithmic pseudocode, an iterative procedure to estimate $\{p_1, p_2, p_3\}$. *(5 marks)*

4. Consider the iterative algorithm to compute singular vectors discussed in the class. Explain why the singular values computed by the algorithm satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. *(5 marks)*

5. Suppose we have a neural network with four input features x_1, x_2, x_3, x_4 and a single output y. As usual, we assume that each pair of adjacent layers is completely connected and there is a single output layer. How many parameters do we have to estimate in the following situations?

   (a) A shallow network with 1 hidden layer consisting of 18 nodes.

   (b) A deep network with 3 hidden layers, where the first two layers have 3 nodes each and the third layer has 2 nodes. *(5 marks)*

6. We made the following assumptions about the loss (cost) function C for neural networks.

   - For each input x, $C(x)$ is a function of only the output layer activation.
   - The total cost across the training set is the average of the individual input costs

   Explain why these assumptions are important for effective learning of the parameters.

   (5 marks)

7. Let $f(x_1, x_2, \ldots, x_k)$ be a boolean formula with k inputs. The set of inputs for which f is true defines a *concept class* $C_f \subseteq \{0, 1\}^k$ of k-dimensional bit vectors. Let $C \subseteq \{0, 1\}^k$ denote an arbitrary concept class. When can such a concept class C be represented by a neural network? Explain your answer. *(5 marks)*

8. Consider a neural network that is layered and completely connected. Suppose we initialize two nodes n_1 and n_2 from the same layer with the same biases and same weights on incoming and outgoing edges. What can you say about the final weights and biases that will be learned for n_1 and n_2 through backpropagation? What can you conclude about initialization strategies for such networks? *(5 marks)*