UNIVERSITY OF
BIRMINGHAM

School of Computer Science

Machine Learning and Intelligent Data Analysis

Main Summer Examinations 2021

Machine Learning and Intelligent Data Analysis

Question 1 Dimensionality Reduction

(a) Explain what is meant by "dimensionality reduction" and why it is sometimes necessary. **[4 marks]**

(b) Consider the following dataset of four sample points $\{\mathbf{x}^{(i)}\}_{i=1}^{4}$ with $\mathbf{x}^{(i)} \in \mathbb{R}^2 \; \forall i$:

$$\mathbf{X} = \begin{pmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{pmatrix}$$

Explain how to calculate the principal components of this dataset, outlining each step and performing all calculations up to (but not including) the computation of eigenvectors and eigenvalues. **[6 marks]**

(c) What does principal component analysis (PCA) tell you about the nature of a multivariate dataset? Explain how it can be used for dimensionality reduction? **[4 marks]**

(d) What are the limitations of PCA and what other dimensionality reduction techniques may be used instead? **[2 marks]**

(e) You are given a dataset consisting of 100 measurements, each of which has 10 variables. The eigenvalues of the covariance matrix are shown in the following table:

Eigenvalue number	1	2	3	4	5	6	7	8	9	10
Eigenvalue	1382.0	508.4	187.0	68.8	25.3	9.3	3.4	1.3	0.46	0.17

What can you say about the underlying nature of this dataset? **[4 marks]**

Question 2 Classification

(a) Consider the Soft Margin Support Vector Machine learnt in Lecture 4e. Consider also that $C = 100$ and that we are adopting a linear kernel, i.e., $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \mathbf{x}^{(i)^T}\mathbf{x}^{(j)}$. Assume an illustrative binary classification problem with the following training examples:

$\mathbf{x}^{(1)} = (0.3, 0.3)^T$, $y^{(1)} = 1$
$\mathbf{x}^{(2)} = (0.6, 0.6)^T$, $y^{(2)} = 1$
$\mathbf{x}^{(3)} = (0.6, 0.3)^T$, $y^{(3)} = -1$
$\mathbf{x}^{(4)} = (0.9, 0.6)^T$, $y^{(4)} = -1$

Which of the Lagrange multipliers below is(are) a plausible solution(s) for this problem? **Justify your answer.**

(i) $a^{(1)} = 0$, $a^{(2)} = 2$, $a^{(3)} = 2$, $a^{(4)} = 10$

(ii) $a^{(1)} = 0$, $a^{(2)} = 44$, $a^{(3)} = 22$, $a^{(4)} = 22$

(iii) $a^{(1)} = 0$, $a^{(2)} = 200$, $a^{(3)} = 100$, $a^{(4)} = 100$

[6 marks]

(b) Consider a binary classification problem where around 5% of the training examples are likely to have their labels incorrectly assigned (i.e., assigned as -1 when the true label was +1, and vice-versa). Which value of k for k-Nearest Neighbours is likely to be better suited for this problem: $k = 1$ or $k = 3$? **Justify your answer.** **[6 marks]**

(c) Consider a binary classification problem where you wish to predict whether a piece of machinery is likely to contain a defect. For this problem, 0.5% of the training examples belong to the defective class, whereas 99.5% belong to the non-defective class. When adopting Naïve Bayes for this problem, the non-defective class may almost always be the predicted class, even when the true class is the defective class. Explain why **and** propose a method to alleviate this issue. **[8 marks]**

Question 3 Document Analysis

(a) In a small universe of five web pages, one page has a PageRank of 0.4. What does this tell us about this page? **[2 marks]**

(b) Compare and contrast the TF-IDF and word2vec approaches to document vectorisation. You should explain the essential principles of each method, and highlight their respective advantages and disadvantages. **[8 marks]**

(c) One possible approach to searching a large linked set of documents is to combine a measure of document similarity such as TF-IDF similarity with a measure of a page's importance such as that provided by PageRank. Suggest three ways in which this could be done and discuss the advantages and disadvantages of each of them. **[10 marks]**

Total Points 59 != Expected 60