Machine Learning and Intelligent Data Analysis

Solutions

Main Summer 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 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 $\left\{\mathbf{x}^{(i)}\right\}_{i=1}^{4}$ 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]**

Model answer / LOs / Creativity:
Learning outcomes a and d. Part e is creative.

(a) Finding a basis (coordinate system) in which the data can be represented in terms of a reduced number of coordinates without loss of significant information. It is necessary because of the curse of dimensionality which leads to problematic phenomena such as convergence of distances and ultra-sparse sampling, and sometimes also because it allows the data size to be reduced. **[4]**

(b) From data matrix $\mathbf{X}$ subtract column means to form

$$\mathbf{X}' = \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{pmatrix}.$$

Form the covariance matrix

$$\mathbf{C} = \mathbf{X}^\top \mathbf{X} = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$$

Find eigenvalues/vectors – NOT needed as we did not do this in class.
Ordered eigenvalues correspond to the variance of the data projected onto the corresponding eigenvector. Usually aim to remove directions with small variance as not informative. **[6]**

(c) Aligns coordinates with natural directions in the data in order of decreasing variance. Directions in which the data does not vary can be considered to be unimportant and thus removed, reducing the dimensionality. **[4]**

(d) PCA can be costly to perform, and is a strictly linear technique. RP is very cheap but it is also linear and interpretation is difficult. **[2]**

(e) 95% variance is in first three dimensions, with ¿ 99% in the first five dimensions. Likely that true dimensionality of data is in this region. Unlikely that data is genuinely 10-dimensional. **[4]**

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]**

Model answer / LOs / Creativity:
Learning outcomes a and d. Part c is creative.

(a) Items (i) and (iii) do not satisfy the constraints of the dual representation of the problem. In particular, (i) does not satisfy $\sum_{n=1}^{N} a^{(i)} y^{(i)} = 0$ [2 marks], whereas (iii) does not satisfy $0 \leq a^{(i)} \leq C$ [2 marks]. Only (ii) satisfies all the constraints, and is thus plausible [2 marks].

(b) The value $k = 3$ is likely to be better suited for this problem. This is because adopting $k = 1$ will cause the classifier to be sensitive to noise [3 marks], whereas $k = 3$ is likely to reduce a bit of this sensitivity [3 marks].

(c) This is going to happen because the class-conditional probabilities are multiplied by the prior probability of the class when computing the probability of a given example belonging to this class. This prior probability is too low for the minority class, bringing the whole probability of the example belonging to this class down and making it rather unlikely that the classifier will predict the minority class [4 marks].

Possible ways to alleviate this issue include forcing the prior probability of the classes to be 50%, no matter the actual prior probability of the class; or oversampling

examples of the minority class; or undersampling examples of the majority class [4 marks for any of these or other suitable proposal].

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]**

Model answer / LOs / Creativity:
Learning outcomes b, c. Part c is creative.

(a) The page is more important than average, as measured by the number of links that point to it. If all pages were equally important, the pagerank would be 0.2. This page is therefore much more heavily linked to than the other pages. **2**

(b) TF-IDF uses pure word frequency information locally (TF) and globall (IDF). Does not capture any semantic information, in particular about word order. Can be applied to relatively small data. Documents and terms can be added on-the-fly. Interpretable representation. **[4]**

word2vec learns the semantic context of words as it is trained to predict either a missing word (bas of words) or a word's context (skip-gram). Generally needs large corpus. Cannot be retrained on-the-fly. Can learn non-linear relationships within the training set. Representation not interpretable.**[4]**

(c) The idea that these two concepts can be combined for high quality information retrival was discussed in lectures, but now how it could actually be done. A number of possible creative solutions are possible here.

Potential approaches that might be discussed are:

- Compute the document similarity vs all document, multiply by the Page rank and sort. This is simple to implements but computationally expensive for very large document sets. It also implicitly means that it's very hard to compensate for a low Page rank which could lead to less relevant documents being returned.

- As above, but adding the two terms. This is more appealing because it allows for a strong similarity to overcome a low Page rank to some extent.
- Sort the documents by Page rank and then compute the similarity of documents against some top fraction of the sorted documents. This is very efficient because the Page rank is precomputed, but it will give very poor results because there is no guarantee that relevant documents will have a high Page rank.
- Compute the similarity of the query against all documents, select the top matches, and then order these by Page Rank. This is likely to give the best matches, but is computationally very expensive.

[2 marks for each of 3 sensible approaches. 4 marks for correctly analysing the likely performance.]

Total Points 59 != Expected 60