CASS
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON
EST 1894

Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Li, Liang 2. Liu, Houhung 3. Yang, Zhengzhao 4. Zhang, Yunqing	**GROUP NUMBER:** 02
MSc in: Business Analytics	
Module Code: SMM636	
Module Title: Machine Learning	
Lecturer:	**Submission Date:**

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

Final Mark: %

SMM636
Machine Learning

Coursework 1

Group 2

Group 2: Li Liang, Liu Houhung, Yang Zhengzhao, Zhang Yunqing

Question1

Please describe one classification task and one clustering task in business applications. Do not use the examples discussed in lecture notes or R exercises. Justify your answers clearly: why it is a classification or clustering task.

Introduction

Machine learning is a subfield of artificial intelligence, which has become increasingly ubiquitous in various industries. From image recognition to product recommendations and fraud detection, machine learning has become a vital driving force for today's technological advances. Machine learning, in its simplest form, is the process of finding and applying patterns in data. It identifies patterns through two approaches: supervised learning and unsupervised learning. This report will examine two real-world business applications to explain the approaches above.

Supervised Learning and Classification

Supervised learning is the most common subbranch of machine learning. In supervised learning, computers are provided with example inputs that are labelled with their desired outputs. After patterns are identified in the training dataset, those patterns are then used to predict additional data labels. Moreover, classification is a type of supervised learning. It specifies the class to which data elements belong to. There are two types of classification: binomial and multi-class.

Classification Business Application

In many industries, classification is applied to improve business performances directly. People.ai (2021), for example, is a startup which aims to use sales data, including emails, calls, and customer relationship interactions to predict preferable actions that would bring better sales results through classification algorithms.

The company's primary purpose is to provide a data-supported handbook for sales representatives (People.ai, 2021). The company collects sales activity data for years and labels those data according to their success results (deal or not deal). The algorithms then analyze those data by focusing on the successful cases and recommend sales reps to perform the desired actions that are shown to have a positive linkage to the successful cases. People.ai also helps drive sales by predicting leads who are most willing or most likely to make a deal by analyzing their interactions data. For example, if a response is exceptionally late or wording in an email indicates an unwillingness to make a deal, the algorithm will class the people associated with these data into the same group. By doing so, the classification algorithm directs sales to focus on people who are more likely to purchase from them and in turn, avoids wasting resources on futile interactions.

The method that People.ai adopts in this process is thus supervised learning because they need to train the algorithms or models by well-labelled datasets first to generate recommendations. It uses the classification method because the algorithm specifies the class to which clients belong according to their likelihood of making a deal.

Unsupervised Learning and Clustering

On the other hand, unlike supervised learning, unsupervised learning uses algorithms to analyze and cluster unlabeled datasets, making unsupervised learning more challenging than supervised learning. It can group unlabeled data based on their similarities or differences into different groups. These groups are known as "clusters" (Mary, n.d.).

Clustering Business Application

Unsupervised learning is one of the most significant technological development in business. The purpose of unsupervised learning is to have the algorithms detect patterns within the training data sets and classify the input objects based on the features that the system itself recognizes. Companies can utilize and bring huge benefits from it, such as customer segmentation. Rather than depending on a marketer's instinct to divide customers into groups for marketing campaigns, customers can be described and categorized into segments characterized by different demographics and physiographic through unsupervised learning (Castle, 2017).

Take Amazon for an example; unsupervised learning helps Amazon personalize customer services and update the recommender engines. In details, following the three activities (Bushkovskyi, n.d.):

A. exploring the structure of the information;
B. finding common elements in the data;
C. predicting trends coming out of data,

By applying three activities, algorithms find associations among data points. It provides a capability that Amazon can use to identify what products are often bought together to customers.

To sum up, the unsupervised learning analyzes the data sets' underlying structure by extracting useful information or patterns from them. Business decision-makers could identify their consumers' behaviour and preferences and manage their targeted advertising content.

Question 2

Identifying spam emails is one important classification task and comment on the results you obtained.

We measured the classification accuracy of our kNN models with the following formula:

$$\mathrm{Acc}_{Te} = \frac{1}{N_{Te}} \sum_{i \in Te} I(y_i = \hat{f}(\mathbf{x}_i))$$

As shown in table 1, there is a significant difference between the classification accuracy of a small k (k = 1, accuracy = 58%) and a large k (k = 25, accuracy = 70%), which indicate that increasing the number of k is likely to boost up the overall classification accuracy of a kNN model.

Besides, since the number of correctly classified spam emails increases when the number of correctly classified non-spam emails has a trough when k = 9, it is possible that with the current model configurations (100 samples in train set, 50 samples in test set), continuously increasing the value of k will be more likely to result in better classification accuracy for spam emails with fluctuated results for non-spam emails. Moreover, it is also possible that k = 9 is not an appropriate number of nearest neighbours which has a higher chance of causing misclassification.

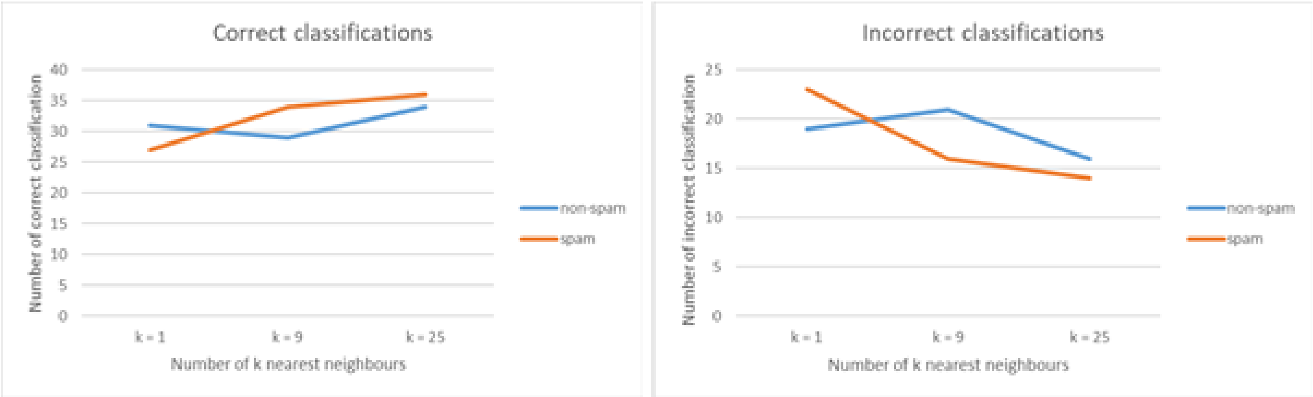
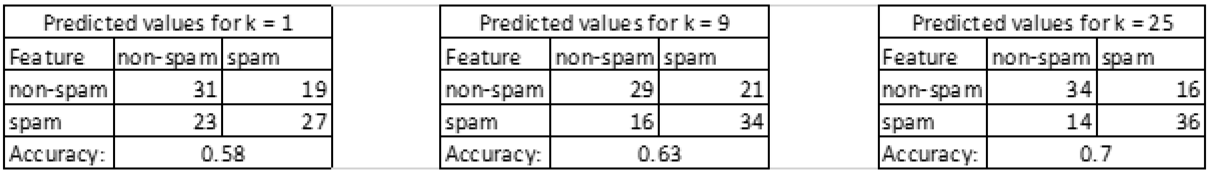


Figure 1. No. of correctly & incorrectly classified cases of the three kNN models



Predicted values for k = 1		
Feature	non-spam	spam
non-spam	31	19
spam	23	27
Accuracy:	0.58	

Predicted values for k = 9		
Feature	non-spam	spam
non-spam	29	21
spam	16	34
Accuracy:	0.63	

Predicted values for k = 25		
Feature	non-spam	spam
non-spam	34	16
spam	14	36
Accuracy:	0.7	

Table 1. Outputs and accuracy of kNN models with k = 1, k = 9 and k = 25.

Reference

Bushkovskyi, O., n.d. *Guide to Unsupervised Machine Learning (With Examples)*. [online] Theappsolutions.com. Available at: <https://theappsolutions.com/blog/development/unsupervised-machine-learning/> [Accessed 2 February 2021].

Castle, N., 2017. *6 Common Machine Learning Applications for Business*. [online] Oracle AI and Data Science Blog. Available at: <https://blogs.oracle.com/datascience/6-common-machine-learning-applications-for-business> [Accessed 1 February 2021].

Mary, K., n.d. *What is Unsupervised Learning?*. [online] SearchEnterpriseAI. Available at: <https://searchenterpriseai.techtarget.com/definition/unsupervised-learning> [Accessed 4 February 2021].

People.ai. 2021. *People.ai | Revenue Operations & Intelligence Powered by AI*. [online] Available at: <https://people.ai> [Accessed 3 February 2021].