National College of Ireland

MSc in Data Analytics – Part-time – Year 1 – MSCDAD_JAN18A/MSCDAD_JAN18B

Semester Three Examinations – 2017/18

Monday 20th August 2018
10.00am – 1.00pm

Advanced Data Mining H9ADM

Dr. Geraldine Gray
Mr. Noel Cosgrave
Mr. Vikas Tomer

Answer question 1 and any 3 of the remaining 5 questions. All questions carry equal marks.

Duration of exam: 3 hours
Attachments: None

1. **Fundamentals of Data Mining (Compulsory Question) [25 marks]**

   a) Describe how your approach to sampling would change in the presence of very large (millions of rows) when compared with sampling very small data (a few 100 rows). [10 Marks]

   b) False Positives and False Negatives: [8 Marks]
      i. What is the difference between a false positive and a false negative; [2 Marks]
      ii. How would they map to type I or II errors; and [2 Marks]
      iii. Does it make a difference if more false positives or false negatives are observed (give an example to justify your answer)? [5 Marks]

   c) What measure(s) of performance would you use to evaluate a model trained on a largely imbalanced dataset? Justify your choices based on how they can inform your view on the quality of a model you have produced. [7 Marks]

2. **Applied Methods I** [25 marks]

You have been hired by a bank to develop models to determine if any suspicious or fraudulent activity has taken place. You have been provided with a sample of selected training data, but no information is available as to how this sample has been created or curated. You should assume that the data has not been cleaned and that there are missing values. You have been provided with:

- account holder demographic details
- a historical transactions data across different accounts,
- details of all accounts a customer has, their balances, monthly in- and out-goings, how long the account has been open,
- and for 50% of your customers a human labelled training dataset with 2 values:
  1) fraudulent activity is suspected
  2) fraudulent activity not suspected

You should evaluate each model against the following criteria:
1) accusations of fraudulent activity are to be binned into sub categories of possible, likely, very likely [1 Mark per method]
2) models should not have a high computational demand [1 Mark per method]
3) models may need to be used as evidence, must be interpretable [1 Mark per method]
4) note if you would consider using this method and why [2 Marks per method]

Evaluate the following methods of tackling this problem against each of the 4 criteria above:

a) An artificial neural network

b) Logistic regression

c) K-means

d) A random forest

e) Association Rule Mining

3. **Applied Methods II** [25 marks]

You are working for a company that provides third-party support for large multinationals. Support requests are received by email or through a web interface. You have been tasked with ways to identify reports as “low priority” and “high priority” based solely on the content of the report. You have been provided with a labelled dataset of previous support requests and their priority. How would you approach this task using any **TWO** of the following methods [8.5 Marks for each method]?

a) Association Rule Mining
b) Naïve Bayes
c) kNN
d) Lasso Regression

Note how you would prepare the data and approach the problem [up to 8 marks]. State any necessary assumptions of the data for the method(s) chosen. Be aware that some methods may be better than others, and some may not work at all. You should, however, only suggest ones that will work.

You may assume the following of the data and environment:

1. You have a large, but imbalanced training set that is biased towards “not important”
2. Support requests are sometimes quite detailed but are all in the same language
3. Compute power is not an issue

4. **Time Series Analysis** [25 marks]

a) Briefly outline ALL of the steps, including model diagnostics, that you would follow when using ARIMA to model a set of time series observations. [10 Marks]

b) Provide the definition for a simple moving average and a weighted moving average. [5 Marks]

c) Consider the following series: 18, 31, 25,14, 26, 16, 29, 21, 18
Apply a 2x2MA smoother to these data and produce a new smoothed series. [5 Marks]

d) Define the terms trend-stationary and difference-stationary and provide an example of each.

5. **Support Vector Machines** [25 marks]

You have been asked to tackle three independent problems using Support Vector Machines:

1- a regression-style problem;
2- a multi-class classification problem: recognition of images of handwritten numbers: 0-9; and
3- a binary classification problem on a wide, but sparse data set

a) Suggest an appropriate kernel to use with each problem and state why it is suitable [12 Marks]

b) Support Vector Machines can be impacted by the curse of dimensionality. In which of the 3 problems above is this most likely to be an issue. Discuss. [5 Marks]

c) Describe, highlighting any assumptions you make and discussing their implications, how you would handle the curse of dimensionality for one of the problems you noted in part b as suffering from the curse of dimensionality. [8 Marks]

6. **Text Mining** [25 marks]

You have a large repository of research documents that cover an extensive variety of topics and topic categories (for example Biochemistry, Applied Physics, Computer Science, Literature, Art, Music). These are written in a number of different languages (for example Spanish, Italian, French, German and English). Describe a data mining process that would permit you to catalogue representative topics present in each category differentiated by language. In addition, critically evaluate your approach paying specific attention to any limitations or specific challenges, and assumptions you have made in order to formulate your approach or simplify the problem as well as the potential impact of said assumptions. Finally, describe an evaluation methodology for your approach.

Mark breakdown:

- 10 marks for the pre-processing, i.e. assignment of categories and language identification
- 10 marks for the topic cataloguing process
- 5 marks for the proposed evaluation methodology