National College of Ireland

NCI DOWNLOAD 26.04.18 16:51:44

National College of Ireland

Masters of Science in Data Analytics – Full-time – Year 1 – MSCDAD_A / MSCDAD_B
Masters of Science in Data Analytics – Part-time – Year 1 – MSCDA1/ MSCDAJANO
Post Graduate Diploma in Science in Data Analytics – Part-time – Year 1 – PGDDSB1/ PGDSBJANO

Semester Two Examinations – 2017/18

Monday 14th May 2018
2.00pm – 5.00pm

Advanced Data Mining

Dr Geraldine Gray
Dr Simon Caton
Dr Cristian Rusu
Vikas Tomer

Answer Question 1, and 3 of the remaining 5 questions.

Duration of exam: 3 hours
Attachments: none

1. Foundations [25 Marks]

   a) Given a sample of mixed data, discuss how would you prepare the data for a model that can only use either numeric or categorical data, but not both. [6 Marks]

   b) Discuss how your approach to sampling would change in the presence of very large (millions of rows) vs. very small data (a few 100 rows). [10 Marks]

   c) How would you handle selection bias in a supervised data mining scenario [6 Marks]? Note: define selection bias in your answer [2 Marks] giving an example to help structure your answer [1 Mark]. [9 Marks in total]

NCI DOWNLOAD 29.04.19 16:51:44

Do 3 of the remaining 5 Questions

2. Applied Methods I [25 Marks]

   Data is abundant in the fashion and retail industry, and companies are increasingly interested in identifying what will be "hot" next season. You have been hired to consult a clothing company on their machine learning strategy, to predict "hot" or "not" for a large set of clothing items. They have a large number of previous sales observations (line items), and the numeric data is extremely wide (1000s of features). There is some missing data, and some extreme outliers (i.e. very "hot" products).

   You are to propose the data mining strategy. The company has the following requirements:
   R1: computation for prediction should be minimal
   R2: false positives are a problem, false negatives are not
   R3: explaining how the model works is not important,
   R4: the resultant model must be both conservative and confident,
   R5: training times are not an issue, but candidate models should be ranked on their prospective computational demands

   a) How will you prepare this dataset for analysis? [6 Marks]

   b) Discuss which measure(s) of performance will you use to evaluate models, and why. [4 Marks]

   c) Discuss in reference to the requirements above, which TWO of the following methods you would propose as the BEST candidates for this problem [10 Marks]
      i) Naïve Bayes
      ii) Logistic Regression
      iii) Support Vector Machine
      iv) kNN
      v) An artificial neural network

   d) The company MD proposes that association rule mining should replace one of your two methods, as it is widely used in analysing customer purchases. How will you address this suggestion? [5 Marks]

3. Applied Methods II [25 Marks]

You work in the complaints department for a services company, and receive a steady supply of strongly worded complaint emails from customers, which over the years you have classified by hand into:

Class	*Response*
Not a real complaint	Send a 10€ gift voucher
Not Sure	Request more information from customer
Serious complaint	Forward to complaints handler

NCI DOWNLOAD 29.04.19 16:51:44

Over the years, you have noticed that there seems to be some sort of pattern in the text, time of day the complaint is received, products complained about etc. You have decided to try to build a machine learning model to filter the complaints you receive.

To ease this process, you have changed how complaints are represented by customers: an electronic form detailing their information (customer ID, full name, email), subject of complaint (i.e. the service), a series of yes/no questions, up to 500 words free text corresponding to the body of the complaint. You also add in some details on the "importance" of the customer to the company from the CRM system.

Note that your company has not authorised this activity, so large numbers of misclassification may raise their suspicions.

Describe in detail your approach to this classification problem. Discuss the implications of any assumptions you make.

a) Are you most concerned about Type I, II or III errors in this problem? Discuss [6 Marks]

b) How will you prepare the data for analysis? [5 Marks]

c) You have heard that some methods that may work well in this area are:
   i) Linear Regression
   ii) Logistic Regression
   iii) K-Means
   iv) Latent Dirichlet Allocation
   v) Sentiment Analysis
   vi) Long Short Term Memory (LSTM)

Which will you choose and how will you build your machine learning model? Note some options are not viable on their own, some are inappropriate, and some can be combined. [10 Marks]

d) How (briefly) will you evaluate your model(s) developed in part c, i.e. which performance metric(s)? [4 Marks]

4. Clustering [25 Marks]

   a) There are many ways to evaluate a cluster. Discuss the following with respect to how to interpret them with respect to evaluating the output of a clustering algorithm. Include in your discussion when they complement or mirror (i.e. are redundant) each other in assessing cluster quality. [9 Marks]
      i) The Davies-Bouldin (DB) index
      ii) Silhouette
      iii) Total/Between cluster sum of squares

   b) Discuss the impact on k-means and mixture models (including potentially any effects on preprocessing) the presence of the following. Note that you may **NOT** explicitly remove columns, only transform them.
      i) Large amounts of numeric data [2 Marks]
      ii) Only categorical data [2 Marks]
      iii) Both i and ii concurrently [4 Marks]
      iv) Non-Euclidean data (give examples!) [4 Marks]

   c) Assuming that we have identified a "good" number of clusters. Discuss (noting key assumptions) how we could identify outliers distinguishing between exclusive and overlapping clusters. [4 Marks]

NCI DOWNLOAD 29.04.19 16:51:44

5. Ensembles [25 Marks]

   a) Compare and contrast the key differences between Bagging, Boosting and Stacking in reference to bias, variance, under/over fitting and sampling methods used. [10 Marks]

   b) Discuss the different roles that "weights" play in Boosting vs. Voting [8 Marks]

   c) In your view, what are the 3 most important modelling considerations when developing an ensemble model, and why (use examples)? [7 Marks]

6. Support Vector Machines and Artificial Neural Networks [25 Marks]

   The perceptron is an integral part of both the Neural Network and Support Vector Machine. Discuss the following (in the context of classification):

   a) Formally describe the fundamental perceptron components [6 Marks]

   b) Compare and contrast how it is used and trained in the context of BOTH a SVM and ANN, noting also how key limitations are overcome [12 Marks]

Pick **one** of c) or d) but do **NOT** do both

   c) Compare and contrast 2 methods how SVMs can handle multi-class classification problems [7 Marks].
   OR
   d) Compare and contrast the key differences between gradient descent vs. stochastic gradient descent [7 Marks]