BT5153 Topics in Business Analytics

Academic Year 2018/2019

Semester 2 Lecturer: DR ZHAO Rui

Midterm Assignment

Updated as of 5 Mar 2019

Important Instructions

- This assignment is released in IVLE on **15 March 2019, Friday** at **19:10.**
- This assignment is due on **15 March 2019, Friday** at **20:40**.
- This assignment consists of **11** Multiple Choice Questions (MCQs). Each MCQ consists of **4** options. There is **NO** negative marking.
- The total possible points awarded for the assignment is **25**. This will constitute **25%** of your overall grade for the module.
- You are to write your own code to attempt the tasks in the assignment. All the required techniques have been covered in the first half of the semester.
- You are to input your answers in this Google Form: https://goo.gl/forms/o2DI31Q52bvXzIXJ2
- *** Please take note that the options (a), (b), (c) and (d), as stated in midterm.pdf, do not follow accordingly in the Google Form. Please choose your answers carefully and not just blindly select them.
- **Please conduct a thorough check before you submit your answers as you are not allowed to resubmit them**. You will need to sign in to your Google account to fill out the response form so as to limit one submission per account. If you spot any discrepancies, please inform me as soon as possible so that we can rectify it.
- **Submit** your solution notebook to “**IVLE Workbin > Midterm Assignment”**. Failure to do so will result in the deduction of marks.
- Before you begin your attempt, please ensure that `scikit-learn` and `pandas` are properly installed on your system.

- The Jupyter Notebook `midterm.ipynb` can be found in “**IVLE Workbin > Midterm Assignment**”. You may use this to help you in your assignment. The data file `mcdonalds.csv` can be found in “**IVLE Workbin > Midterm Assignment**”.

- This is an individual assignment. Please attempt the assignment on your own, **NO DISCUSSION** is allowed.

- All the best for your assignment! I hope you learn something meaningful and useful through your attempt.

Assignment: McDonald’s Sentiment Data Analysis

Problem

McDonald’s receives thousands of consumer comments on their website every day and many of them are negative. Their corporate employees do not have the time to browse through every single comment, but they do want to read a subset that they are most interested in. In particular, articles about the rude service of their employees have recently surfaced on social media. In order to take appropriate action, they would now like to review comments about **rude service**.

You are hired to develop a system that ranks each comment by the **likelihood that it is referring to rude service.** They will use this system to build a “rudeness dashboard” for their corporate employees, so that the employees can spend a few minutes each day examining the **most relevant recent comments.**

Data

McDonald’s used the CrowdFlower platform to pay humans to hand-annotate approximately 1500 comments with the type of complaint. The list of complaint types can be found below, with the encoding used listed in parentheses:

- Bad Food (BadFood)
- Bad Neighborhood (ScaryMcDs)
- Cost (Cost)
- Dirty Location (Filthy)
- Missing Item (MissingFood)
- Problem with Order (OrderProblem)
- Rude Service (RudeService)
- Slow Service (SlowService)
- None of the above (na)

You will be asked to perform some tasks. In the midst of these tasks, some MCQs will be asked. You are to select the best possible option as your answer. Please answer them accordingly.

Task 1

Read `mcdonalds.csv` into a pandas DataFrame and examine it. (Instructions: `mcdonalds.csv` can be found in “IVLE Workbin > Midterm Assignment”)

A description of the more important columns to get you started:

- The **policies_violated** column lists the type(s) of complaint. If there is more than one type, the types are separated by newline characters.

- The **policies_violated:confidence** column lists CrowdFlower's confidence in the judgments of its human annotators for that row (higher is better).

- The **city** column is the McDonald's location.

- The **review** column is the actual text comment.

[Question 1]

Which option below gives the first sentence of the 303th review in the dataset?

(a) “Came here because I had to feed my daughter and well, we love McDonald's.”

(b) “Went here after work for a quick Mickey D fix.”

(c) “The Mickey D's breakfast vibe was strong in the work space and several of us were looking for some extra calories to fuel the early morning mental dash.”

(d) “Ok I’m waiting for like 10 minutes to place my order with the staff walking back & forth just looking at me like I’m crazy.”

Task 2

Remove any rows from the DataFrame in which the **policies_violated** column has a **null value.**

- **Note**: Null values are also known as “missing values”, and are encoded in pandas with the special value “NaN’. This is different from the “na” encoding used by CrowdFlower to denote “None of the above”. Rows that contain “na” should **not** be removed.
- **Note:** pandas.notnull() can return true if the object is not null and false if the object is null

[Question 2]

What is the shape of the DataFrame after removing the rows in which policies_violated has a null value?

(a) (1471, 10)

(b) (1525, 11)

(c) (1525, 10)

(d) (1471, 11)

Task 3

Add a new column to the DataFrame called “**rude**” that takes value 1 if the policies_violated column contains the text “RudeService”, and 0 if the policies_violated column does not contain “RudeService”. The “rude” column is going to be your response variable, so check how many zeros and ones it contains.

- **Note**: .iloc[] function can be used to select dataframe rows by position.

[Question 3]

What proportion of the DataFrame has reviews that are not complaining about rude service in the first 500 reviews?

(a) 66.3%

(b) 65.1%

(c) 65.8%

(d) 65.4%

Task 4

Define X using the **review** column and y using the **rude** column. Split X and y into training and testing sets (using the parameter `random_state=1`). Use CountVectorizer (with the **default parameters**) to create document-term matrices from X_train and X_test.

- **Note**: Please remember to follow the instructions carefully by setting the parameters as required for reproducibility of results.

[Question 4]

How many unique features do you arrive at after tokenizing X_train?

(a) 1103

(b) 1203

(c) 7300

(d) 7400

Task 5

Fit a Multinomial Naive Bayes model to the training set, calculate the **predicted probabilities** for the testing set, and then calculate the **AUC.** Repeat this task using a logistic regression model to compare which of the two models achieves a better AUC.

- **Note:** McDonald's requires you to rank the comments by the likelihood that they refer to rude service. In this case, classification accuracy is NOT the relevant evaluation metric. Area Under Curve (AUC) is a more useful evaluation metric for this scenario, since it measures the ability of the classifier to assign higher predicted probabilities to positive instances than to negative instances.

[Question 5]

What is the predicted probability of the 5th review in X_test belonging to rude service for the Multinomial Naive Bayes model?

(a) 99.7%

(b) 73.2%

(c) 26.8%

(d) 0.3%

[Question 6]

How much better is the AUC score under the Naive Bayes model as compared to that under the Logistic Regression model?

(a) The Logistic Regression model performs better.

(b) 0.0192

(c) 0.0221

(d) 0.0203

Task 6

Using Naive Bayes, try **tuning CountVectorizer** using some of the techniques we learned in class. Check the testing set AUC after each change, and find the set of parameters that increases AUC the most. (This is meant for your own learning experience, you may skip the tuning and go straight to Question 7.)

- **Hint**: It is highly recommended that you adapt the `tokenize_test()` function from class for this purpose, since it will allow you to iterate quickly through different sets of parameters.

[Question 7]

If you were to remove English stop-words, words whose document frequency (count) is lower than 4 and words whose document frequency (proportion) is higher than 0.3 in CountVectorizer, How many unique features do you arrive at?

(a) 7300

(b) 3237

(c) 1461

(d) 1732

[Question 8]

In Question 7 Model, what is the new AUC score you can achieve for X_test?

(a) 0.857

(b) 0.778

(c) 0.734

(d) 0.862

Non-coding Questions

For questions 9 to 11, please refer to the Section 2 in the google form.

You have reached the end of the assignment. Congratulation