BT4222 Mining Web Data for Business Insights
Academic Year 2016/2017 Semester 2
Lecturer: DR Jovian LIN

Midterm Assignment

Updated as of 21 Feb 2017

Important Instructions

- This assignment is released in IVLE on **10 March 2017, Friday** at **18:00.**
- This assignment is due on **12 March 2017, Sunday** at **23:59**. You are given the weekend to complete the assignment at your own time.
- This assignment consists of **15** Multiple Choice Questions (MCQs). Each MCQ consists of **4** options. Each MCQ carries **2 points**. There is **NO** negative marking.
- The total possible points awarded for the assignment is **30**. This will constitute **20%** 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/9on8j99dN61UYb5O2
- *** 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. I will publish on IVLE, a list that contains the names of those who have submitted from **12 March 2017, Sunday, 1100 onwards**. If you spot any discrepancies, please inform me as soon as possible so that we can rectify it.
- Before you begin your attempt, please ensure that `scikit-learn` and `pandas` are properly installed on your system. If you encounter any difficulties, please refer to the instructions under the “Useful Resources” section.
- 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”.
- Please attempt the assignment on your own. Of course, I would not be able to stop you from discussing and comparing your answers, but I am sure all of you know how the bell curve works.

- 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 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 128th review in the dataset?

(a) “I’m not a huge mcds lover, but I’ve been to better ones.”
(b) “Terrible customer service.”
(c) “It is what it is… so I’m mclovin’ it.”
(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.

[Question 2]

How many null values does the city column contain?

(a) 0
(b) 54
(c) 87
(d) 1525

[Question 3]

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

(a) (1471, 10)
(b) (1471, 11)
(c) (1525, 10)
(d) (1525, 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.

[Question 4]

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

(a) 65.8%
(b) 66.0%
(c) 66.2%
(d) 66.4%

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 5]

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

(a) 7100
(b) 7200
(c) 7300
(d) 7400

[Question 6]

How many non empty features are there in the 25th row of X_test?

(a) 3
(b) 21
(c) 40
(d) 64

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 7]

What proportion of the first five reviews in X_test is predicted to be of rude service?

(a) 20%
(b) 40%
(c) 60%
(d) 80%

[Question 8]

What is the AUC score under the Naive Bayes model?

(a) 0.734
(b) 0.778
(c) 0.809
(d) 0.842

[Question 9]

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) 0.0136
(c) 0.0192
(d) 0.0236

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 10.)

- **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 10]

If you were to allow for English stopwords removal, and set the following parameters as `max_df=0.3, min_df=4` in CountVectorizer, how many features do you get in your X_train?

(a) 1471
(b) 1732
(c) 2023
(d) 2536

[Question 11]

If you were to allow for English stopwords removal, and set the following parameters as `max_df=0.3, min_df=4` in CountVectorizer, what is the new AUC score you can achieve for X_test?

(a) 0.734
(b) 0.778
(c) 0.842
(d) 0.862

Task 7

The **city** column might be predictive of the response, but we are currently not using it as a feature. We will now explore to see if we can increase the AUC by adding **city** to the model. You are to do the following:

1. Create a new DataFrame column, **review_city**, that concatenates the **review** text with the **city** text. One easy way to combine string columns in pandas is by using the `Series.str.cat()` method. Make sure to use the whitespace character as a separator, as well as replacing null city values with a reasonable string value such as 'na'.
2. Redefine X using the **review_city** column, and re-split X and y into training and testing sets (using the parameter `random_state=1`).
3. By allowing for English stopwords removal, and setting the following parameters as `max_df=0.3, min_df=4` in the CountVectorizer, check whether it has increased or decreased the AUC.

[Question 12]

If you were to allow for English stopwords removal, and set the following parameters as `max_df=0.3, min_df=4` in CountVectorizer, is there any improvement in the AUC for this set of X_test that utilizes review_city over that mentioned in Question 11?

(a) Yes, by approximately 0.003
(b) Yes, by approximately 0.005
(c) No, the AUC decreased by 0.003
(d) No, the AUC decreased by 0.005

Task 8

The **policies_violated:confidence** column may be useful as it is a measure of the training data quality. You are to calculate the **mean confidence** score for each row of your McDonald's dataset (i.e. X_train together with X_test) and store these mean scores in a new column. For example the confidence scores for the first row are 1.0\r\n0.6667\r\n0.6667, so you should calculate a mean of 0.7778. Here are some of the steps you can follow:

1. Using the `Series.str.split()` method, convert the **policies_violated:confidence** column into lists of one or more "confidence scores". Save the results as a new DataFrame column called **confidence_list**.

2. Apply a function that can calculate the mean of a list of numbers, and pass that function to the `Series.apply()` method of the **confidence_list** column. Save those scores in a new DataFrame column called **confidence_mean**.

[Question 13]

How many reviews have a confidence_mean value of 1.00?

(a) 368
(b) 475
(c) 688
(d) 785

We will now like to remove lower quality rows from the **training** set to reduce noise. You are to remove all rows from X_train and y_train that have a **confidence_mean lower than 0.75**.

[Question 14]

How many reviews are you left in X_train that has confidence_mean higher or equal to 0.75?

(a) 569
(b) 799
(c) 939
(d) 1089

[Question 15]

Using the X_train that you have arrived at after filtering away confidence_mean values lower than 0.75, if you were to allow for English stopwords removal, and set the following parameters as `max_df=0.3, min_df=4` in CountVectorizer, what is the new AUC score you can achieve for X_test?

(a) 0.822
(b) 0.835
(c) 0.850
(d) 0.862

You have reached the end of the assignment. Congratulations!