Machine learning by Weka - Assignment Requirement

Mark as done

Machine learning by Weka - Assignment Requirement

Assignment Dataset: "assignment_data.arff"

Basic information of the assignment dataset - The dataset concerns TripAdvisor hotel reviews.
Availability: You can download the "assignment_data.arff" from Weka - Dataset folder.
Data structure: The dataset contains 3,118 rows of instances and 22 attributes.
Source of data: The dataset is a small sample extracted from a large open TripAdvisor Review dataset from: http://www.cs.cmu.edu/~jiweil/html/hotel-review.html.

Definition of the variables in the assignment dataset:
Hotel_id: The Id of the hotel
num_helpful_vote: The number of helpfulness vote that a review has received. **This measures how influential a review is.**
id: The id of instances
via_mobile: Whether a review is compiled on a mobile device (TRUE) or not (FALSE).
Revisit: Whether the customer revisited the hotel after writing this review. **Values: "Trigger_revisit"; "No_revisit"**
Rating_overall: Customer's overall rating on this hotel in the review.
Rating_service: Customer's rating on hotel service in the review.
Rating_cleanliness: Customer's rating on hotel cleanliness in the review.
Rating_value: Customer's rating on hotel value in the review.
Rating_location: Customer's rating on hotel location in the review.
Rating_sleep_quality: Customer's rating on hotel sleep quality in the review.
Rating_rooms: Customer's rating on hotel rooms in the review.
Rating_check_in_front_desk: Customer's rating on hotel check_in_front_desk in the review.
Rating_business_service_(e_g_internet_access): Customer's rating on hotel business service in the review.
author_id: id of review author/customer
Author_username: Username of the review author
Author_num_cities: the number of cities the review authors have visited before. The larger the number, **the more influential the review author is in real life.**
Author_num_helpful_votes: the number of helpfulness votes the review authors have so far visited based on all the reviews s/he produced before. The large the number is, **the more influential the review author is in online environments.**
Author_num_reviews: The number of reviews that the review authors have produced at TripAdvisor.
Author_location: The location where the review author is living at.
title: The title of a review
text: The review text

Hints:
1. Please pay attention to the variable format. A variable (e.g., rating_overall) set to be a numeric variable would yield different results by converting it to a nominal variable.
2. Please pay attention to the variables included in the analysis. J48 does not allow its predictors to include a string variable. In other words, if a string variable is included in the variable list, J48 cannot be activated. Other machine learning algorithms may have similar requirements.
3. Please consider using the attribute selection function if you want to include text mining in your analysis. You will likely have a big data problem if you turn all the words of reviews into attributes and include them all in the analysis.
4. SMOTE can be a very good solution dealing with unbalanced data.

Assignment Task Requirements:
You are now applying for a job as a machine learning analyst at a company. The employer requests the shortlisted applicants (you are one of them) to analyze a dataset and offer a report, which will be used as the basis for assessment and for making the final recruitment decision. Thus, you aim to offer an attractive report to the employer to demonstrate your ability to do machine learning. The evaluator of your report is expected to be someone with a basic knowledge of machine learning.

You are given a list of possible tasks to select from analyzing the data.
- Business scenario: The company actually purchased a large amount of similar reviews from Facebook on the company, which are pure texts without giving any numeric ratings on the service product. The company wants you to develop machine learning models and use the models derived from TripAdvisor review data to understand Facebook reviews.
1. Using review title and text to model the **hotel's overall ratings or ratings on the hotel's different attributes** (e.g., service, location, etc.)
- Please choose to model overall rating or one of the attribute ratings (e.g., service, location). You don't need to model all different attribute ratings.
- You may consider the cost of classification. For instance, classifying a low-rating (high-rating) review to a high-rating (low-rating) review will incur a relatively high (low) cost. You are free to decide whether to include a cost matrix in the analysis, as well as to decide the values in the cost matrix.
2. Using review title and text to predict how **influential a review** will be **[This question is slighted updated on March 28 by removing irrelevant text and adding a explanation below]**.
- The company is very interested in the reviews that received more than 15 helpful votes. You predict whether a review would receive more than 15 helpful votes or not (Yes vs. No).

3. Utilize all the variables available (you are free to choose which ones to use) how influential the review author is in real life **(Author_num_cities).**
- The company is very interested in the review authors who visited more than 15 cities in real life. You predict whether a review author has visited more than 15 cities or not (Yes vs. No).
4. Utilize all the variables available (you are free to choose which ones to use) to predict how influential the review author is in the online environment **(Author_num_helpful_votes).**
- The company is very interested in the review authors whose reviews received more than 100 review helpfulness vote. You predict whether a review author has visited more than 100 review helpfulness votes (Yes vs. No).

For all the tasks, you are free to decide whether to include a cost matrix in the analysis, as well as to decide the values in the cost matrix. If you use a coast matrix, please briefly explain the reasons for using the cost matrix, which should make sense from a business perspective.

Please specify the tasks you want to address in your assignment report, and you can decide which task(s) to address! You can choose a combination of tasks, such as, two tasks from business scenario 1 and the task of business scenario 2.

You could use other tools for data processing, such as Excel or Python. However, please only use WEKA for machine learning!

Assignment Format Requirments:
- Page limits: No more than **20** pages. You can include screenshots of the Weka if you want.
- Content: You are free to decide what content to include, e.g., pictures or screenshots. Mathematic formulas are not encouraged!
- **Pdf**: Please submit the report in PDF format.
- **Anonymity**: Please don't include your personal information (e.g., name or student ID) in the report or in the file name.

Assessment Methods:
We will use the **peer assessment** method for the assignment. Each student will request to evaluate **four randomly assigned** assignments of other students. You will learn by evaluating how other students do the tasks.

AI use policy: Using AI to proofread the text and to generate ideas is fine, but copying and pasting the ideas generated by AI into the report **is not allowed!** The use of AI in the assignment should be transparent by openly describing how AI is used in the assignment.

Assessment rubric
1. Key assessment dimensions:
- **Knowledge**: The demonstrated knowledge of using different machine learning methods**.**
- Performance/Accuracy: The performance of the computed machine learning model
- **Informativeness:** Detailed information is offered in the report on how the models were constructed and compared.
Note: please evaluate the report as if you are the employer who will decide on the candidate to recruit.
2. Grading scale: 1 - 5
- 1 (Poor); 2 (below average), 3 (average), 4 (very good), 5 (Excellent)
3. Final grade: The final grade will be the average of the received grades, after removing the lowest grade.
Note that: as not all the students will complete the course, it may happen that some assignments will receive less than four assessments. We will try to guarantee that each assignment receives at least three peer assessments.

Deadline for the assignment submission: April 08
- Please submit the assignment before the deadline.
- Penalty for late submission: i) less than 3 hours, a penalty of 5%; ii)less than 24 hours, a penalty of 20%; iii) 24 - 48 hours, a penalty of 40%; iv) over 48 hours, the assignment will not be evaluated.
Because the peer assessment method will be used, a late submission after 48 hours will not be accepted.

Deadline for peer assessment submission: April 17.
- Peer assessment will start on **April 11** (~12:00).
- Please evaluate **four** reports of other students.
- No doing peer assessment will lead to a minus of **0.25 point (out of the total 5 point for weka assignment) for each report not assessed**.

Previous activity
◄ Learning Diary Requirement and Evaluation Rubric

Next activity
Weka Assignment Submission and Peer Assessment ►