STATS 202: Data Mining and Analysis

Instructor: Linh Tran

FINAL PROJECT

Due date: August 14, 2019

Stanford University

Introduction

The goal of this class project is to give you experience in real life statistical analyses and data mining. By the end of the project, you will have learned how to identify and interpret types of different attributes in a dataset, visualize the attributes and relationships between attributes of different types, understand how those relationships could affect your models and analyses, and finally build regression and classification models. Throughout the project, I will be available via e-mail to help out and provide advice.

The final project is broken up into four parts, each with their own separate objectives. To complete the project, you are expected to satisfy all four objectives. Each of the objectives are explained in further detail below.

The class project is worth 200 points and is optional. It can replace your score in the Final Exam if you choose not to take it. If you choose to also take the Final Exam, your score for the Final Exam portion of your grade will be the maximum of the Final Exam and the Class Project.

Background

You will be working with real data from five randomized controlled trials for patients with schizophrenia. The same (anonymized) drug is being evaluated over all five trials for its efficacy in treating schizophrenia. Patients in the trials are followed for varying amounts of time (depending on the criteria of the study) and observed for symptoms related to schizophrenia.

At a high level, patients with schizophrenia are screened to decide if they meet the criteria to participate in the study. Upon satisfying the requirements and enrolling in the study, patients attend an initial visit (i.e. visit day 0) where a baseline measurement is taken. At this time, they are randomized into one of the two treatment groups. Throughout the study (which ranges from weeks to months), the patient comes back for follow-up visits to have the same measurements repeatedly taken.

The outcome being measured is a standardized scale commonly used for measuring symptom severity, called the Positive and Negative Syndrome Scale (PANSS). The scale is known as the "golden standard" that all assessments of antipsychotic behavioral disorders should follow.[1] The PANSS is measured during a relatively brief interview, requiring between 45 to 50 minutes to administer.[2] The patient is rated a score on 30 different symptoms based on the interview as well as reports of family members or primary care hospital workers.[3] Before being able to assess the patient, the interviewer must be trained to a standardized level of reliability.[4]

The PANSS has three classes of items that are assessed: Positive symptoms (7 items) which refer to an excess or distortion of normal functions (e.g. hallucinations and delusions), Negative symptoms (7 items) which represent a diminution or loss of normal functions, and General Psychopathology symptoms (16 items). Each PANSS item is rated on an ordinal scale from 1 (i.e. absent) to 7 (i.e. extreme). Details on each of the items can be found within the references (e.g. PANSS Institute[5]).

Despite attempts to maintain high reliability, there can be a noticeable amount of occurrences of bad ratings. Examples include the patient assessment (as a whole) not making any sense, assessments that are inconsisent with previous ratings, and an outcome assessment trajectory that is infeasible. Consequently, clinical auditing firms are typically hired to validate the collected patient assessments. Assessments that are potentially erroneous are either flagged for review or assigned to a clinical specialist for follow up and confirmation.

Data

You will be provided 5 data sets, corresponding to each of the 5 trials that were run (Study A - E). The data sets contain unique rows for each PANSS assessment conducted on a patient within the study. A "unique" assessment is comprised of the patient being assessed, the rater making the assessment, the site that the assessment is made at, and the date of the assessment (all within the study that is being conducted).

A unique identification number is given to each unit that can be identified uniquely. This includes the patient being assessed, the rater making the assessment, and the site that the assessment is made at. Additionally, a unique identification number is given to each unique assessment made. Other attributes include the patient's country, their (randomly) assigned treatment group, and the relative day of assessment (compared to the initial baseline visit). Outcome variables include ratings on each of the 30 PANSS items, the sum of the 30 ratings, and the outcome of the assessment's audit.

Keep in mind that, as with most real world data, there are a number of nuances that have to be accounted for. Examples include (but are not limited to) patients being assessed multiple times in the same day by different raters or at different sites, patients skipping some of their assessments or dropping out of the study prior to the end, or patients being assessed at odd intervals of time.

As part of the four objectives, you will be asked to make predictions on Study E for the total PANSS score (at the 18th week) as well as the audit outcomes for all of the assessments made. Consequently, the data set for Study E does not contain these values.

Please download the data from the links provided at the course website. The list below describes the explicit variables included in the data:

1. *Study* - A character indicating which of the five studies the data represents.
2. *Country* - The country where the assessment was conducted.
3. *PatientID* - An identification number given to each unique patient.
4. *SiteID* - An identification number given to each unique assessment site.
5. *RaterID* - An identification number given to each unique rater.
6. *AssessmentiD* - An idenfication number given to each unique assessment conducted.
7. *TxGroup* - A string corresponding to the patient's (randomly) assigned treatment group.
8. *VisitDay* - An integer corresponding to the number of days that have passed since the baseline assessment.
9. *P1-P7* - The scores corresponding to each of the 7 positive symptoms of the assessment.
10. *N1-N7* - The scores corresponding to each of the 7 negative symptoms of the assessment.
11. *G1-G16* - The scores corresponding to each of the 16 general psychopathology symptoms of the assessment.
12. *PANSS_Total* - The sum of of the ratings across the 30 PANSS items.
13. *LeadStatus* - A string indicating whether the assessment's audit passed, was flagged, or was assigned to a CS (i.e. clinical specialist).

Write-up

As part of the final project, you are expected to submit a final report (in PDF format) covering the approaches, details, and results of each of the four objectives. The report should be no longer than 10 pages (excluding figures, tables, and code) and capture the steps you took throughout the data mining process (Figure 1). Table 1 provides further details on each step of the process. **Make sure to reference your team's Kaggle leaderboard name in your report**. The code used to generate the results should be either attached as additional scripts in your final project submission, appended to the end of your report as an appendix, or referenced to in the report via a link to the uploaded git repository. Note that 10% of your grade will be based upon the organization, readability, reproducibility, and efficiency of your code.

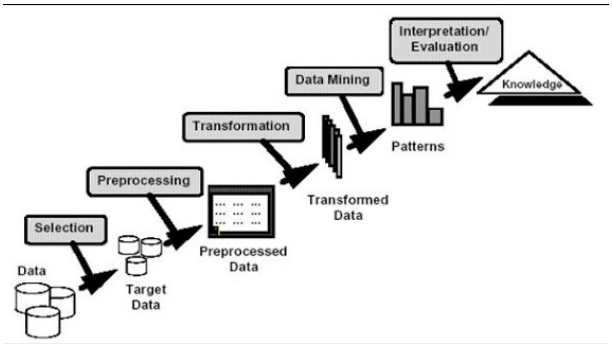

Figure 1: Steps in the data mining process.

Selection	Explain which attributes you used to build the model, and why you chose those attributes
Preprocessing	Explain whether you pre-processed any of the attributes by modifying them in any way
Transformation	Explain whether you created new features from the existing attributes, or from pairs of the existing attributes. Did you transform any of the attributes into another representation of data? Remember, you do not need to use all of the attributes in your model. Try to evaluate which attributes you think will be useful, and use those attributes.
Data Mining	Explain how you built your regression / classification model. There are many kinds of models that may work for this problem. You are welcome to use whatever regression / classification approach you would like, but remember, you need to end up with a prediction or probability.
Interpretation/Evaluation	Understand what your model is doing and how it is performing. This may require you to separate your training data into different groups so that you can test your models performance on a “hold out” group.

Table 1: Steps in the data mining process.

1 Treatment effect (35 points)

Does the (anonymized) treatment have an effect on schizophrenia? Your goal is to use the data to make a case for or against the effect of treatment. Note that this objective is intentionally left somewhat ambiguous, as to give you flexibility on how to evaluate the treatment effect.

Disclaimer: Typically, in regulated industries (e.g. the pharmaceutical sector), a statistical analysis plan has to be submitted a-priori stating the exact details of the experiment and the planned statistical analyses to conduct. Once the data is collected, only the statistical analyses that have been approved can be conducted. Conversely, here we have already run our experiment and you are being asked post-hoc to conduct the statistical analyses. This opens the potential issues of *data snooping* and conducting multiple statistical tests, leading to a larger than expected Type I error rate. Your write-up should state clearly how you are evaluating the treatment effect, what (if any) other statistical tests or data analyses you did before arriving at your conclusions, and your conclusions regarding the effect of treatment.

2 Patient segmentation (55 points)

Stakeholders typically want to understand the distribution of the population that they are working with, in an effort to better serve them. Specifically, it is (often) valuable to understand the different types of patients with schizophrenia that are being treated by the proposed treatment. To get a clear idea of this many times stakeholders will ask to have the population segmented into k groups, assign descriptions to each of the k groups, and to summarize the proportion of the population that each group is comprised of.

Your goal is to segment the schizophrenia patients into these k groups (where you decide on the value k) and to describe each of the groups, as well as the approach and settings that were used to assign patients to each of the groups. In doing so, you should only be relying upon the baseline measurements for each patient (i.e. their visit day 0 measurements), as this will capture their functional status prior to experiencing any (potential) treatment effects. You should justify all of the decisions made, from the value k used to why you are describing the groups the way that you are.

3 Forecasting (55 points)

Because clinical experiments typically cost millions of dollars to run, stakeholders will typically want to predict and understand the potential trajectories of patient assessments prior to running the study. Approaching this in a more simplified manner, you are being provided all of the patient assessments leading up to the final 18th-week assessment for Study E. Your goal is to predict the total PANSS score (across the 30 PANSS symptoms) for the 18th- week assessment. Specifically, you will be putting together a csv file that contains the `PatientID` and the predicted 18th-week PANSS score and submitting the file to the *Kaggle leaderboard*. *For reference, a sample file has been included with the data demonstrating what your submission should look like.*

During the course of the final project, the (public) leaderboard will show your results on (a random) 25% of the patients. At the end of the final project, the (private) leaderboard will show you the results on the remaining 75% of the patients.

Your write-up should include the data processing that was done, the features used, the modeling approaches chosen, and the results of your work (including your Kaggle username and public leaderboard score).

Note that patients can drop out of the study prior to their 18th week. Consequently, we observe 18th-week assessments from only 379 (of the 512 total) patients in Study E. Due to limitations from the Kaggle leaderboard, you are only allowed to submit predictions on these 379 patients. Please refer to the sample submission file for the corresponding list of the 379 patients to submit predictions for.

4 Binary classification (55 points)

Having humans audit all of the PANSS assessments can be time consuming and expensive. It is therefore reasonable to wonder if a machine learning algorithm can reliably predict which of the assessments are erroneous. Having a reliable algorithm would allow the clinical auditor to focus only on the subset of

assessments that have issues, instead of all of the assessments uniformly. Therefore, your goal is to predict which of the assessments in Study E will be either flagged for review or assigned to a CS.

Specifically, you will be putting together a csv file that contains the `AssessmentID` and the probability of the assessment being either flagged or assigned to a CS. The file will be submitted to the *Kaggle leaderboard*. Similar to Objective 3, the public leaderboard will show your results on (a random) 25% of the assessments during the course of the final project and the private leaderboard will show results on the remaining 75%. *For reference, a sample file has been included with the data demonstrating what your submission should look like.*

Similar to Objective 3, your write-up should include the data processing that was done, the features used, the modeling approaches chosen, and the results of your work (including your Kaggle username and public leaderboard score).

References

[1] M. G. Opler, C. Yavorsky, and D. G. Daniel, "Positive and negative syndrome scale (PANSS) training: Challenges, solutions, and future directions," *Innovations in Clinical Neuroscience*, vol. 14, no. 11-12, pp. 77–81, 2017.

[2] S. R. Kay and L. A. Qpjer, "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia," *Schizophrenia bulletin*, vol. 13, no. 2, pp. 261–276, 1987.

[3] S. R. Kay, *Positive and Negative Syndromes in Schizophrenia: Assessment and Research*. New York: Brunner/Mazel Inc., 1991.

[4] L. A. Opler and P. M. Ramirez, "00741069-199805000-00005.pdf," *Jrnl Prac Psych and Behav Hlth*, pp. 157–162, 1998.

[5] PANSS Institute, "Positive and Negative Syndrome Scale (Panss) Rating Criteria," Tech. Rep. 2, 1987.