MSSP Portfolio

Yifan Liu

Department of Statistics

Boston University

Contents

Model Selection for Behavioral Studies

Introduction

The client is a student from Biomedical Engineering Department of Boston University. She is working on a dataset obtained from a behavioral study to look at people's learning behavior. The task she conducted consists of 4 (or 3 for some subjects) blocks. In each block, the subject learned what the targeted response is for the 8 stimuli (4 visual and 4 auditory) based on feedback (correct/incorrect). There are two keypads, one for visual stimuli and the other for auditory stimuli, which switched between blocks (e.g., L-aud, R-vis for block 1 and L-vis, R-aud for block 2). The 4 auditory stimuli correspond to up, down, left, and right keys respectively and the 4 visual stimuli correspond to left, right, up, and down keys, respectively. The stimuli were presented in random order in each block for around 200 iterations. The rules changed between blocks but can be the same.

The client is using RLWM model which is a combination of RL (Reinforcement Learning) and WM (Working Memory) models to predict whether or not the subject will press the correct key given a particular stimulus at a particular iteration. The client wants to know which loss function should be used for the model and how to use BIC and exceedance probability to filter parameters to yield better prediction accuracy and model interpretability. Also, she would like to analyze the evolution of each parameter value throughout the experiment.

Data and Methods

We first examined the logic behind the model selection in the RLWM settings and proposed BIC to be the appropriate procedure. The RLWM model will initialize Q matrices of two models at Q=1/n and then use prediction error to update the Q matrices and weight. Then the model will calculate the probability matrices based on Q matrices using the SoftMax function. The final probability is the weighted average over the RL probability and the WM probability. So, the loss function which measure the difference between the experimental outcomes (correct/incorrect) and the estimated probability should be minimized to update parameters.

The BIC is short for Bayesian Information Criterion, a criterion for model selection among a finite set of models. Bayesian model comparison is immune to overfitting by imposing a penalty on more complex and flexible models. The classical model comparison provides an inflated measure of how well a model predicts a dataset. Since the data obtained by the experiment is timeseries and in timeseries data, it is difficult to define a second dataset is truly independent of the first, cross validation is not recommended in this area. Likelihood ratio test cannot be used to compare model that are not nested in one another either.

Results and Discussion

We decided to recommend using BIC to do model selection. First, we suggested that the client should establish RLWM models M using different sets of parameters. Then we can use BIC to approximate model evidence. According to Bayes rule, the posterior probability of a model is proportional to the product of two factors: the likelihood of the data and the prior probability of the model. Since the model evidence is intractable and must be approximated, we recommended to use BIC as the approximation method. Other penalized score like AIC (Akaike information criterion), though this has a similar form to BIC, is not advocated because it does not arise from an approximation to the log of the model evidence and thus cannot be used to approximate Bayes factors. Then we can use Bayes factor to compare models. When comparing two models, we hope to get a statistical claim about the relative fit of one model over another. A standardized measure is the Bayes factor, which seems to be the most reasonable and standard metric to report, defined as ratio of their posterior probabilities.

Conclusion

In the client's study, the response variable is binary, and we can regard it as a logistic regression where the dependent variable is a binary variable (whether the action is correct or not) and the independent variable is the learning history. For the model selection method, BIC is an appropriate method we recommended.

Exploration of Double Negation in Russian Language

Introduction

The client is interested in exploring the availability of Double Negation (DN) readings in Russian. DN readings are not supposed to exist in Russian; however, working with one other linguist, the client has indeed found DNs in Russian when the negation element “ni za chto” is not fronted.

The client ran a survey with two different versions which contain the same sentence structure but under different context. By inserting target sentences into specific contexts, the client wanted the participants to rate how well the target sentence fit in on a scale of 1 to 6. There were 49 participants in total, all of which were adult native Russian speakers. (24 for version 1 and 25 for version 2). The client also included other variables such as the broad or narrow focus, naturalness of the target sentence, whether the negation word, “ni za chto” is fronted or not, and single negation or double negation etc.

Our client needs help with data visualization and guidance on statistical analysis of the data for statistical significance.

Data and Methods

From our understanding of the project, the client would like to explain the availability of DN readings in Russian, and specifically would like to compare the fitting scores of Double Negations (DNs) and Single Negations (SNs) and compare 2 negation words in different situations.

We first read-in the dataset and then created two separate datasets for the two interested question we wanted to explore. For each question, we used four different types of plots for visualization and exploration of the data. The four types of plots are: violin plot, violin plot combined with box plot, pie chart, and ridge plot. At last, we provided some additional resources for exploring statistical significance and how to implement in R. To answer both questions, since the fitting in scores are ordinal dependent variables, we would recommend starting from the ordered logistic regress (also called proportional odds model).

Results and Discussion

The violin plots are a method of plotting numeric data and can be considered a combination of the box plot with a kernel density plot. Compared to the box plot, violin plots can also show the entire distribution of the data aside from showing the general statistics. The wider sections of the violin plot represent a higher probability of observations taking a given value, the thinner sections correspond to a lower probability.

The points in the plots are jittering points of the real data. By adding random noise to data (jittering) we can prevent overplotting in graphs below. Because there are multiple repeated values in the real data, plotting without jittering will result in scores with the same value overlapping with each other. Jittering helps us to better visualize the density of the data and to find clusters in the data. After combining violin plots with box plots, statistics such as median, interquartile range, and the lower/upper adjacent values are clearly represented.

Pie charts are useful for displaying data that are classified into nominal or ordinal categories. Nominal data are categorized according to descriptive or qualitative information while ordinal data are similar, but the different categories can also be ranked. Pie charts are generally used to show percentage or proportional data and good for displaying data for around 6 categories or fewer.

Ridge plots show the distribution of a numeric value for several groups. Distribution can be represented using histograms or density plots, all aligned to the same horizontal scale and presented with a slight overlap. Ridge plots work well when there is a clear pattern in the results, which in our case, the main distributions of fitting scores corresponding to different expressions.

Conclusion

When comparing the fitting scores of Double Negations (DNs) and Single Negations (SNs), from the plots we can tell in version 1 survey, the expressions with 1 negation word in narrow-fit (whether fronted or not) have generally higher fitting score than expressions with 2 negation words not fronted in narrow-fit. Most of expressions with 1 negation word in narrow-fit (whether fronted or not) have fitting scores distributed at 6, while expressions with 2 negation words not fronted in narrow-fit have a broader distribution from 1 to 6. In version 2 survey, we can tell that most of expressions with 1 negation word in narrow-fit (whether fronted or not) and expressions with 2 negation words not fronted in narrow-fit have fitting scores distributed at 5 and 6. Generally speaking, compared to version 1 survey, version 2 survey has more fitting scores distributed between 5 and 6 which in graph, we can say that the violins in version 2 are more compact.

When comparing 2 negation words in different situations, from the plots we can tell in both version 1 and version 2 survey, expressions with 2 negation words have fitting scores cluster differently. To be more specific, expressions with 2 negation words not-fronted in double-negation-narrow-fit have median score at 5, first quartile score at 4 and third quartile score at 6 in version 1 and have median score and first quartile at 5 and third quartile at 6. However, expressions with 2 negation words not-fronted in single-negation-narrow-fit have median score and first quartile score at 1 and third quartile score at 2 in version 1 and have first quartile score at 1 and median and third quartile score at 2 in version 2.

Trinity Project

Abstract

Statistical learning methods were applied to the COVID-19 related text dataset on the web. According to the World Health Organization, the coronavirus (COVID-19) was worsening around the globe and has resulted in thousands of lives lost and significant incremental cost to the healthcare system. The COVID-19 related text data have enormous potential to drive epidemic-fighting progress to success. A variety of learning techniques were explored and validated.

Introduction

To explore the COVID-19 related text dataset on the web, we were divided into two directions. Direction 1 was using text to understand impact of COVID-19 on healthcare industry. Under Direction 1, two tasks were created: Task 1 was focusing on text mining on COVID-19 academic research using data from PubMed which is a free search engine accessing the Medline database of references and abstracts on life sciences and biomedical topics; Task 2 was focusing on analysis of telemedicine-related topic using data from Reddit.

Direction 2 was exploring the contents that are related to COVID-19. Specifically, Direction 2 found people's considering trends during pandemic and explored their correlations and coincidence with COVID-19 spreading. Also, Direction 2 analyzed sentiment attitudes of such trends using data from Twitter and Reddit.

Data and Methods

For Task 1 under Direction 1, we collected the data on relevant topic from PubMed and created vocabulary frequency table for words used by researchers. Also, we conducted sentiment analysis and clustering on our data.

For Task 2 under Direction1, we collected data on relevant topic on Reddit and used LDA to extract main topics. Then we conducted sentiment analysis on the data that had been collected.

For Direction 2, we created an auxiliary tool to narrow the range of research. The tool used data from COVID-Related Tweet IDs and detailed Tweets Info, then stored in SQLite Database. Then we used simple NLP to do sentiment score calculation, computed keywords frequency statistics, and reversed geocoding with Bing. The display of our tool was completed in the interface by Shiny App.

Results and Discussion

The results were mainly displayed in word frequency tables such as word clouds, sentiment analysis plots, and Shiny App. However, there were still certain limitations in our study. For example, most researchers tend to not include any emotions and to be subjective which makes the sentiment analysis to be difficult. Also, tweets themselves are not independent and there would be correlations and be influenced by the network structure.

Conclusion

For Direction 1 Task 1, based on the results we provided above, we can conclude that: researchers are interested in topics such as the patients' clinical response, the medical research, and the relevant information in the hospital. Words like patients, clinical, hospital and medicine appear very often. For the keyword "China", words like characteristics, Wuhan, center, and cases appear very often. In other words, researchers are interested in patient characteristics and solutions about the epidemic in China, because China is the first country in the world to experience and control the epidemic. According to our result, over the five types of vaccines, it seems that Pfizer's COVID-19 vaccine is the most promising vaccine by researchers. In fact, Pfizer's COVID-19 vaccine is exactly the vaccine authorized by FDA and recommended by the CDC for use in the US for a limited population right now. Through clustering, one interesting point we find is that the researcher's focus and interests are constantly changing.

For Direction 1 Task 2, after analyzing the top 10 active users, we found that most of them have a positive attitude towards telemedicine. For people who have a positive attitude, themselves, or someone close to them, they have a mental illness or chronic illness, like anxiety and migraines, which means that they need to go to the hospital periodically for the treatment. For people who have a negative attitude, had some experiences of waiting for a long time when getting telemedicine service instead of traditional one or the elderly in their family do not know how to get diagnosed online, which means they have experienced telemedicine, but found it inconvenient or unable to get the same quality of service over the phone call or the online video as they did in the hospital. According to the analysis above, we provided some business suggestions for the healthcare providers: it might be beneficial to look at patients who have mental illness or chronic illness based on their current personal profiles. And people are worrying about the quality of services in telemedicine. Therefore, we need to seriously enhance services to make it easier for people to experience telemedicine. For example, a mobile application can be developed to provide in-time response and communication by professional teams and to record patients' detailed medical information for further consultations.

For Direction 3, we created our database on a network disk and uploaded all data and the written R files on GitHub. Also, Shiny App was published.

Statistical Analysis of Burial Effects on Fetal Piglets Decomposition

Introduction

This project is about providing statistical analysis to the research of burial effects on fetal piglets' decomposition conducted by our client from the Department of Anatomy and Neurobiology of Boston University. The client is interested in analyzing the effects of the burial depth and plastic wrapping on the decomposition of the fetal piglets so that the findings of this experiment can be applied to human remains related studies. The client is also interested in comparing the effects of these two variables to determine if one has a more significant effect on the decay rate.

To conduct the experiments, 56 fetal piglets were weighted and buried at the same time in 7 different points in the same geographical location. The client is recording the average mass of the piglets and a scoring variable Total Body Score (TBS) which measures the condition of the remains as units and using the burial depth and plastic wrapping as variables. The burial depth is divided into two groups: shallow (20 cm) and deep (60 cm). And the plastic wrapping is also divided into two groups which indicate the remains are wrapped with plastic or not before burial. The total experiment lasts for 1 year and the buried piglets will be unearthed 1 month after the burial, 2 months after the burial, 3 months after the burial, and 6, 9, 12 months accordingly. After each set of piglets were unearthed, the client will record the corresponding burial depth, wrapping or not, average mass loss, TBS score, the percentage of adipocere, and the percentage of skeletal.

Data and Methods

In order to explore the significance of effects of burial depth and wrapping conditions, we first conducted Exploratory Data Analysis (EDA). After the EDA, we decided to use the percentage loss in mass and the Total Body Score (TBS) as our target variables and we also create a dummy variable to indicate whether the burial time of each sample is greater than 6 months or not. To prepare the data for modeling, we cleaned and re-organized the raw data and removed two outliers which have negative mass loss and one high leverage point. Also, we used 1 to represent the wrapping condition "wrapped", burial depth "shallow", burail time greater than 6 months, and 0 to represent the wrapping condition "unwrapped", burial depth "deep", and burial time less than or equal to 6 months.

The model we decide to use is the mixed-effects model and the predictors are the burial depth, wrapping condition, the interaction term of burial depth and wrapping condition, and dummy variable burial time. Also, location is classified as random effects and the burial depth, wrapping condition, and burial time are classified as fixed effects. After the EDA, we have found that both the intercepts and slopes between predictors and target variables are random. In other words, each location can

have their own intercept, random slopes influenced by wrapping condition and burial depth, and their interaction between wrapping condition and burial depth. Importantly, all random slopes and intercepts are correlated. We used two estimation methods are maximum likelihood estimation (MLE) and Bayesian Estimation.

Results and Discussion

For the MLE methods, we were using the Q-Q plot to validate the assumptions and the residual plot to check the goodness of fit. From the Q-Q plots for both models we can tell that all points roughly fall into a straight line which indicates the assumption of normality is satisfied. According to the residual plots for both models, we can tell that the residuals were evenly distributed around the center line and there are no particular patterns or cluster which indicates our models had good fits.

For the Bayesian estimation, we were using R-hat and effect size to check the convergence of our model and graphical posterior predictive check and residual plots to validate our models. With the R-hat less than 1.1 and effective size is greater than 0.5 we could say that the assumption of convergence is satisfied. In the posterior predictive check plots, the actual values by our models were completely covered by the predicted values which indicates our models performed well. In the posterior predictive interval plots, the actual values were completely fall in the range of the intervals of predicted values which also indicates our models performed well. The residual plots also showed that both of our models have good fits.

Conclusion

When using the MLE method, we can use the p-value as an indicator for the statistically significance of variable effects. Wrapping condition has a statistically significant effect on the Mass Percentage Loss but not on the Total Body Score at a significance level of 0.05. Depth has a statistically non-significant effect on the Mass Percentage Loss and Total Body Score at a significance level of 0.05. (However, it has a statistically significant effect on the Total Body Score at a significance level of 0.1). The interaction between Wrapping and Depth has a statistically non-significant effect on the Mass Percentage Loss and Total Body Score at a significance level of 0.05. The Bayesian Estimation provided the same conclusions as above.

Impact of Race on Cancer Therapy

Introduction

Our client is interested in knowing what sample size is needed in order to perform a study looking into the effect of race on the survival of patients with stage 3 and stage 4 melanoma. Sample size can be determined using power analysis with a given effect size, level of significance, and power of study desired. For this project, since our client did not provide an effect size that could be used to calculate sample size and since there are some uncertainties regarding the population sizes of the two groups being studies, we approached situation from a different angle.

Data and Methods

We used two power tests for this analysis: power test for equal sized populations and power test for unequal population sizes. For the purpose of this analysis, the proportions are defined as the following: Population 1 – Caucasian melanoma cancer patients and Population 2 – Non-Caucasian melanoma cancer patients. The null hypothesis of our study is: There is no difference in treatment outcome between population 1 and 2; while the alternative hypothesis is: There is a difference in treatment outcome between population 1 and 2. We also defined effect size as the standardized mean difference between two groups. The significance level we were using is 0.05 and the power of the test is 80% (as minimum and standard used for study to be statistically powerful).

In order to perform power test, we needed to assume what the proportion of survival outcomes for population 1 and the proportion of survival for population 2. Then we subtracted these proportions from each other which will give us the smallest difference that can be detected between the two populations given the significance level, and 0.8 power. Then we will use the R function to perform the test.

Results and Discussion

Suppose for proportion 1, 75% of Caucasian population survived from the treatment; as for proportion 2, 48% of Non-Caucasian population survived from the same treatment. The results show that with a sample size of 50 for each population, a difference in outcome between the two proportions will need to be at least 27% for results to be meaningful. Actually, as the sample size gets larger, the effect size, or difference between the populations detected, can be small and results will still be meaningful. If a small difference in outcome is detected between the populations, a larger sample will need to have been used for the results to be meaningful.

As for two unequal population sizes, the test would provide the second population needed if only one population size is known. If the size of population 1 is 50, the size

of population 2 will need to be at least 84 for a medium effect size of 0.5. If the size of population 2 is very small, a large effect size will need to be detected. Therefore, although the size of population 2 can be small, the difference in outcome between the two populations will need to be 0.8 or more for results to be meaningful. According to our study, with sample sizes less than 250 for each population, a difference between the two populations studied of greater than 20% will be needed.

Conclusion

The above analysis shows that although a small sample size of 50 can be used for each of the populations, or for one of the populations, a large difference in the treatment outcomes between the two populations studied, will need to be detected. Our client did mention the possibility that population 2, non-Caucasian melanoma patients, may be used. We would like to kindly point out that although a small sample size for the second population can still be used with a power of 80% of the test, a large effect size, as defined by Cohen, will need to present itself in the results for any test statistic to be considered meaningful.