Capstone Project Proposal
Machine Learning Engineer Nanodegree

Shohei Nagamine
Nov 2, 2017

Domain Background

It's really unhappy when I couldn't have adequate sleep. It makes me forgettable, slow, short-tempered and so on. Several studies also show the importance of getting adequate sleep. According to an article of CBS NEWS, drivers who get just one to two hours less than the recommended seven hours of sleep in a 24-hour period nearly double their risk for a car crash (*1). Some studies reveal that sleeping five hours or less per night increased mortality risk from all causes by roughly 15 percent (*2). Moreover, according to an article of "Sleep, Learning, and Memory" in Healthy Sleep, a research suggests sleep itself has a role in the consolidation of memory, which is essential for learning new information (*3).

Therefore, taking time for sleep is obviously important. However, although we spend enough time for sleep, sometimes we cannot sleep well and cannot get the advantages of it. According to the National Institute of Neurological Disorders and Stroke, about 40 million people in the United States suffer from chronic long-term sleep disorders each year and an additional 20 million people experience occasional sleep problems (*4). Therefore, it seems to be valuable to know what brings us good sleep. That way, a lot of people can live healthier qualitied life without wasting our time for not efficient sleep,

Problem Statement

To know what brings us good sleep, I'd like to find out the correlation between good sleep and some features by making valid models which predict respondents' quality of sleep by using the data of "2013SleepinAmericaPollExerciseandSleepRawDataExcel" from National Sleep Foundation (*5).

Since the target variables are discrete and have multiple classes, this is a multiclass

classification problem. For example, “q12” which is the one of the target variables and indicates a respondent’s the quality of sleep by multiple classes, and this value will be predicted by several input variables such as the ones indicates the length of sleep time, whether a respondent drinks alcoholic or caffeinated beverages, the frequency of doing exercise and so on.

One of the potential solutions to this problem is to apply AdaBoosted DecisionTree (*6), to build models that will predict the respondents’ the quality of sleep, because it can handles multi classification problems and generally need less computation.

Datasets and inputs

The dataset I will use is “2013SleepinAmericaPollExerciseandSleepRawDataExcel” (*5 from a study of National Sleep Foundation (*7) .

According to “Summary of findings”(*8) of this study, this dataset is made by National Sleep Foundation and WB&A Market by conducting a national survey of Americans regarding their sleep habits. This poll is an annual review of habits, behaviors and attitudes pertaining to sleep and sleep quality. The study includes measures of sleepiness, drowsy driving, sleep disorders and general health.

The num of instances is exactly 1000, and the most of the dataset is made by answering the questions of “SIAQuestionnaire2013.pdf” (*9). For example, the column of **q1** in the dataset contains the answers of the question 1 in the questionnaire. If the answer to the question is “01”, “1” will be in the cell. However, some answers for the questionnaire are converted to continuous values such as **Q1VALUE**, **Q2VALUE** and so on.
Since it’s impossible to detail all the features in a permissible length, refer the “SIAQuestionnaire2013.pdf” (*9) to know the details of the features. However, I will introduce some of important features which I guess strongly correlated with target variables which indicate quality of sleep.

- **q29c**: Quantity of drinking caffeinated beverages between 5:00 PM and 5:00 AM the next morning in 12 ounce servings. (continuous)
- **Q36**: Amount of time a respondent spent for vigorous physical activities in the past 7

days in minutes. (continuous)
- **q20**: If a respondent snores loudly or not. (classification)

Since there are a lot of features and it will likely require a lot of computation, I will reduce the number of inputs to at least 100. I will remove following unnecessary features from the dataset.

-features which obviously not correlated with target variables.(Ex. **caseid**, **source**, **market**)
- duplicate features.
- features which are not in use.
- features which have too many empty values.
- features which I cannot understand what they indicate.
- features which are relatively weakly correlated with the target variables. I will measure the strength of correlations by using sklearn's Adaboost and its feature_importances_ attribute.

Since our aim is to find out some features which are correlated with the quality of sleep, the target variables will be following:

- **q11**: The number of worknights or weeknights a respondent feel like "I had a good night's sleep".
- **q12**: The number of nights you do not work or on weekend nights a respondent feel like "I had a good night's sleep".
- **q30**: The overall sleep quality during the past two weeks.

Solution Statement

To find out some features which are correlated with the quality of sleep, I'd like to make valid models which predict the target variables of **q11, q12, q30,** and find the features correlated with the quality of sleep by using feature_importance_ attribute. Then, I will analyze the result, and make a hypothesis of what brings us good sleep.

A benchmark model

Since it is a classification problem and there will be many features, I'd like to try to use Gaussian Naive Bayes which needs less computation and works well with many features.

Evaluation metrics

Since it is a classification problem and this problem doesn't attach importance to precision nor recall, I'd like to use accuracy as a metric.

Project Design

1, Preprocess the data.
- 1.1, Drop duplicate or unknown or empty columns.
- 1.2, Deal with empty values
-- 1.2.1, For a empty value that the feature requires us to input a continuous value, insert the median value of the feature.(Imputation)
-- 1.2.2, For a empty value that the feature requires us to input a discrete value, Insert 94, because this classification value is never used in the questionnaire
-1.3, Apply one-hot-encodings to discrete features
-1.4, Make premodels to apply feature selection, then, drop the features whose strength of correlation with each target variable is not in top 100.

2, Make models
-- 2.1, Separate training sets and test sets. Since the number of target variables is currently 3, I'm planning to make three models. Since I'm planning to use AdaBoosted Decision Trees, perform the grid search to find out the best parameters.

3, Test the models, and evaluate them by accuracy.

4, Analyze the result, and make a hypothesis of what brings us good sleep.

Citations

*1, “Missing just a couple of hours of sleep doubles car crash risk” from CBS NEWS
https://www.cbsnews.com/news/sleep-deprivation-doubles-car-crash-risk-aaa/

*2, “Sleep and Disease Risk” from Healthy Sleep
http://healthysleep.med.harvard.edu/healthy/matters/consequences/sleep-and-disease-risk

*3, “Sleep, Learning, and Memory” from Healthy Sleep
http://healthysleep.med.harvard.edu/healthy/matters/benefits-of-sleep/learning-memory

*4 “Overview of Sleep Disorders” from healthcommunities.com
http://www.healthcommunities.com/sleep-disorders/overview-of-sleep-disorders.shtml

*5, “2013SleepinAmericaPollExerciseandSleepRawDataExcel” from National Sleep Foundation
http://www.sleephealthjournal.org/pb/assets/raw/Health%20Advance/journals/sleh/2013SleepinAmericaPollExerciseandSleepRawDataExcel.xls

*6, “Multi-class AdaBoosted Decision Trees” from sklearn
http://scikit-learn.org/stable/auto_examples/ensemble/plot_adaboost_multiclass.html

*7, ” 2013 EXERCISE AND SLEEP” from National Sleep Foundation
https://sleepfoundation.org/sleep-polls-data/sleep-in-america-poll/2013-exercise-and-sleep

*8, ” Summary of Findings” of “2013 Sleep in America® Poll” from National Sleep Foundation
https://sleepfoundation.org/sites/default/files/RPT336%20Summary%20of%20Findings%2002%2020%202013.pdf

*9, “NATIONAL SLEEP FOUNDATION
2013 SLEEP IN AMERICA POLL: PHYSICAL ACTIVITY AND SLEEP SCREENING QUESTIONNAIR” from National Sleep Foundation
https://sleepfoundation.org/sites/default/files/SIAQuestionnaire2013.pdf