Problem 1: TriangleWell

A Machine Learning Approach to Predicting Sleep

Soren Davis
Annie Do
Nathalia Compres Perez
Summary

Given data on how many steps people take in each 15-minute interval of the day, we aimed to accurately predict the sleep patterns of students in the dataset. However, the data includes a great deal of error, as the step counters erroneously assume movement equals steps. This further complicates the model because people will in some cases have over 90 steps even while they recorded that they were asleep.

Our model seeks to account for this problem by using a machine learning algorithm to find the patterns and predict when a student is asleep given their step count. We implemented this using logistic regression and a machine learning model. This will allow the algorithm to predict whether the student is asleep or not.

In order to train the ML model, we used 70% of the dataset that recorded the students and their sleep patterns. To test the accuracy of our model, we commanded the machine to predict the remaining 30%. Ultimately, it fulfilled its role with 66% accuracy.

As a result, we have a model that can be applied to the remaining 400 students in order to predict at what times they are asleep. While there are certainly a lot of improvements to be made, this is an excellent starting place for the programming of the TriangleWell app.
Introduction

Sleep is integral to people’s wellbeing. Even though there is still much to learn about the process, sleep deprivation has been shown to lead to cognitive impairments and other health complications.\(^1\) As a result, TriangleWell aims to create an app to give students information about the connection between their sleep habits and activity levels. Given information about the steps students take every 15 minutes, our model aims to accurately detect sleep patterns in students. To do this, we fit a machine learning model using the scikit-learn Python package. Our goal is to develop a model that can, given a dataset of students and their step counts at different time intervals, accurately guess when they are sleeping.

Background Information

In order to accurately detect sleep in students, we first needed to familiarize ourselves with the sleep cycle. It typically takes a person 10-20 minutes to enter light sleep after settling into bed.\(^2\) Sleep varies between NREM (Non-Rapid Eye Movement) and REM (Rapid Eye Movement) sleep. It is then subdivided into 4 or 5 separate sections.\(^3\) The first stage falls under light sleep, it is the process of transitioning to heavier sleep. On average, this stage lasts 5-10 minutes and features sudden muscle contractions.\(^4\) The second stage’s primary feature is the introduction of sleep spindles, or bursts of activity in your brain waves in otherwise slow-wave sleep.\(^5\) This period lasts approximately 20 minutes. Some researchers lump stages 3 and 4 together into what is known as deep sleep. During this time, your brain primarily produces delta

\(^{1}\) https://my.clevelandclinic.org/health/articles/12148-sleep-basics
\(^{2}\) https://nmcab.org/how-long-does-it-take-to-fall-asleep/
\(^{3}\) https://www.sleepassociation.org/about-sleep/stages-of-sleep/
\(^{4}\) https://www.ncbi.nlm.nih.gov/books/NBK526132/
\(^{5}\) https://www.sleepfoundation.org/how-sleep-works/sleep-spindles
waves, making it more difficult to wake up during this time. Deep sleep usually lasts for 20 to 40 minutes.\textsuperscript{6}

While it may be intuitive to think that we would transition directly to Stage 5 (or REM sleep after stage 4, the body actually returns to stage 3 and 2 afterwards.\textsuperscript{7} This is because the brainwaves associated with REM are closest to waking state. REM sleep is entered approximately 90 minutes after falling asleep, and the first REM cycle lasts about 10 minutes.\textsuperscript{8} However, it gets longer the longer you sleep. During REM, there is no skeletal movement, a phenomenon described as muscular atonia.\textsuperscript{9} After REM, the body returns to stage 2 and the cycle begins again.\textsuperscript{10}

**Modeling Approach**

When we first began approaching this problem, we used the information we gathered in our research to make a set of assumptions. We first assumed that the movement recorded when people are sleeping would be much lower than when they’re awake. Therefore, the step count should be lower in spite of system error. By looking at the dataset that gave the step count and sleep count of the 100 students, we set a threshold of error of about 100 steps. This is a result of looking at a couple of students and finding that during the time that they were asleep, their steps consistently stayed under 100. We also assumed that there would probably also be fluctuations throughout the night as students progressed through the stages of sleep.

With this in mind, we began generally plotting charts of the step counts and the time that they were asleep to see if these were fair assumptions to make. One such graph is shown below.

\begin{itemize}
\item \textsuperscript{6} \url{http://www.weknowmattresses.com/2011/05/sleep-stages/}
\item \textsuperscript{7} Ibid.
\item \textsuperscript{8} \url{https://www.ncbi.nlm.nih.gov/books/NBK526132/}
\item \textsuperscript{9} Ibid.
\item \textsuperscript{10} \url{http://www.weknowmattresses.com/2011/05/sleep-stages/}
\end{itemize}
We also wanted to see the distribution of steps based on whether the students were asleep or not, this is depicted on the following graph. Here we see that most of the asleep data is concentrated in one area.

After making a couple of these, we realized that it would be near impossible to go through each one in the time allotted, much less manually make our guesses about the set of 400 students. Therefore, we used another resource at our disposal: machine learning.
To analyze the data, we used a supervised machine learning approach. We started with the 100 student dataset, and cleaned up our data by turning each row into a column to ensure the machine could read it. Then, we split the dataset into two, with the first 70% of the data used to train the algorithm, while the remaining 30% was used to verify its accuracy. We applied logistic regression using step count data to fit the model. We implemented the algorithm with the scikit-learn package. Ultimately, we found that our model was 66% accurate at predicting the remaining 30% of the data, while the logistic regression model was 59% accurate.

As a baseline, we decided to compare it to one of our assumptions alone. We told the program to make the assumption that if the step count was under 100, they were probably asleep. When we ran this test, we found that the accuracy was approximately 43%. Just to ensure that we did not just make an incorrect guess on the threshold of error, we played around with the numbers. We raised the step count value to be 110, 120, and then lowered it down to 90 and 80, but the percent accuracy would never be higher than 43% using this method. Therefore, we sided with the machine learning model that is 23 percentage points more accurate than using this assumption alone.

Model Analysis

A strength of our machine learning model is that it is very general. It has the ability to make decisions on complex combinations of factors that would be difficult to program manually. It is also quite fast to implement once the data has been prepared appropriately.

One significant weakness of our approach was that, the way we had set it up, the ML algorithm did not know whether any two rows of data were from the same person or different people. So it would not be able to make decisions based on learning the sleep schedule of an
individual. Ideally, we would be able to personalize it for each student to account for the different levels of movement that are individual to a person.

Another weakness of our model is that it does not increase the probability that the person is asleep late at night. Since most people are relatively consistent in their sleep patterns and tend to sleep at night, then an ideal model will account for this higher probability during these hours.

If we had more time, we would also aim to compare our model with an addition of one more assumption. Since we know that people take 15-20 minutes to fall asleep, we would want to make a model that accounts for both the step count and the intervals of time that have passed before the next time the step count is over 100. Under this model, we would make the algorithm say that they are asleep when the current step count is under 100, and the past 2 intervals have also been under 100. Unfortunately, we were not able to complete this analysis in the allotted time and have no way to compare this method to the existing model.

Results

We were able to use the step count data to predict the student’s sleep patterns with over 66% accuracy. The steps involved in this process are to convert the data into a usable format, by turning each of the rows into columns and give them an appropriate name within the program. Then, we use these new labels to run the logistic regression. After that, we used the scikit machine learning methods to create an ML model to predict sleep patterns. The code for these processes is found here.

This code can be expanded to predict the sleep patterns of the 400 student dataset given the presence of a properly formatted table.
Dear President Price and Wellness Center Staff,

We are honored to have the opportunity to work on the TriangleWell app. We believe that it will be integral to providing students with the information necessary to make informed decisions regarding their health, activity levels, and sleep patterns.

Our goal was to design this model being cognizant of the differences in student’s accessibility to technology. We understand that not every student has access to the most advanced health monitoring devices, therefore we sought to be able to predict sleep patterns through step count alone. However, with this comes different levels of complications. For one, not every device is as advanced in accurately quantifying step count, therefore the data that we were working with has steps recorded when none were taken. This is likely due to a student’s movement being counted as a step, even when they were asleep. As a result, we needed a model that could account for this error.

Enter machine learning. In our model, we used the data that was available to us to train the computer to be able to detect when students were asleep or not. We split the data of the students we had sleep information for into 2 parts. We used 70% of the data to ‘train’ the machine, and we made sure it was accurate by seeing how well it predicted the remaining 30%. We found that it was over 66% accurate!

This is of course only a preliminary model for the app, but it is of course a great start. We can apply this code to the step count information of students using the app and this will give us whether they were asleep or not. With this code, we aim to lower the effects of the technological disparity by allowing all students to learn more about their bodies and wellness, regardless of the quality of technology they have access to.

We look forward to continuing on this endeavor with your support.

Sincerely,

The ⅔ Triangle Coalition