Exploratory analysis of health data

G. Jönemo
E17, Lund Institute of Technology, Sweden
gu1673jo-s@student.lu.se

Abstract

This paper explores the process and possibilities of extracting health data from Apple watches given to organ transplant patients. Using the extracted data, the paper further explores how it can be used, in part to illustrate patient behavior but also to make predictions about patient health.

1 Introduction

When a patient has undergone intense organ transplant surgery, post operation health is of great importance. Nowadays extensive and expensive regular checkups are required during recovery to continually monitor for signs of illness and organ rejection. To assist in this health monitoring a medical study has been carried out over the course of a typical recovery period where health data has been collected through Apple watches. The project sets out to explore two things. First and foremost, it sets out to process the large amount of data being exported from the watches and present them in a more intuitive manner, so that medical professionals can assess how the patients' health changes. Second, the project sets out to explore the possibility of predicting how the health of the patients will change, ahead of time, given already collected data. If this were to be achieved, it would result in greatly reduced strain on the health care system and vast monetary savings. In this paper there will be an overview of the methods used to achieve the aforementioned goals in section 2. This includes a short paragraph about file formats, a dive into how the data was collected and, in conjunction, how the data was processed. This is followed by an explanation of the predictive model and the results. In Section 3 the conclusions of the project are presented, and section 4 contains acknowledgements of parties involved in the project.

2 Method

For the sake of simplicity and ease of use, Jupyter notebook was utilized in this project. This together with Python 3 as it has advantages when it comes to fast development of intelligent systems as well as extensive libraries for numerical operations, data handling and plotting. Notable libraries include NumPy, Pandas & TensorFlow. NumPy was mainly used for matrix manipulation, Pandas in order to manage the data using data frames and TensorFlow for the creation of the predictive model. In order to import the data of the XML-files that was extracted from the Apple watches the xmltodict module was used.

2.1 The XML format

Data exported from Apple watches are formatted in the file format "XML" which stands for Extensible Markup Language. The data is tree structured and follows a similar language composition as HTML, but other than that it's up to the author of the file to structure the data. Apples convention of creating these files come from their "HealthKit" API, but this API has little to no documentation which makes understanding the data more difficult. The files do include standardized encoding information which allows xmltodict to import the data in full into a data frame, but a lot of redundancy is introduced with this approach and the full data set itself is not suitable to work with.

2.2 The data

The data was collected from patients over a recovery period of a little less than two years. During this time, the quality of the data has largely been affected by the watch use of the patients. This means that days without wear, periods of forgotten heart monitoring and unusual movement patterns all contribute to errors that result in an incorrect representation of the patient's health. For example, when analyzing heart rate, large variations can be found. This is most likely due to the occasion during which the measurements were taken, such as if they were taken after a workout or during rest. Furthermore, due to patient confidentiality, information regarding recovery was not available at the time of the project which made predictions of absolute patient recovery infeasible.

2.3 Processing

In order to compensate for the errors in the data some manipulation was required. This was done in part by imputing through interpolation, which essentially fills in gaps between datapoints my assuming a linear relation. In addition the sometimes empty ends of the datasets had to be trimmed since interpolation only works between established datapoints. To get clean data sets to interpolate the original data also had to be properly datetime-formatted, filtered for relevant fields and saved in the CSV (Comma Separated Values) file format for faster load times.

3 The predictive model

The predictive model was designed as a proof of concept. It to take two inputs, in this case step count and flights climbed, and outputs a prediction of what the corresponding heart rate measurement during the same time frame should be. Using a sliding window, the data is fed through different layers using TensorFlow. First, a batch normalization layer which further helps mitigate data point errors. It then uses an LSTM layer, which stands for Long Short Term Memory, in conjunction with a dense layer in order to achieve the desired single output.

LSTM, which is the main technology behind the predictive model, is a type of RNN (Recurrent Neural Network). RNNs allows information to pass between steps of a network, thus giving the network the ability to “remember”. The problem arises with long sequences where the distance between the prediction and prediction-critical data is too large, as the memorizing effect of RNNs favor recent information. LSTM solves this problem by introducing a built-in long-term memory, capable of remembering information for long periods of time. LSTM accomplishes this nu using four interacting layers within each repeating module, as opposed toa single one in a typical RNN. These layers makes it so information can be added or subtracted at each step, allowing for different types of information to pass through and thus certain data properties to be remembered longer than others.[1]

Figure 1: Example of basic RNN architecture

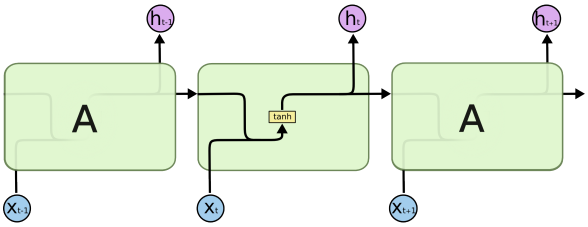
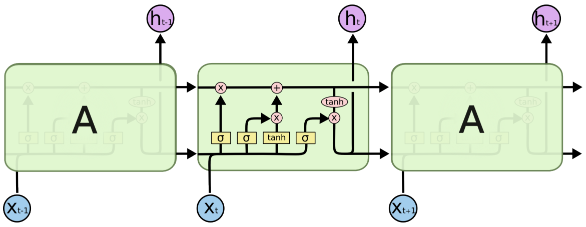


Figure 2: Example of LSTM architecture



4 Results

A graphical representation of the data was achieved and contributed in large to the understanding of the patient’s movement patterns. Figure 3 depicts a cronological representation of flights climbed from a particular patient, displayed on a daily basis with a monthly mean in order to better understand the trend of the otherwise erratic datapoints. Flights climbed is apples name for vertical height traveled, referring to flights of stairs climbed. In figure 4 we see step count by month for the same patient. Certain trends are apparent in this particular data set such as decreased activity in December, as shown by the two different metrics.

Figure 3: Flights climbed

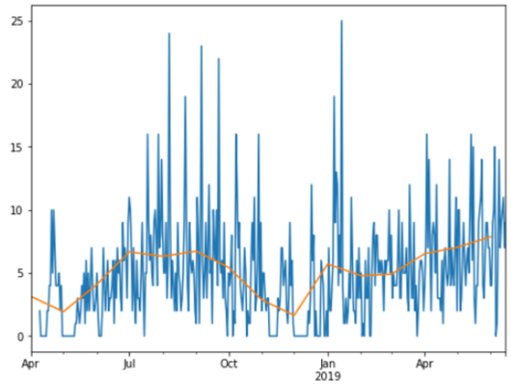
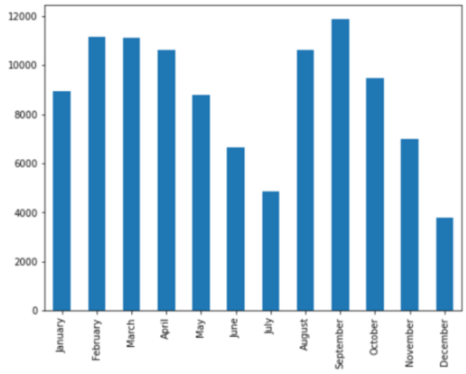


Figure 4: Step counts by month



As for the predictive model, it was a good start but needs some work. While the model was learning correctly and making predictions in the same order of magnitude as the training data the mean square error was still relatively large, and stagnated at around 360 after 35 epochs. There are many underlying sources of error as to why the LSTM, a model very well suited for time series like this, underperformed. This is further discussed in the next section.

5 Sources of error

There were many sources of error in this project. The main ones were inconsistent and insufficient data. Had there been clear information regarding how and in what circumstance the measurements were taken this would have been greatly useful in understanding the data. For example, as previously mentioned, it's of great importance to understand weather or not a heart rate measurement was taken before, during or after a workout in order to process it correctly. The study was also relatively limited in size with only a handful or two of patient records available. For accurate predictions more data would be beneficial.

6 Conclusions

I believe predictions using the algorithms outlined in my project can be useful in understanding patient health and possibly even make predictions ahead of time. For this to work however I think a broader study with more patients would be required, with stricter guidelines regarding data collection. A more fine-tuned predictive model can most likely be created that would be better adapted for the application. All in all, I think the project was a success and given more time I believe more discoveries can be made within the area.

Acknowledgements

As for my acknowledgements I would like to thank my supervisor, Pierre Nugues, for keeping the project on track by organizing our weekly meetings and making sure realistic goals were set. I would like to thank Marcus Klang for all the technical assistance when my code wasn't working, Johan Nilsson for his invaluable patient data and insight, and Dennis Medved for his expertise and support during the project. Last but not least I would like to thank Jacek Malec for organizing the course and getting me in touch with my supervisor and associated team.

References

1. Christopher Olah. Understanding LSTM Networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/