Predicting Sleep Using Consumer Wearable Sensing Devices

Miguel A. Garcia
Department of Computer Science
Stanford University
Palo Alto, California
miguel16@stanford.edu

1 Introduction

In contrast to the explosion of consumer wearable health device market, which includes millions of units sold worldwide, the efforts undertaken to validate the accuracy and reliability of these monitoring apps and devices have been minimal [1]. To reach this understanding there are several barriers. In the domain of sleep tracking, sleep researchers have yet to define standard metrics for validation, and wearable companies contribute to a general lack of availability of technical information by keeping information about sensors' raw data, accuracy, and algorithms hidden as trade secrets [2].

It is increasingly becoming more common for patients to present data acquired from consumer wearables to clinicians (who due to obstacles like those presented before, struggle to interpret this data). Finding reliable devices is a challenge since the acquired data can be at best unregulated and at worst dubious. A widely accepted way to track sleep duration is found in clinical actigraphy, which is a non-invasive method of measuring of motor activity, usually by a small device worn on the wrist. Marino, et al. (2013) validated actigraphy for detecting sleep and wakefulness by using it to yield accurate results resembling those of polysomnography (PSG), which provides the gold standard for this task. However clinical actigraphy devices are expensive ($1000 per unit).

To this end, we study the possibility of employing accelerometer-based consumer wearable devices, which are as widespread as they are affordable (as low as $15).

2 Dataset and Features

The dataset for this project was obtained from the Emergent Innovative Global Health Technologies (EIGHT) lab at Stanford. For about 20 different test subjects in the span of 6 months, the Basis Peak watch was used to record 7 different body measurements by the second, including time, heart rate, and galvanic skin response. Since the feature set was already small, when PCA was ran it was not used to project the feature space into a smaller subspace, but rather to analyze what features were most important to categorizing between sleep stages.

There were 2 separate datasets used as inputs to the models, both sets of recordings from the same test subject. The first input (dataset A) trained on observations made for the entire month of January 2016, and then tested on observations from the first 2 weeks of February. The second input (dataset B) trained on observations made from January 2016 until March 31, 2016, and then tested on the month of April. Each training example was either unlabeled or labeled with the sleep stage that the test subject was in if they were asleep (deep, rem, or light sleep). This label was also calculated by the watch and can be assumed to be true.

2.1 Pre-processing

We re-labeled unlabeled training examples as "awake", scaled individual samples to have unit norm, and filtered out training examples with missing values in any of the features. An extra output vector

was made for the binarized task of predicting wakefulness vs general sleep, as all training examples during times of wakefulness were assigned to the value "0", and those during times of sleep in any stage were assigned the value "1".

Dataset (A) contained 44609 total training examples before being reduced by preprocessing to a final total of 7972 training examples. The test set size for A was 18479 after preprocessing. Dataset (B) contained 130889 total training examples before being reduced by preprocessing to a final total of 21112 training examples. The test set size for B was 38925 after preprocessing.

After experiencing issues in classification, it was discovered that an imbalance in the dataset in the form of over-representation of wakefulness training examples and under-representation for rem/deep sleep training examples was resulting in flawed results. Thus, an extra step was added to preprocessing to balance the dataset by ensuring there were an equal number of examples for each label. This balancing was not done for the test data.

An initial attempt was made at conducting PCA to see whether the feature space could be reduced further, but the explained variance for any feature was not insignificant enough to consider removing it. The following PCA plot was made for our single test subject on dataset A. It is important to note how sleep stages at this point are still inseparable.

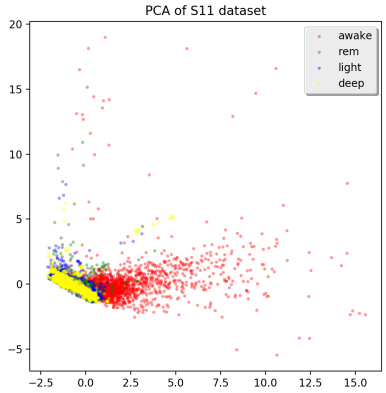

Figure 1: 2D projection of dataset A (training examples) using PCA. The first principal component is heart rate with an explained variance of 0.44, and the second principal component is galvanic skin response with an explained variance of 0.16.

3 Models

The models are split into 2 different categories for the 2 different tasks in this project: finding a binarized classifier to distinguish between wakefulness and sleep, and finding a multi-class classifier to classify examples based on the different stages of sleep.

3.1 Binary classifiers

The labels for this task were "1" for asleep and "0" for awake.

3.1.1 Logistic Regression

Logistic regression works by maximizing the log-likelihood of the training examples. We learn the parameters θ_i of the binary model by maximizing

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} \log(h(x^{(i)})) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \tag{1}$$

3.1.2 Support Vector Machines

Support vector machines (SVMs) try to separate data points with a vector while maximizing the margin between training examples of different classes. Given training vectors $x_i \in \mathbb{R}^p, i = 1 \dots n$ and a vector $y \in \{1, -1\}^n$, SVM's can solve the binary classification problem

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j K(x^{(i)}, x^{(j)}) \tag{2}$$

s.t. $0 \leq \alpha_i \leq C, i = 1, \dots, m, \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$

The following kernels were used for K: $linear : \langle x, x' \rangle$, $polynomial : (\gamma \langle x, x' \rangle + r)^d$

3.2 Multi-class classifiers

The labels for this task were "awake", "light", "rem", "deep".

3.2.1 K-Nearest Neighbors

Assign class j to $x^{(i)}$ that maximizes the following:

$$P(y^{(i)} = j | x^{(i)}) = \frac{1}{k} \sum_{i \in N} 1\{y^{(i)} = j\} \tag{3}$$

3.2.2 Softmax Regression

Softmax regression is an extension of logistic regression to the multi-class problem. Instead of calculating probabilities for 2 classes, we calculate them for multiple classes. The class predicted is the class with the probability that maximizes the likelihood. We learn the parameters θ_i of this model by maximizing

$$l(\theta) = \sum_{i=1}^{m} \log \prod_{l=1}^{k} \left(\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right)^{1\{y^{(i)} = l\}} \tag{4}$$

3.2.3 Support Vector Machines

SVMs can also be used to build a multi-class predictor for k classes by using the "one vs. rest" approach (OVR). In OVR, the algorithm builds an SVM for each class, and the class assigned to each data point is that which maximizes its distance from its respective margin.

4 Experiments

The first series of tables below demonstrate the first experiment ran without balancing the dataset. Since this experiment was done without balancing, the training set is larger. While the best binary classifier performs reasonably well, the multi-class classification has artificially high accuracy. Because of the unbalanced dataset, the less frequent stages of sleep "light" and "rem" sleep are almost entirely misclassified. A better measure of performance would be the F1-score, so that we can evaluate the predictor on precision and recall over all classes.

The second experiment balanced out the datasets. For the binary case, this meant including as many examples as the sleep state. For the multi-class state, this meant including as many examples as the "deep" sleep state. After looking at the training data I determined these were the least represented categories. There were improvements in the binary case, but in the multi-class task error approached 60% for nearly all the models. Here are the results below for the binary case (with k-nearest neighbors performed this time around):

kNN train err: 3.581620%
kNN dev err: 17.570322%
kNN test err: 24.882299%

Classifier	Train	Dev	Test
Logistic Reg. (L2 penalty)	6.99%	7.32%	8.86%
SVM (linear)	6.83%	7.19%	9.15%
SVM (d = 2)	9.17%	9.68%	11.5%
SVM (d = 3)	6.61%	7.09%	8.09%

Table 1: Binary classifier error (hold-out cross validation)

Classifier	Train	Dev	Test
K-Nearest Neighbors	9.03%	13.43%	20.66%
Softmax Reg.	17.29%	17.25%	15.03%
Linear SVM	17.31%	16.88%	15.42%
SVM (d=2)	19.18%	19.03%	17.28%
SVM (d=3)	17.13%	16.91%	14.94%

Table 2: Multi-class classifier error (hold-out cross validation)

	'awake'	**'asleep'**
'awake'	12826	1029
'asleep'	466	4158

Table 3: Actual vs predicted for binary SVM(d=3)

	'awake'	**'rem'**	**'light'**	**'deep'**
'awake'	13032	3	820	0
'rem'	189	1	680	2
'light'	468	8	2685	4
'deep'	60	0	527	0

Table 4: Actual vs predicted for multi-class SVM(d=3)

Test set f1 score: 0.751177

LR train err: 6.64492458214%
LR dev err: 20.5462698736%
LR test err: 29.4063531576%
Test set f1 score: 0.627604

SVM (linear, 0) train err: 6.493506%
SVM (linear, 0) dev err: 19.431988%
SVM (linear, 0) test err: 28.118405%
Test set f1 score: 0.637556

SVM (poly, 2) train err: 11.991148%
SVM (poly, 2) dev err: 29.936133%
SVM (poly, 2) test err: 35.602576%
Test set f1 score: 0.580555

SVM (poly, 3) train err: 6.609982%
SVM (poly, 3) dev err: 22.638946%
SVM (poly, 3) test err: 31.543915%
Test set f1 score: 0.611581

5 Discussion

Although the results are acceptable for the binary case, there is much room for improvement in the multi-class state. Since the models do not generalize well, using more data might improve them. However, much of the models are also hampered by the limited feature set. Perhaps adding more features by collecting more information about the user with the Basis Peak watch might help, though data collected by another device, such as a smartphone, could be combined someway with the data we have collected already as well.

6 Future Work

With more time, I wold have liked to improve upon these models. Much of my time was spent understanding the models and learning how to evaluate the results of the experiments. Once the models are improved, I would like to use a mixture of Guassians model to combine data from multiple people, and see whether these distributions fit the Gaussian model.

Acknowledgments

I would like to thank the EIGHT lab for their help in providing me with their datasets and Dr. Katarzyna Wac for her guidance.

References

[1] de Zambotti M, Godino JG, Baker FC, Cheung J, Patrick K, Colrain IM. *The boom in wearable technology: Cause for alarm or just what is needed to better understand sleep? Sleep. 2016;In Press:1761-1762. doi:10.5665/sleep.6108.*

[2] Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, Dulin H, Berkman LF, Buxton OM. Measuring Sleep: Accuracy, Sensitivity, and Specificity of Wrist Actigraphy Compared to Polysomnography. Sleep. 2013;36(11):1747-1755. doi:10.5665/sleep.3142.

[4] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. Journal of Neuroscience **15***(7):5249-5262.*