**The Lone Ranger - Project A**

**Samantha L. Misurda**

**Week 2 (4/2)**

As it is only week 2, my first week was spent mostly getting acclimated to the class and project expectations. My personal interests lie heavily in the areas of fitness and nutrition, so the human activity recognition immediately piqued my interest.

From reading the project overview, it appears that my task is first to use the dataset to determine if we can use accelerometer data to see if a user is participating in some sort of physical activity. Following that, I will select features, focusing on the timed nature of the events cataloged in the dataset. Finally, I hope to apply some of the techniques that we will learn in the class, in order to build a model capable of detecting said activity.

My experience with fitness trackers suggests that they are a little finicky, and that noisy, or otherwise anomalous readings may affect my results. Task number 4, which is determining the optimal location for the sensor may also be a little difficult, as this likely varies from person to person.

Over the next week, I plan to begin exploring the dataset, addressing the areas posed in the guidelines. For reference, these are recorded verbatim below:

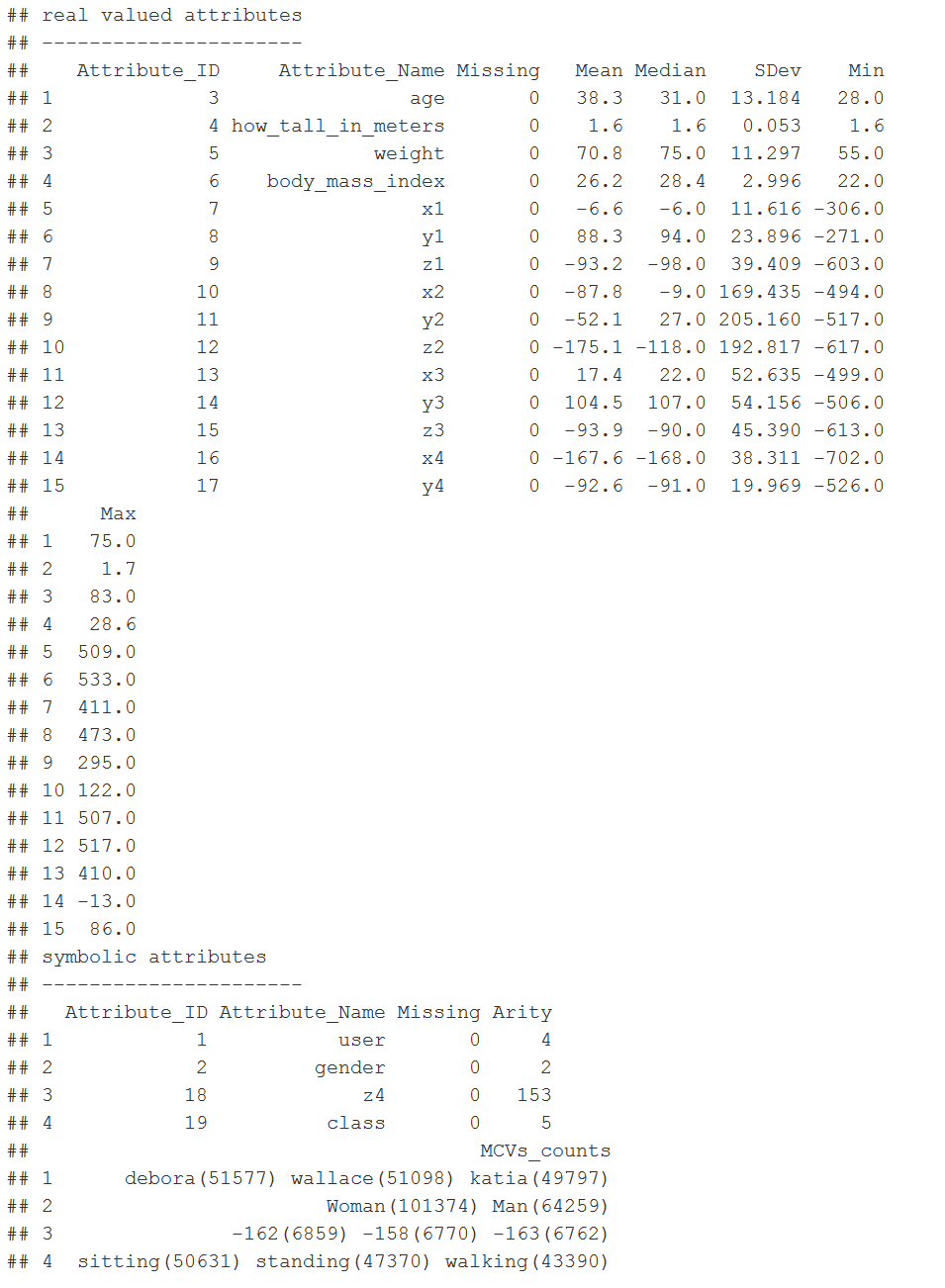
* do you have a grasp of the problem? what's the motivation behind it? get yourself familiar with the background by doing a little bit of research on the internet.
* data profiling: the shape of the data set (long skinny, fat wide);  plot distribution (histogram) of each variable; missing values? anything strangeness of the data? what are you curious about this data?
* what is the input and what is the output in your model?
* without using any modeling, how do you assess the predictive utility of a single variable as related to the output? any statistical test can be used, for example, t test, chi-squared test, correlation etc.

At this point, the only obstacle standing in my way is time. I’m relatively new to data mining, and will need some time to progress through the class in order to learn what techniques may be suitable for use.

**Week 3 (4/9)**

1. Create a profile of the dataset, you may use the brief function you wrote in homework 1, or use summary function, create a few plots to illustrate what you learned from your dataset, just like you did in Homework 1 problem 1b.

Below is the output of running the brief function on the Project A dataset:



The acceleration data isn’t going to be very useful for just a human reading it, so I am going to ignore it in my initial basic analysis. It will be extremely essential in the future. Additionally, with only four users, visualizations aren’t going to show much. This is just a very shallow exploration, to hopefully learn a bit about our subjects. I used R to extract the following (age, BMI) tuples for each user:

(debora, 46, 28.6)

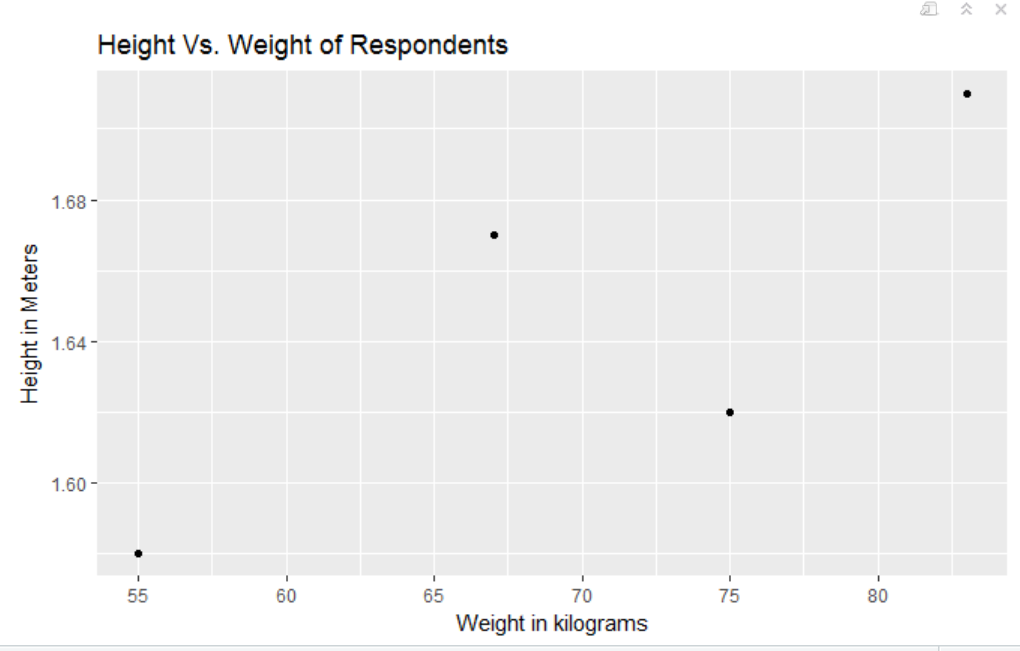
(jose\_carlos, 75, 24)

(wallace, 31, 28.4)

(katia, 28, 22)

I was interested to see that there was a wide range in ages. However, it was a bit discouraging to see the BMI ranges. With a minimum value of 22 and a maximum of 28, we are not necessarily focusing on athletic users, but exclusively on the overweight, or slightly above average. Although BMI is not the sole indicator of health, it is usually a good indication, aside from a few extreme cases, such as the incredibly tall (BMI kind of falls apart for basketball players).

A scatterplot of height vs. weight may help to illustrate any anomalies a bit better.



1. **If your problem is a classification problem, try to run a few classifiers and compare the results; if this is a regression problem, you may try a linear model.  Don’t forget to compare to default predictor.**

I do believe this is a classification problem, but I haven’t started looking into the acceleration data yet. With that being said, I’m not sure how useful a linear model would be for summary statistics on the four users, as I’m not really trying to predict anything. I am hoping to repeat the approach outlined in Homework 1 on my acceleration data during the following week, as part of addressing Task 3.

1. **What is your plan for next week?**

When considering the schedule for the remainder of the course, I would like to start by trying to address one of the tasks in the project description each week. For next week, this would mean Task 1: Determine whether or not the activity of an individual can be correctly detected from the given accelerometer data. I’m apprehensive to start dealing with the accelerometer data, as I’m imagining there is a bit of overhead associated with understanding motion.

1. **Any questions to client and mentor?**

Nothing at the moment, but I may have questions about how best to visualize movement, if I choose to include plots that show that. I’m thinking some sort of facet wrap for each user over time.

**Week 4 (4/16)**

1. **If your problem is amenable to classification, please continue to experiment with classifier if you haven’t done it. Try to produce ROC diagram for different model and compare. Ready to interpreter the result.  What kind of classifier metrics should you really care about given the context of the problem? Verify this with client;**

For task 1, I am attempting to figure out if we can detect movement from the accelerometer data. I plan to refine the models and better prepare the report in the last two weeks of the course, but I am going through the steps weekly in order to serve as an additional check of my understanding.

Initial work with classifiers has produced the following models and results. Note that my training and testing sets have been naively crafted, and I may want to rethink the split, and retrain models. The models were built to predict class from the x, y, and z coordinates. For motion, we are concerned with sitting down, sitting up, and walking. I have recoded the data for model building to be a binary “is moving” or “is not moving”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | TPR | FPR | Accuracy | Precision | Recall |
| Logistic Regression | .73 | .08 | .84 | .85 | .73 |
| SVM | .68 | .08 | .81 | .86 | .68 |
| Naïve Bayes | .53 | .02 | .79 | .92 | .53 |

SVM takes n^2 time, which is not reasonable given the size of our dataset. To be able to use it, I had to perform sampling, which may have skewed the model.

Naive Bayes has a decent precision score, but a laughable True Positive Ratio.

Logistic Regression was simple to implement, fast, and performed respectably.

Given the nature of the data, I am not overly concerned about the true and false positive rates. It is not medically critical that we are super accurate in our detections. Also, movement is an event that occurs over time. Our data may not immediately capture the relationship as “moving”, but may take a few time steps. With that being said, accuracy might be a good metric to be concerned with, as the number of targets in our data are fairly small.

**2. You may start to think about how you plan to go about with feature engineering, given your knowledge about PCA or feature selection techniques**

In terms of feature engineering, I think choosing to encode the different types of activity classes in a similar way may be useful. As I said above, we need to decide what we consider to be movement. We currently have five levels of activity to reflect change: walking, standingup, sittingdown, standing, and sitting. Three of those activities represent actively moving, whereas two are stationary.

**3. As always, plan for next week and questions to client and mentor.**

For next week, I plan to begin examining task 2.

As I said above, for motion, we are concerned with sitting down, sitting up, and walking. Is this correct, or do we only consider “walking” to be motion?

Are there any other areas I should consider with regard to feature engineering?

**Week 5 (No Report Due)**

**Week 6 (4/29)**

There weren’t any questions to answer this week, unless I have missed them. I will use this week to just give a brief update on progress, and leave the floor open to you, Yixin, to ask any questions as the client.

Over this week and the past week, I have finished my precursory work addressing the tasks posed in the project description. I say precursory, because I have performed very shallow explorations at this point, and plan to perform a full and thorough experiment while writing up the report over the next few weeks.  Just a reminder of what these tasks were:

* Explore the temporal nature of the data
  + Look into which features are most helpful
  + Compare them to only acceleration data for a time slice
* Determine if I can reliably detect activity (I can)
  + Investigate how long it takes to detect a change
  + Derive new features to improve detection
* Determine the best location for the accelerometer sensor
  + Waist, left thigh, right arm, or right ankle
  + Examine both accuracy and change point detection

The answers to these questions will be addressed in my final report, but initial results are encouraging.

Yixin, is there anything else you would like to see demonstrated that the tasks do not directly address, even as an extra “nice to have” feature? Can you advise me on any of the pitfalls you ran into while working on this project when you took the class?