Sensor Data Project

Alex Nguyen (ana58), Jordan Bengco (jbengco)

The problem you are addressing, particularly how you refined the provided idea.

Our group chose to analyze if we could collect walking data and differentiate between subjects in our test group. To refine the question, we collected data for 3 different locations on the body: ankle, hand, and pocket. As a secondary problem we also tried to see if we could differentiate which part of the body the data was collected. The focus of our study is to analyze which data collection techniques and classification methods can best answer the question.

The data that you used: how it was gathered, cleaned, etc.

How the data was gathered:

The data was gathered on a mobile phone application called 'Physics Toolbox Sensor Suite' that records the acceleration of the device. The data was recorded while the device was located on the tester's ankle, hand and pocket.

For the ankle test, the device was taped on to ensure that it did not have any movements that added unnecessary noise while data collecting. For the hand test, we simply ask the testers to hold the device in a constant position during the duration of the recording. The pocket tests had the device sitting inside the testers front pant pockets. The choice of left or right pockets were decided by the participant.

Each test lasted 30 seconds and was repeated three times for each location on the body. This gave us nine total tests per tester.

How data was cleaned:

The data was cleaned by first removing any extra information that was not needed for our analysis. The original data had five columns: time, x acceleration, y acceleration, z acceleration, and total acceleration. We removed the total acceleration column and the time column was made redundant after we cropped any data beyond 30 seconds. We cropped all data after 30 seconds to get rid of the noise that was created while we were trying to stop the recording at the end of each test. It also makes data analysis more clear and easier to compute as it uniforms the data. Lastly we calculated the means and standard deviations for each column.

Techniques you used to analyse the data.

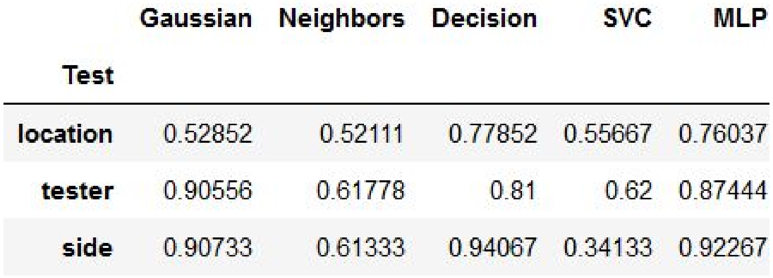
To analyze the noise heavy data, we took the means and standard deviations for the x, y, and z accelerations for each test. This helped us get a better general understanding of our data. We noticed that the same test for each participant will have relatively the same standard deviation. This lead us to do a Mann–Whitney U-test on the data. We also used different machine learning classification techniques to see if we could classify each participant, tester or side using the means and standard deviations from their data.

Your results/findings/conclusions.

We found that certain classifiers were very good at determining which tester and side the data was collected from. Gaussian, decision trees, and the multilayer perceptron classifiers all had scores above 80% determining which tester recorded the data and scores above 90% scoring which side of the body the data was recorded on. On figure 2 we see the x and y acceleration graphed and color coded per user. We can see that most of the red points lie near the middle and upper left side of the graph, green points clumped near the center, and blue points towards the right side of the graph. The concise green clump is particularly interesting considering it is the average summation of ankle, hand, and pocket data and as figure 1 shows, the data is wildly different per location. Figure three shows a plot of which side the data was collected on. Interestingly there is an even split at x = 0 with only one outlier jumping over the line. The split in x acceleration could be caused by a flaw in data collection. When doing data collection the phone was always facing towards the user so they could easily start and stop the data collection. This means that when collecting data for the left side, the phone is facing right. When collecting data for the right side, the phone is facing left. The divide in data points explains how the classifiers were able to score very highly on what first seemed like a difficult task.

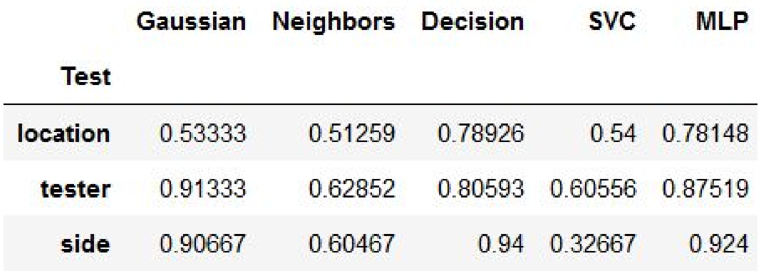
Chart 3 shows the results of running a Mann-Whitney U test on every permutation pair of data. The most significant differences overall was the difference in x acceleration with only pocket data not having a significant difference to its comparisons. Interestingly, tester two and tester three had significantly different {x,y,z} triples which scored very poor for every other pairing.

Some appropriate visualization of your data/results.



	Gaussian	Neighbors	Decision	SVC	MLP
Test					
location	0.52852	0.52111	0.77852	0.55667	0.76037
tester	0.90556	0.61778	0.81	0.62	0.87444
side	0.90733	0.61333	0.94067	0.34133	0.92267

Chart 1: Classifiers on raw data



Test	Gaussian	Neighbors	Decision	SVC	MLP
location	0.53333	0.51259	0.78926	0.54	0.78148
tester	0.91333	0.62852	0.80593	0.60556	0.87519
side	0.90667	0.60467	0.94	0.32667	0.924

Chart 2: Classifiers on LOESS cleaned data

	x, y, z	x	y	z
Ankle / Hand	0.5	**0.0303**	**0.00147**	**0.0202**
Ankle / Pocket	0.19137	0.23519	**0.00036**	0.13006
Hand / Pocket	0.19137	0.07856	**0.04423**	**0.00029**
P1 / P2	0.33126	**0.00669**	0.05589	0.12538
P1 / P3	0.33126	**0.00669**	**0.00029**	0.36197
P2 / P3	**0.04043**	**0.0002**	0.36189	0.12538
Left / Right	0.33126	**0.00021**	0.18861	0.36197

Chart 3: Comparing pairs of means with the Mann-Whitney U Test

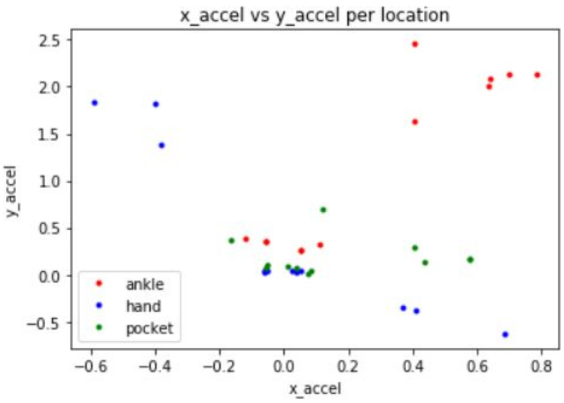


Figure 1: Acceleration per location

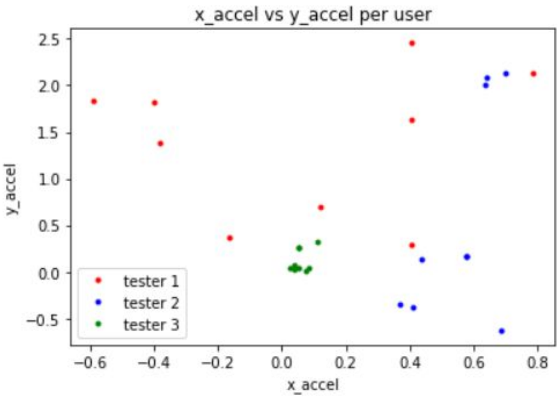


Figure 2: Acceleration per tester

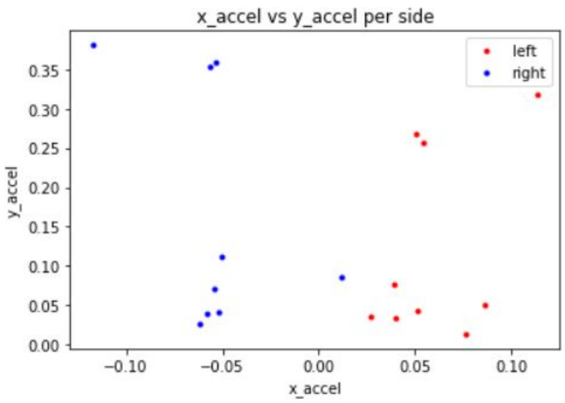


Figure 3: Acceleration per side

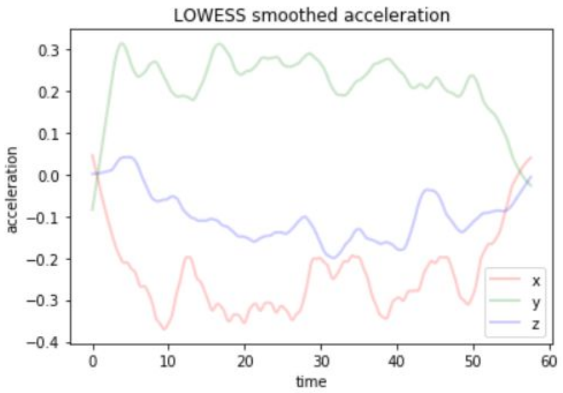


Figure 4: LOWESS smoothed acceleration

Limitations: problems you encountered, things you would do if you had more time, things you should have done in retrospect, etc.

Problems encountered:

One limitation we encountered was that we could only record the acceleration of a certain direction. For example we could not tell if the phone was rotated unless we also recorded using a gyroscope. Another limitation was that starting and stopping the recording required us to interact with the phone. Therefore some data collected at the beginning and end of the recordings were not related to the tester’s walking. Another problem we faced was debating whether to run some inferential stats tests on our data because we were unsure if the results from the normal and equal variance test were “good” enough.

Things to do if we had more time:

If we had more time we would have tried to implement Kalman filtering. Due to how much we needed to understand the error in our data collection, we settled for using Lowess filtering. If we had more time and resources we could have collected more data from more testers and done more tests per location per tester. We collected data from four users who each recorded three tests per location for a total of 36 data samples to analyze and test. Ideally we collect data from many more users and do

Things you should have done in retrospective:

One thing we should of done in retrospective filter and clean the data in a more organized way or in a single csv file. For our project we had a csv file for each person and each position where the device was recording. Instead we could have just added new columns to identify the tester and where the device was recording the data.

Project Summary - Alex Nguyen (ana58)

- Saved group member's time by creating a general project plan. Wrote a text file that shared general task and communicated the code that needed to be implemented in order to finish the project.
- Fixed an issue with a final report by proofreading and identifying missed assumptions within the data analysis technique used to create results for the report. Resulting in a more correct final report.
- Identified/implement a fix for possible errors from data collection. Identified any time testers were not performing actions during data collection and reference during cleaning data process. Resulting in cleaner data for analysis.

Project Summary - Jordan Bengco (jbengco)

- Collected data in a consistent format that maintained integrity of the research and enabled us to accurately perform statistical tests and analyze conclusions.
- Fit, tested and scored the data using various machine learning classifiers to determine the best classifier for each task.
- Tested every permutation pair of the data using a Mann-Whitney U test to verify which pairs had different means and analyzed the results.
- Co-wrote a report that summarized and explained our results using graphs and charts that displayed the data in a way that is easy to read and understand.