Classification and Repair of Low Integrity Heart Data

Abstract
A common feature of smart watch devices is the ability to measure heart rate. This is done via a process called photoplethysmography, where green light is shone through the skin, and absorbed by the red colored blood cells. The change in how much light is reflected back by the blood allows the sensors in the watch to measure its wearer's heart rate. This method is prone to errors. If the watch is moved slightly, it could cause the sensors to not receive the light that is reflected back. These errors are most commonly reflected in drastic spikes or dips in heart rate.

The goal of this research is to classify low integrity data points present in heart rate data. We also attempt to repair erroneous heart rate data in order. Heart rate data can be used to calculate various metrics, such as Total Daily Energy Expenditure (TDEE) and Basal Metabolic Rate (BMR). These metrics are useful for tracking fitness, so being able to measure them using data from a personal device is incredible valuable. Thus, these metrics are used in this research to show how repairing the data can have a great improvement on measurements.

Technologies Used
- Python – Main language used for programming
- Pandas – Database library used for processing raw data
- Numpy – Python library supporting math operations on big data
- Scikit-learn – Python library with implementations of common machine learning algorithms
- Matplotlib – Data visualization tool used to create graphs and diagrams of data

Methods
Many algorithms were used to try and detect erroneous data, to varying levels of success:
- Clustering: Create a k-dimensional scatter plot of heart rate data with each point representing heart rate from time t to t + k. Determine where clusters of data are, and fix data in outlying clusters. This failed because clustering was too inaccurate on smaller data sets.
- Stability: Separate heart rates into four different zones (see right). Determine how stable a zone is, and how likely heart rates are to transition to new zones. Find outliers via unlikely transitions. This failed because it was too general
- Census of Zones: Formalize the last technique by taking recording data around every zone transition. Determine outlier using zones with extreme data. This failed to find all transitions, as some happen outside of transitions.
- K-Nearest Neighbors: Create a n-dimensional scatter of heart rate data with each point representing heart rate from t – n/2 to t + n/2. Calculate the distance to the k nearest neighbors of each point. Take points above the 95th percentile of distance as outliers. This worked to find the majority of points.

To fix heart rate data, we simply remove outlying data, and replace it with a linear approximation between endpoints. This creates a decent approximation of heart rate change over short intervals.

Heart Rate Zones: Inductive or Deductive?
While working with heart rate zones, we had to choose whether to use an inductive or deductive approach to determining these zones. The inductive approach involve using machine learning algorithms to determine these zones. The deductive approach was to use heart rate zones as defined by the CDC (listed below). Both the inductive and deductive approaches yielded similar results, so we used the deductive measures, as they were easier to implement in code.

Heart Rate Zone | Heart Rates (bpm)
--- | ---
0 – Resting Heart Rate | < 92
1 – Moderate Heart Rate | 92 – 129
2 – Intense Heart Rate | 130 – 157
3 – Extreme Heart Rate | > 157

Conclusion
After trying many different methods, we were able to successfully classify errors in heart rate data. The methods used to repair the data also worked fairly well. This project was done as some preliminary research for Dr. Davis’s lab, and can easily be extended in the future. Possible extensions are to:
- Test repairs on aberrant data, such as data from patients with sleep apnea or atrial fibrillation
- Generate fake heart rate data and inject errors to more rigorously test repairs
- Measure how much more accurately data measures metrics such as Total Daily Energy Expenditure or Basal Metabolic Rate.