Technology Review

Ebola Virus Project

CS 461, Fall 2017, Group 34

Bianca Beauchamp

Abstract

There is currently no way to take a persons temperature without being within a close proximity to their bodies, putting health care workers at a great risk of infection. The purpose of this project is to reduce this risk by creating a device that will be able to quickly take a persons core body temperature from a distance using a thermal camera. In order to do this the thermal image must be processed and a model needs to be created. To process the image the pixel data must first be extracted and sorted, then the outliers will be removed. To create the model, summary statistics will be preformed to consolidate the data into a single value. That value will be used to model the relationship between skin temperature and core body temperature. Then the model will be tested by calculating the error of the model. This document will discuss some of the possible options and come to a conclusion best option for implementing the summary statistics, model, and model test. The best option for the summary statistics was determined to be the mean because it is the most accurate option and it is the simplest to find. The best option for the model was determined to be the linear regression because it is the simplest type of model and using it is a good starting place to see if it works of if a more complex model is needed. The best option for testing the model is absolute error because it provides a simple evaluation of if the model is accurate or not.

1 Summary Statistics

1.1 Overview

Once the pixels of skin have been isolated from the image, they need to be simplified into a single value that is a good representation of the entire set of data. This single value would represent the temperature of the subjects skin. There are three good options to consider for this purpose. The first is taking the median of the data, the second is to find the variance of the data and the third is to find the mean of the data.

1.2 Criteria

The option chosen has to be the most accurate representation of the data. This is important because this value will represent the skin temperature and will be used in relation to core body temperature. The model that creates the relationship is the most important part of this project. Therefore, the value produced by this step must be a good representation of the skin temperature.

1.3 Potential Choices

1.3.1 Median

The median is the value that separates the upper and lower halves of the data. The median is found by sorting all of the data points from least to greatest or from greatest to least and then finding the value that is in the middle of this set of data. If the number of data points is odd then a value will exist in the middle of the data set but if the number of data points is even the two data points that are closest to the middle are averaged. One advantage of using the median is that it minimizes the impact of outliers in its representation of the data set.[1]

1.3.2 Variance

The variance measures how far a set of random values from the data set are from the average value of the data set. Variance is calculated by squaring the standard deviation of the set of data. The variance is typically used to identify the causes of variability in a data set.[2]

1.3.3 Mean

In this case the mean used would be the arithmetic mean which is the average of the entire data set. The arithmetic mean is found by adding up all of the values in a data set and then dividing by the number of data points in the data set. The mean is the best used as a representation of the entire data

set because it minimizes the sum of squared deviations from the typical value, meaning that it is a value that is the closest to all values in the set. [3]

1.4 Discussion

The key characteristic of the median is that it minimizes the impact of the outliers in that data set. This is not very important in this case because by the time the data is ready have the summary statistics taken, the outliers have already been removed. The key characteristic of the mean is that it is the best representation of every value in the data set. This is very important because the value produced by the summary statistics will be used to represent the the data set in the model. The key characteristic of the variance is identifying the causes of variability. This is not very useful in this case because the value from the summary statistics will be used to represent the data and not analyze it.

1.5 Conclusion

The mean is the best choice for the summary statistics because it is the most accurate representation of all the data points and it is simple to find. This will be important because the value from the summary statistics will be what is used to relate the data from the camera to the measured value in the model.

2 MODELING

2.1 Overview

The condensed data needs to be related to the measured value in order to produce an equation that represents the relationship between skin temperature and core body temperature. The value produced by the summary statistics represents the skin temperature of a subject and the measured core body temperature of the same subject will be provided. These two pieces of data will be taken from many subjects and then used to create a model. The model will then be able to take just the summary statistics, which is the subjects estimated skin temperature, and produce the core body temperature.

2.2 Criteria

The model must be able to represent the relationship as accurately as possible. This means that the best option for the model will need to be determined by trial and error since there is no known relationship. The order in which the different models are tried should go from least complex to most complex.

2.3 Potential Choices

2.3.1 Linear Regression

A linear regression model is used for modeling the relationship between one dependant variable and one or more explanatory variables. This type of model is typically used for predictions, forecasting and error reduction. The predictive model is made by finding the equation of the line of best fit for a set of dependant and explanatory variables. Then the model can be given an explanatory variable and predict the dependant variable.[4]

2.3.2 Polynomial Regression

A polynomial regression model is used for modeling the relationship between dependant and explanatory variables as an nth degree polynomial. The model is made by fitting a nonlinear relationship between explanatory variables and the corresponding conditional mean of dependant variables. Polynomial regression models are used for nonlinear phenomena.[5]

2.3.3 Ridge Regression

Ridge regression is a type of regression model that is used when there are explanatory variables that have a high correlation. It is very similar to the linear regression but it is more complex since in order to account for the explanatory variables with a high correlation it uses the prediction errors.[6]

2.4 Discussion

The linear regression is the least complex of the models since it is finding a simple line to represent the relationship between core and skin temperatures. The polynomial regression is mildly complex since it is used for modeling nonlinear relationships. The ridge regression is more complex since it is similar to the linear regression but considers independent variables that are correlated and takes the error into account.

2.5 Conclusion

The linear regression will bet the best model to try and test first because it is the simplest of all the models. If the linear regression does not produce a high enough accuracy during testing, the next model tried and tested will be the polynomial regression since it is slightly more complex. If the polynomial regression is not accurate enough, then the ridge regression will be tried and tested. If all three do not produce the accuracy that is desired then other models besides these three may need to

be considered or it may not be possible to produce an accurate relationship between skin and core body temperature.

3 MODEL TESTING

3.1 Overview

Once the model has trained on the set of data that has both the subjects estimated skin temperature and their measured core temperature, it needs to be tested on a new set of data to determine how accurate the model is. This will be done by providing the model with only the estimated skin temperature and comparing the core temperature the model calculates to the measured core temperature. To quantify this comparison it is best to calculate the error between the calculated core temperature and the measured core temperature. This will need to be done for a large set of data to get a good idea of how well the model is working.

3.2 Criteria

The calculation used to find the error needs to be the best representation of the accuracy of the model. To do this it needs to provide valuable information that will allow for an accurate evaluation of any error in the relationship between the calculated core temperature and the measured core temperature.

3.3 Potential Choices

3.3.1 Absolute Error

Absolute error averages the size of the error and it weights each error the same. This is done by subtracting a measured quantity from a calculated quantity and taking the absolute value of the result, repeating this for all sets of measured and calculated values, adding all of these values together and dividing by the number of sets.[7]

3.3.2 Squared Error

Squared error measures the differences between the calculated and measured values by putting a larger emphasis on values that are more consistent throughout the data set. This is done by subtracting a measured quantity from a calculated quantity and squaring the result, repeating this for all sets of measured and calculated values, adding all of these values together and dividing by the number of sets.[7]

3.3.3 Root Mean Square Error

Root mean square error represents the spread of the calculated and measured values from each other. This is done by subtracting a measured quantity from a calculated quantity and squaring the result, repeating this for all sets of measured and calculated values, adding all of these values together, dividing the result by the number of sets, and then taking the square root of the result.[8][9]

3.4 Discussion

The absolute error represents the size of the error of all the data sets and puts the same emphasis on all data sets. Unlike the absolute error, the squared error puts more emphasis on the errors that occur more often in the data sets. The root mean square error measures the magnitude of the error much like the absolute error but it puts a larger emphasis on larger errors.

3.5 Conclusion

The absolute error is the best error calculation for finding the error of this particular model. Any error at all is detrimental in this case. Putting emphasis on larger errors or on errors that occur more often won't offer any extra benefit overall. They do offer benefits for fine tuning the model because they offer extra information but are not needed to know the overall error.

REFERENCES

[1] "Median," https://en.wikipedia.org/wiki/Median, (Accessed on 1/10/2018).

[2] "Variance," https://en.wikipedia.org/wiki/Variance, (Accessed on 1/10/2018).

[3] "Mean," https://en.wikipedia.org/wiki/Mean, (Accessed on 1/10/2018).

[4] "Linear regression," https://en.wikipedia.org/wiki/Linear_regression#Applications_of_linear_regression, (Accessed on 1/10/2018).

[5] "Polynomial regression," https://en.wikipedia.org/wiki/Polynomial_regression#Interpretation, (Accessed on 1/10/2018).

[6] S. Ray, "7 types of regression techniques you should know!" https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/, August 2015, (Accessed on 1/10/2018).

[7] "Squared or absolute? how different error can be." http://archive.is/0hHpF, September 2013, (Accessed on 1/10/2018).

[8] S. Holmes, "Rms error," http://statweb.stanford.edu/~susan/courses/s60/split/node60.html, November 2000, (Accessed on 1/10/2018).

[9] "Mae and rmse-which metric is better?" https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d, March 2016, (Accessed on 1/10/2018).