HUMAN ACTIVITY RECOGNITION

Matt and Prava

ABOUT THE DATASET

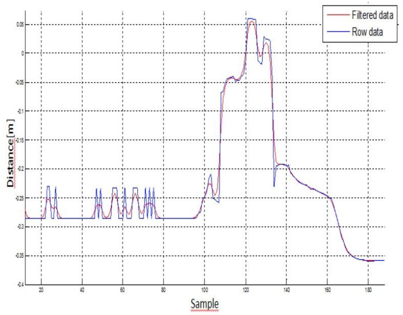
https://devmesh.intel.com/projects/human-activity-recognition

The **Human Activity Recognition** dataset consists of accelerometer and gyroscope data, recorded with a smartphone, of thirty people doing six tasks: walking, walking upstairs, walking downstairs, sitting, standing, and lying down. The dataset is split into 70% train and 30% test data.

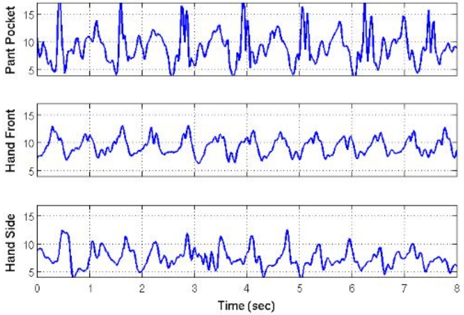
ABOUT PREDICTORS

- 561 predictors
- The time-series smartphone data was filtered and broken up into 2.5 second windows.
- Features such as mean, median, correlation, etc. were calculated for acceleration in X, Y, and Z directions.
- The Fast Fourier Transform was applied to transform the data into the frequency-domain and statistics were calculated from that
- A Butterworth Low-Pass Filter was applied in order to filter out high-frequency noise.

Butterworth Low Pass Filter



Original Time Series Data



https://www.researchgate.net/figure/Result-of-filtering-Kinect-data-with-a-Butterworth-low-pass-filter_fig6_233828357
https://www.edn.com/Home/PrintView?contentItemId=4412351

OUR QUESTION:

How can we build model to predict what human activity is being performed?

FULL DATASET:

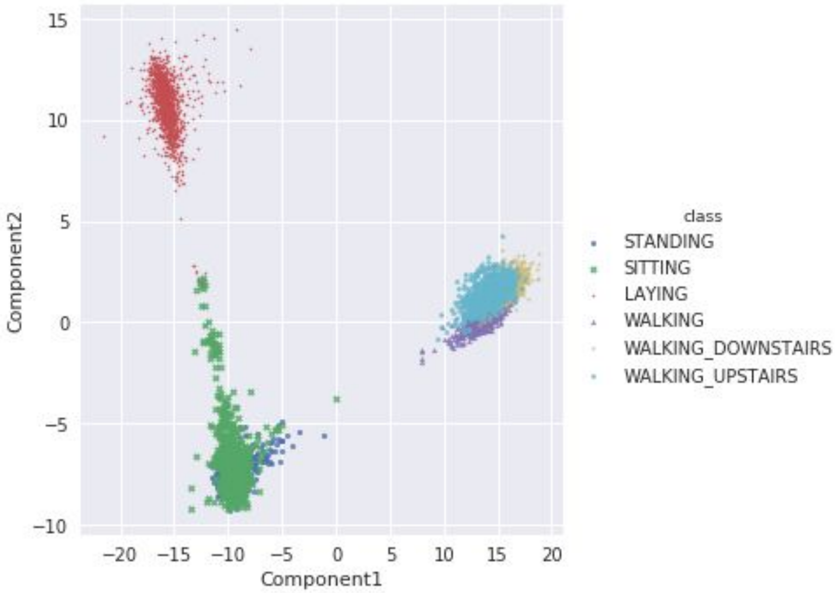
All 561 variables

ANALYZING FULL DATASET: kNN

Here, we perform kNN classification using all 561 predictors across a range of values of k. We obtain a maximum test correct classification rate of 81.7%, when k=10.

ANALYZING FULL DATASET: LDA

Next, we perform LDA classification using all predictors. We obtain a test classification rate of 96.4%.



ISSUES WITH LDA

Although LDA gives a high correct classification rate, this is somewhat misleading. Because the dataset has high collinearity (correlation between the predictors), LDA gives a low error rate, but the model isn’t actually interpretable.

FULL DATASET: SVM AND TREE-BASED

Performing SVM and tree-based classification using all predictors, we obtain test correct classification rates of 93.1% and 85.1%, respectively.

SUMMARY

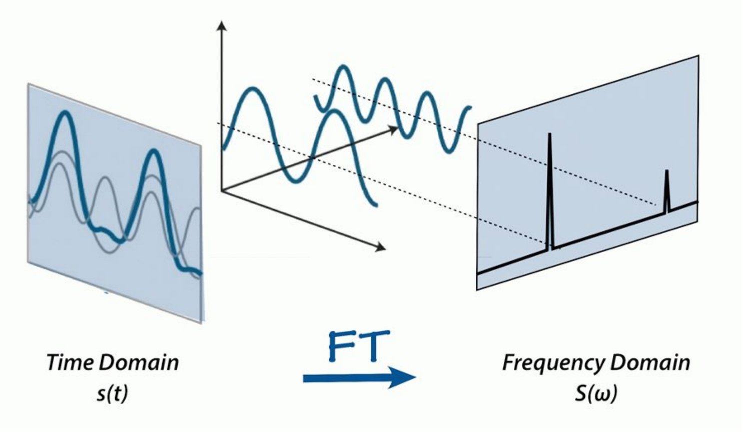
With the entire dataset, we performed KNN, LDA, SVM, and a Tree-Based approach in order to perform human activity recognition. As can be seen, rate of correct classifications are fairly high for all of the different statistical methods performed, noting that LDA has an obscenely high rate of correct classifications (96.4%)!. However, this uses all of the data and thus the model is fairly complex.

Model	Test % Correct
KNN	**81.7%**
LDA	**96.4%**
SVM	**93.1%**
Tree-Based	**85.1%**

TIME AND FREQUENCY:

Splitting into time and frequency subsets

SPLITTING THE DATASET



https://aavos.eu/glossary/fourier-transform/

In order to perform more meaningful analysis on the data set, we wanted to subset the data further into time and frequency data. The given dataset had both, and we wondered if splitting up the data in this way would substantially change our rate of correct classifications.

ANALYZING TIME AND FREQUENCY: kNN

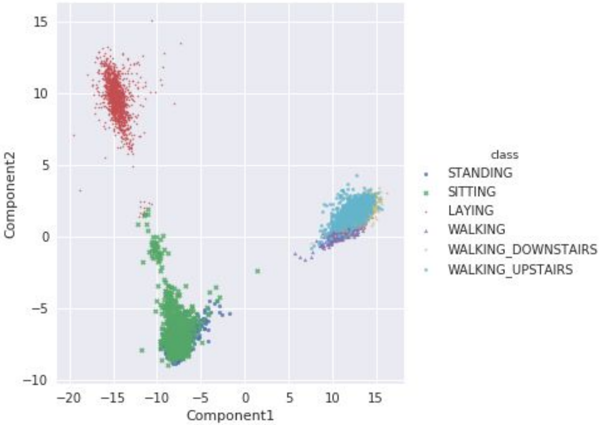
Here, we perform kNN classification using both the time- and frequency-domain predictors.

- Maximum success rate with time predictors: 88.2%, at k=15
- Maximum success rate with freq predictors: 79.3%, at k=25

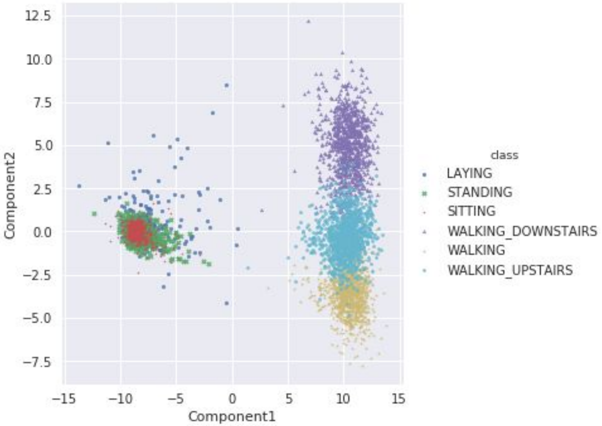
ANALYZING TIME/FREQUENCY: LDA

Next, we perform LDA classification using time- and frequency-based predictors. We obtain test correct rates of 96.3% and 86.8%, respectively.

Time Predictors



Frequency Predictors



TIME/FREQ: SVM AND TREE-BASED

Performing SVM and tree-based classification using all predictors, we obtain test correct classification rates:

- Time-domain predictors:
  - SVM: 83.8%
  - Tree-based: 76.6%
- Frequency-domain predictors:
  - SVM: 93.8%
  - Tree-based: 82.8%

SUMMARY

As can be seen in the results, there's no clear subset of data that is better. Although all models for each subset of data perform reasonably well, KNN and LDA have higher classification rates for the time subset of the data, and SVM and Tree-Based have higher classification rates for the frequency subset of the data. Neither are clearly superior to each other or to the full subset - although their complexity is less then the full subset.

Model	Time Test % Correct	Freq. Test % Correct
KNN	88.2%	79.3%
LDA	96.3%	86.8%
SVM	83.8%	93.8%
Tree-Based	76.6%	82.8%

DIMENSION REDUCTION

Feature selection using SVM

DIMENSION REDUCTION

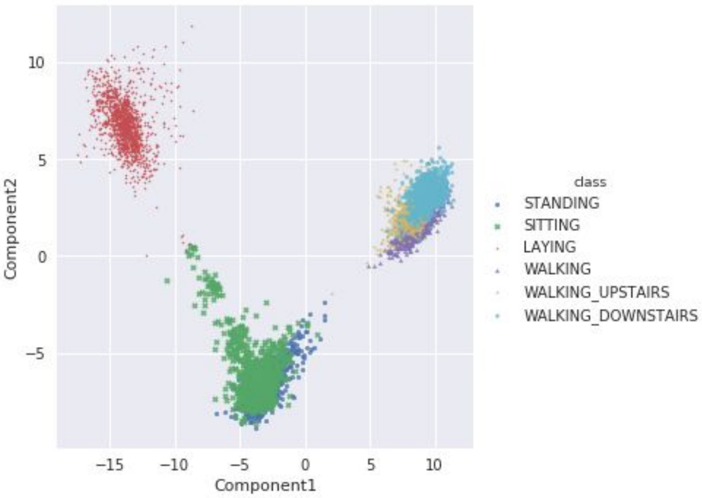
The dataset has 561 variables in it, so using all of the variables as predictors leads to a fairly complex model. We used dimension reduction techniques, specifically, we performed feature selection using SVM. This feature selection reduced 563 variables down to 100 variables - still a fair amount of data but significantly less complex.

ANALYZING PREDICTOR SUBSET: kNN

Here, we perform kNN classification using the predictors subset. We get a maximum test correct rate of 82.0%, at k=5.

ANALYZING SUBSET: LDA

Next, we perform LDA classification using the subset of predictors. We obtain a test correct rate of 95.0%.



SUBSET: SVM AND TREE-BASED

Performing SVM and tree-based classification using all predictors, we obtain test correct classification rates of 92.8% and 85.4%, respectively.

SUMMARY

As seen on the table to the right, the results were also relatively similar. However, the models themselves are less complex and more interpretable, making them a better overall statistical model. Of these, LDA is the best, having a 95.0% accuracy rate.

Model	Test % Correct
KNN	82.0%
LDA	95.0%
SVM	92.8%
Tree-Based	85.4%

IN CONCLUSION

- When applying statistics to dataset for a certain real-life end goal, it’s important to assess tradeoff between accuracy and model complexity
- We can make a small tradeoff in accuracy in order to reduce complexity
- Speed is important for things like human activity recognition