Data Collection for Machine Learning

By Emily Moise, ECE '20

Introduction

The objective of this report is to provide a detailed description of data collection methods and strategies for successful machine learning algorithms. My Senior Design project involves collecting data to train, test, and finally, implement a machine learning algorithm in our application. Machine learning is a tool that can be used (in our case) to classify data based on past patterns and inferences that have been made from feeding the algorithm several samples of data relevant to the type of "thing" that is being classified. To have success with machine learning, my team needs to collect enough data in order to correctly train the algorithm. Therefore, this report will highlight how to best collect data to accomplish our goal of a functioning machine learning algorithm for our Senior Design project.

Basics of Data

What is data?

What are data? More specifically, what are data for machine learning purposes? Datum is a statistic or fact that is used for reference or analysis. In machine learning, the computer learns how to correctly identify relevant data based on the patterns recognized from the training data sets. Data sets are a matrix collection of data points. For projects involving machine learning, the bulk of the time is spent collecting data, to ensure that the computer has sufficient information to make accurate decisions. According to towardsdatascience.com, "If a data set is not good enough, the entire project will fail." This emphasizes the significance of data collection.

Validation and Testing

Three main steps are undertaken to create useful machine learning algorithms: training, validation, and testing. Within this, there are two groups of data sets that are used: training and validation. The training set comprises approximately 60% of the data and it is used to train an algorithm to understand how to apply concepts such as neural networks, to learn and produce results. The validation set comprises approximately 20% of the data and it is used to fine tune the algorithm and to judge the effectiveness of the algorithm based on the training set. It is important that these two sets are different because the algorithm already has classified the data in the training set. If we run the algorithm again on the training set, the results would not show anything new because that is the only data the algorithm knows. Having a validation set that is close to, but different than the training set can identify flaws in the algorithm. Likewise, the validation set should not be too close to the training set or this will cause overfitting. As a result, the algorithm will not recognize any data that falls reasonably outside of the curve. The last step to build confidence in machine learning is testing.

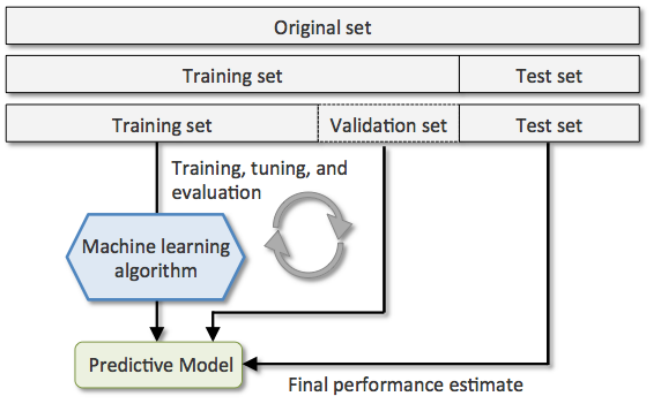

Figure 1. Breakdown of data.

Once the algorithm is working as expected for the training set and the validation set, the designers can use new, raw data to see if the algorithm will still perform well. As expected, the testing data will be less fitted to the curve as the training set and validation were, but as the algorithm interacts with more data, the curve will become better fitted. Figure 1 shows the breakdown of data into training, validation, and testing sets to yield an accurate predictive model.

Data Coverage

How much data do we need?

The last important step in data collection for machine learning is data coverage. In this, it is necessary that we understand how much data we need and the quality of the data that is required for a strong algorithm. According to towardsdatascience.com, we need to collect 10 times as much data as the parameters in the model that is being built. Within this data, we need to make sure that we are collecting the right data. It is necessary to have a data distribution that covers all target classes (things that we want to predict or detect - in our case the squat and deadlift) or values as well as other variables that could have a significant impact on the results. The purpose of collecting data this way is to ensure that any inherent variation that may come up is covered in the training of the algorithm.

Does collection ever stop?

Figure 2 shows how effective it is to have a dynamic data set. This means that as customers use the product and create other data sets, the machine learning algorithm should still be training itself based on the additional user's data. The more users that use the product create more data that goes in the database, which then creates smarter algorithms, which in turn, creates a better product.

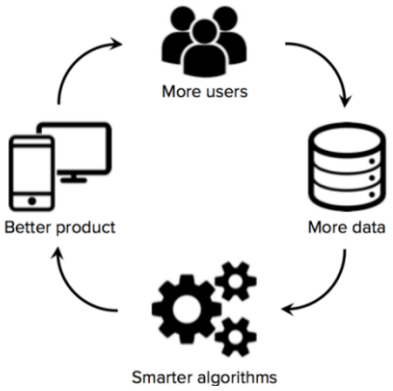

Figure 2. Illustration of the effectiveness of a dynamic data set.

Where to go from here?

Future project ideas

As mentioned before, my project encompasses two weightlifting movements that will be classified by the conclusion of the project. The next phase of this project would be to incorporate more weightlifting movements so that user is able to record their entire workout. In terms of the data collection aspect, some weightlifting movements follow the same pattern (i.e. shoulder press and deadlift both move up and then down). In order to distinguish the two movements, other sensors may need to be used to have an accurate classifier. This would add to the complexity of data collection, as the training and validation groups would now need to include large sets of data for multiple sensors.

Future project technology

In an article published in the 2019 International Conference on Opto-Electronics and Applied Optics, engineers described a project in which they built a classifier and machine learning algorithm to distinguish between eight movements (walking, walking with weight, running, running with weight, sitting, sitting with weight, climbing, and climbing with weight). They found that when just using accelerometer data, the classifier was not able to

distinguish between the movements and the movements with weight. To mitigate this issue, they included heart rate data from a wearable heart rate monitor. They were then able to accurately classify the data, noting that, “in reality, sensor readings for some activity may vary due to different user behavior and intensity level. Thus, in detailed human activity recognition (HAR), the intensity level of activities should be taken into account using heart rate of individual uses.” From the prior paragraph, if we took data from a shoulder press versus a deadlift, since the deadlift is a higher intensity movement, I would be interested to see if our current classifier is able to distinguish between the two movements or if our team would need to integrate another sensor into our design.

Conclusion

In conclusion, data collection is a critical step in the machine learning process. Without thorough data, the project will fail. For my Senior Design project to be successful, my team needs to dedicate hours of time to collect samples of data. By doing this, we will generate a data set that is hopefully expansive enough to train and test a machine learning algorithm that can accomplish the classification of two weightlifting movements.

References

1. Gonfalonieri, A. (2019, February 14). How to Build A Data Set For Your Machine Learning Project. Retrieved from https://towardsdatascience.com/how-to-build-a-data-set-for-your-machine-learning-project-5b3b871881ac

2. Successful Data Collection for Machine Learning with Sensors - Part 1. (2019, March 22). Retrieved from https://reality.ai/successful-data-collection-for-machine-learning-with-sensors-part-1/

3. A. Nandy, J. Saha, C. Chowdhury and K. P. D. Singh, "Detailed Human Activity Recognition using Wearable Sensor and Smartphones," 2019 International Conference on Opto-Electronics and Applied Optics (Optronix), Kolkata, India, 2019, pp. 1-6