Thesis Proposal

Elizabeth Ash

October 16, 2015

Thesis Goals:

- Plan to have all research to be reproducible -> go to GitHub to get recipe to reproduce what I got
- Build an honest predictive model- one to use in real life
- This report is always publicly available on (URL for GitHub)

Research Question

Predictive modelling is a process used to create a statistical model of future behavior. A predictive model is made up of a number of *predictors*, which are variable factors that are likely to influence future behaviors or results. (For example, in marketing could be a customer's age, gender, and purchase history may be predictor variables in the likelihood of a future sale.) In predictive modelling, data is collected for the relevant predictors (the data from the Velloso et. al study was gathered using sensors on common weight lifting equipment), a statistical model is formulated, predictions are made, and the model is revised as additional data becomes available. My research project deals with predictive models and applications of such modelling.

The article *Qualitative Activity Recognition of Weight Lifting Exercises* describes a study involving six male subjects, all in their twenties and with little weight lifting experience. The subjects were taught how to lift a dumb-bell correctly and were also taught how to perform the same movement in four incorrect ways. The five categories of lift data collected were:

```
* Class A: correct lift movement
```

* Class B: throwing the elbows to the front
* Class C: lifting the dumbbell only halfway
* Class D: lowering the dumbbell only halfway
* Class E: throwing the hips to the front

The subjects repeated each lift ten times and during each lift the researchers recorded a number of inertial measurements from sensors in the users' glove, armband, lumbar belt, and dumbbell (these are equipment that are commonly used by weight lifters).

Using this data, Dr. Homer White aimed to devise a random forest model to predict activity-type from variables of the data set. [Random forest is a statistical algorithm that is used to cluster points of data in functional groups. When the data set is large and/or there are many variables it becomes difficult to cluster the data because not all variables can be taken into account. Therefore the algorithm can also give a certain chance that a data point belongs in a certain group. (This is just a basic definition, we will go into further detail in the Methodology section)]. Basically, Dr. White wanted to see if he could make a model that predicted the activity the subject was performing based on the data gathered during each of the lift types. The final model he constructed was estimated to be correct about 99.7% of the time. However, when the model was used to make predictions for new subjects, the results were terrible. The model would not be good to use to predict the activity of a new subject.

This is where my research project begins. The main question I would like to investigate using the data from Velloso, et. al is "Can we predict how someone is performing a weight lift motion based on the numeric information provided from instruments?" I want to tweak Dr. White's model to build an honest predictive model- one to use in real life.

Significance of Project

Activity recognition is an increasingly important technology because it can be applied to many real-life, human-centric problems. Activity recognition can be applied not only in home-based proactive and preventive healthcare applications, but also in learning environments, security systems, and a variety of human-computer interfaces. The goal of activity recognition is to

recognize common human activities in real-life settings.

In the case of my research project, I want to see if a predictive model can be made to recognize certain weight lift motions. Regular physical activity is one of the most important things that can be done for overall health. It can help control weight, lower risk for heart disease, Type 2 diabetes, and some types of cancers, strengthen bones and muscles, and increase chances of longer life. However, if the activity is performed incorrectly, there is a greater risk of injury. To benefit most from a fitness routine, the activity should be performed as accurately as possible. Some people can go to a gym and work with a certified trainer, but many people cannot or will not work with a personal trainer. These people may be doing the correct lift motion, but there is no way to really know unless they are taught the correct motion by a professional.

If an honest predictor model could be made, then the model could be integrated into the weight lift equipment and used to determine if the lift was done correctly or incorrectly. This technology could be used to help reinforce the correct weight lift motion by commending the user for a correct movement, or making a comment when the user made an incorrect movement.

Methodology

The main goal of this project is to build a predictive model using data gathered from inertial measurement units (IMUs) of the Velloso, et. al study. To do so requires several steps of data cleaning, separation, and analyses.

A new model should be able to be used to classify new data. Thus, it is important to have high model performance with new data. The performance of a model is measured in terms of its *error rate*: percentage of incorrectly classified instances in the data set. The simplest way to get a handle on the ability of a predictive model to perform on future data is to try to simulate it. Even though it is impossible to gain access to the future before it occurs, some of the current data can be reserved and treated as if it were data from the future. Such data is referred to as cross-sectional.

The data set I am working with will have to be divided into two sets (a training set and a test set). The training set is used to build the model (determine important parameters) and the test set is new data that is used to measure the model's performance (holding the parameters constant). When the original data set is separated into the training and test sets, the simplest partition is a two-way random partition, careful to avoid introducing any systematic differences. The reasoning behind this type of division is that the data available for analytics fairly represents the real-world processes and that those processes are expected to remain stable over time so a well-constructed model will perform adequately on the new data. The *resubstitution error* (error rate on the training set) is a bad predictor of performance on new data because the model was built to account for the training data, so the model may not generalize to the new data. Thus, to really know if the model would be a good predictor of the weight lift motion, it must be measured on the test data set, not the training set.

Since there are six subjects in the study, a total of 300 lifts were performed and recorded (each subject did 10 repetitions of the 5 lifts). However, during **each** lift the IMU measurements were gathered using a sliding window approach with different lengths (from 0.5 to 2.5 seconds), with a 0.5 second overlap. This resulted in a large data set (over 19,000 observations); a single observation in the data set corresponds to a specific time window for a specific subject performing one of the specified lifts. In his report, Dr. White separated the original data set into training and test sets using a random partition (a typical separation). However, in my project I will attempt to further expand on the partition by ensuring that no single lift appears in both sets. In real life, each new person does a new lift and no lift is the same. While the error rate for the model will most likely be worse with new data, I want to make the model as honest as possible, which means none of the measurements from a single lift should be in both the training and test set.

The two main approaches for building a predictive model are random forests and nearest neighbors. Random forests are more sophisticated, but if I have time I will look at both approaches.

Random forests are a method for classification, regression, and other tasks, that operate by constructing a multitude of decision trees using a training set of data and outputting the

class that is the mode of the classes (for classification) or mean prediction (for regression) of individual trees. The working of a random forest algorithm is as follows:

1. A random seed is chosen which pulls out, at random, a collection of samples from the training data set while maintaining the class distribution
2. With this selected data, a random set of attributes from the original data set is chosen based on user defined values
3. In a data set, where M is the total number of input attributes, only R attributes are chosen at random for each tree where R<M
4. The attributes from this set creates the best possible split. The process repeats for each branch until the termination condition stating that the nodes are too small to split any further

Random forest follows this same methodology and constructs multiple trees for the forest using different sets of attributes.

The nearest neighbors approach, or better known as k-Nearest Neighbors algorithm (k-NN for short), is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. Basically, it works based on minimum distance from what we are predicting (query instance) to the training samples to determine the k nearest neighbors. After gathering the k-nearest neighbors, the majority of the neighbors are taken to be the prediction of the query instance. This is a simple method, but the results can be extremely helpful in determining where to begin for the Random Forest method.

Bibliography

Breiman, Leo, and Adele Cutler. *Random Forests Leo Breiman and Adele Cutler.* Random Forests. 29 June 2007. Web. 20 Sept. 2015. https://www.stat.berkeley.edu/~breiman/

RandomForests/cc_home.htm#intro

Breiman, Leo. *Random Forests.* Machine Learning 45.1 (2001): 5-32. Web. 20 Sept. 2015.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R.* 2013. Print.

Mitchell, Tom. M. 1997. *Machine Learning.* New York: McGraw-Hill.

Steinberg, Dan. *Why Data Scientists Split Data into Train and Test.* Why Data Scientists Split Data into Train and Test. 3 Mar. 2014. Web. 27 Sept. 2015.

Velloso, Eduardo, Andreas Bulling, Hans Gellersen, Wallace Ugulino, and Hugo Fuks. *Qualitative Activity Recognition of Weight Lifting Exercises.* Proceedings of the 4th Augmented Human International Conference on - AH '13 (2015).

White, Homer. *Predicting Movement-Types: Quick Model-Making with Random Forests.* Https://github.com/homerhanumat/WeightLifting. 4 Aug. 2015. Web. 1 Sept. 2015.

Witten, Ian H., and Eibe Frank. 2000. *Data Mining: Practical Machine Learing Tools and Techniques with Java Implementations.* San Diego, CA: Morgan Kaufmann.

Xie, Yihui. *Dynamic Documents with R and knitr.* 2014. Print.

Working Outline

1. Introduction
    a) Background information on project data
    b) Background information on the various approaches of model making
    c) Significance of this project
    d) Brief summary of what was done to make the model
2. Body
    a) Include the steps taken to create training/test data sets
    b) Include the steps taken to create the model

- Did I look at multiple model approaches?

c) Include any figures, graphs, charts, etc.

3. Conclusion

    a) Discussion on what everything means

    b) Any further investigations or applications?

Timeline for Completing Project

- Oct 1st: Complete first draft of Thesis Proposal
- Oct. 16th: Complete final draft of Thesis Proposal and send to Dr. Burch
- By end of October finish the Coursera courses