Project 7

Raj Shrestha, Zach Ellis

Introduction

Our slides on the analysis can be found here.
The datasets that will be analyzed have benefits to two specific groups of people. The first one is for people who enjoy working out, and would like to do it even better. The second set is for anyone who uses a credit card. The first issue that we decided to look into is correcting form for workouts. A lot of people would like to work out and begin but their form is incorrect. This can end up doing more damage than good. The purpose was to attempt to classify incorrect types of bicep curls so the person performing the workout can receive guidance. The second dataset is important for credit card companies and credit card users alike. Being able to classify if a charge is fraudulent or not is very important. The purpose of using machine learning on this dataset is to provide proper feedback for classifying fraudulent charges.

Dataset

For our analysis we choose two different datasets. First one is a Weight Lifting Exercise Dataset. The dataset was collected from 4 different sensors placed on Arm, Forearm, Dumbbell and Waist. This dataset helps to investigate aspects that pertain to the correct execution of dumbbell exercise and can be used to provide feedback on quality of execution through classification of Biceps Curls. The dataset consists of 39242 rows and 159 different attributes with 5 different class labels: A (Correct Execution), B (throwing elbows to front), (lifting dumbbell halfway) C, (lowering dumbbell halfway) D and (throwing the hips to the front) E. There were a lot of attributes with “NA” values. We dropped all columns with null values for our analysis. We also removed non-numeric attributes that had no predictive value. The remaining 53 attributes were data (quantitative) measured by accelerometers and gyroscopes of sensors and one class label of the position.

For the second set of data we chose Credit Card Transactions. These are transactions made by credit cards in September 2013 by european cardholders over the course of two days. The dataset that we obtained has only numerical input variables which are the result of a Principal Component Analysis (PCA) Transformation. This is because of two reasons: a lot of the data was confidential and making the input variables numeric instead of strings, makes using a machine learning algorithm much more viable. There was no need to clean the data because most of the cleaning had already been performed by the PCA Transformation. The only categories that the PCA wasn't applied to is the time and the amount of the charge.

Analysis Technique

For our analysis we started to explore and understand the dataset, their attributes, values and importances in prediction. For the Weight Lifting dataset there were different gyroscopes and accelerometers. SInce we are trying to identify the position of dumbbells in space we decided to use x, y, z values of all sensors (Arm, Forearm, Dumbbell and Waist). We also have "roll", "pitch" and "yaw" data of 4 sensors. We plotted 3D scatter plots of some attributes (Fig 1 & 2) and decided whether to include those attributes in our classifier or not. We then used a logistic regression and linear SVM to compare the results. We also used GridSearchCV to find the best type of SVM and parameters for our dataset.

For the analysis of the credit card charges dataset, we initially began by visualizing the data by simply creating a scatter plot (Fig 3a and 3b) of the fraudulent charges and all the charges. as was mentioned previously, the PCA transformation made it easier to perform the SVM. However, the one takeback is that we don't really get to understand what factors play into the fraudulent charges. There are 30 different columns or categories that play into the fraudulent charges. These would be categories such as "Where the person lives", "Where the charge was made", and "." We pulled just over half of the data entries for the training set. Using the training dataset, we trained a Support Vector Machine with a linear kernel. Then we tested the fit against the other half of the dataset.

Results

- Which variables are predictive of the target variable?
    - For the Weight Lifting Dataset the values measured by 4 different sensors are predictive of the target variable. In addition to that we saw 3d scatter plots of multiple variables that helped us to decide whether to include those attributes or not. Fig 1 and 2 shows the 3D scatter plots where we can see clear boundaries between class labels for 6 attributes.
- Can logistic regression or a linear SVM predict well?
    - For the Weight Lifting Dataset both logistic regression and linear SVM predicted poorly. Logistic regression works best for independent attributes. But since our data is generated from sensors and the position of the sensor is highly dependent with other attributes SVM is suitable for our analysis. However, linear SVM performed poorly compared to 'rbf'. The f1-score for linear SVM was: [0.95, 0.89, 0.88, 0.86, 0.93] for 5 classes and f1-score for logistic regression was: [0.8, 0.63, 0.63, 0.63, 0.63].
    - The SVM with a linear kernel predicts decently well for the fraudulent charges.
- What do plots of selected pairs of variables look like? Where is the decision boundary in those plots?

    - For the Weight Lifting Dataset there were more than 50 attributes on the dataset collected from sensors (Gyroscopes & Accelerometers). Furthermore from the 3d scatter plots with yaw, roll and pitch of (arm, forearm, belt and dumbbell), we decided to include those attributes to. Fig **1 & 2** shows the scatter plots where we can see the blobs of data points of different classes.
- Is there a difference between the polynomial and RBF SVMs?
    - For our exercise dataset RBF SVMs performed well better than polynomial SVM. Table 1 shows the comparison of f1-scores of different types of SVM for exercise dataset. From the table we can see that RBF SVM has higher f1-scores than other models.
- What effect does changing the class_weight in an SVM have on your data? How might this be important for this data?
    - Class weights are great when handling the biased or skewed distribution of the classes. Changing the class weight on Exercise Dataset decreased the f1, precision and recall score of the classifier. Later changing the class weight to "balanced" increased the f1-score, recall and precision of the classifier. One of the reasons might be the balanced distribution of data across all classes as seen in the confusion matrix. (Fig 3)
- Is there a difference in runtime performance?
    - Linear SVM took significantly longer time to fit on training data than Poly and RBF SVMs.
- Logistic regression and LinearSVC use one-vs-rest (OVR) for multi-class classification. SVC uses one-vs-one (OVO). Where n is the number of classes, OVR learns n models, whereas OVO learns n(n-1)/2 (n choose 2) models. What effect does this have on performance?
    - Using a Logistic Regression and LinearSVC comparatively takes less time on our dataset as there are 5 different classes and we only have to build 5 (OVR) models. Whereas, SVC took significantly longer time to complete the training especially LinearSVM.

For the Credit Card dataset, without knowing which variables are which it is hard to guess which ones are predictive of the target variable. We would run an analysis where we take out one variable from the data at a time and run the same analysis, record the F1 scores and then see which one had the greatest difference to find the variable most predictive of the target variable. However, with the quantity of the data this was not completed in time. The SVM with a linear kernel predicted decently well considering how few fraudulent charges there are in general.

Ultimately, we discovered that using RBF SVM for the weight lifting dataset can provide very high F-1 scores. This is beneficial because it means that using the RBF SVM algorithm one could, with high accuracy discover incorrectness in form of a bicep curl, and quite probably other work out forms. As for the Credit Card charges, predicting 130,645 of 134,807 charges correctly is accurate. More importantly it only missed 22 charges deeming them not fraudulent when they actually were. Considering that this is out of 134,807 charges over a two day period, that is a high quality algorithm.

Model	Class 1	Class 2	Class 3	Class 4	Class 5
Logistic	0.79	0.63	0.62	0.63	0.62
Linear SVC	0.73	0.48	0.25	0.37	0.46
Poly SVM (degree=2)	0.95	0.87	0.85	0.81	0.9
Poly SVM (degree=4)	0.96	0.9	0.88	0.87	0.93
RBF SVM	0.99	0.97	0.94	0.94	0.98

Table 1: Comparison of F-1 scores for 5 classes of different Classifiers on Exercise Dataset.

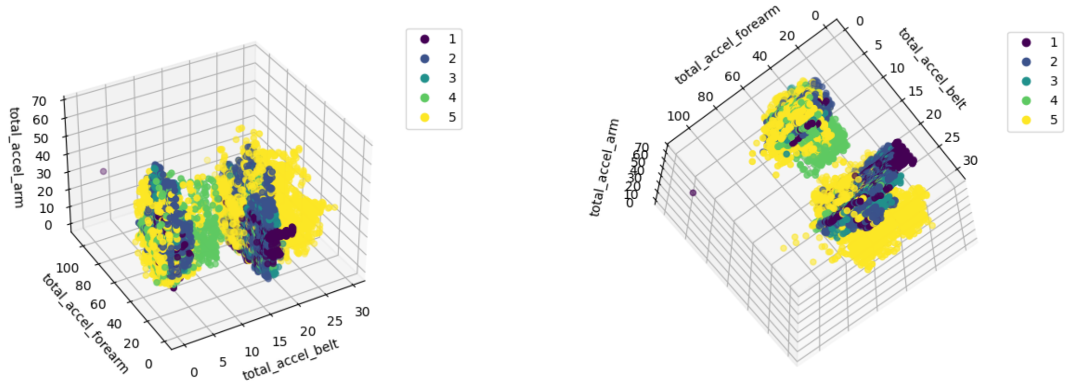


Fig 1: 3D plot of (total_accel_arm", "total_accel_forearm", "total_accel_belt) with class Labels

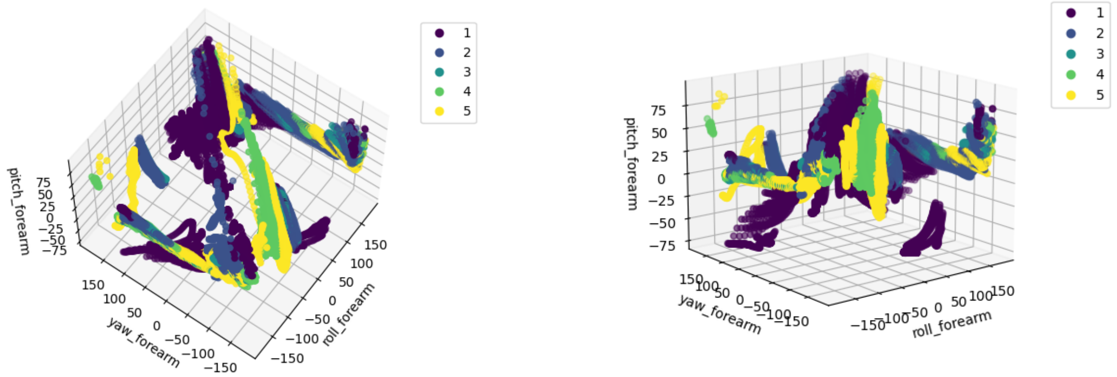


Fig 2: 3D plot of (yaw_forearm, roll_forearm, pitch_forearm) with Class Labels.

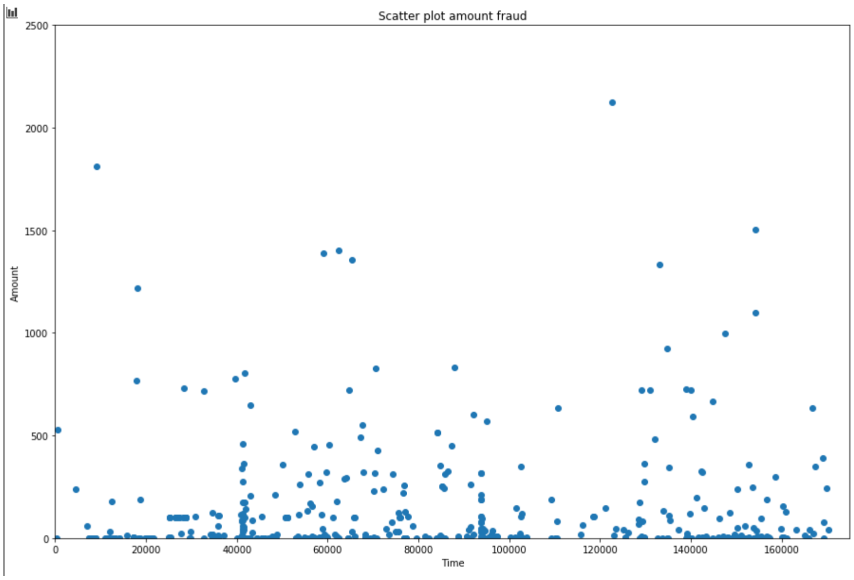


Fig 3a: Fraudulent Charges (Amount over Time)

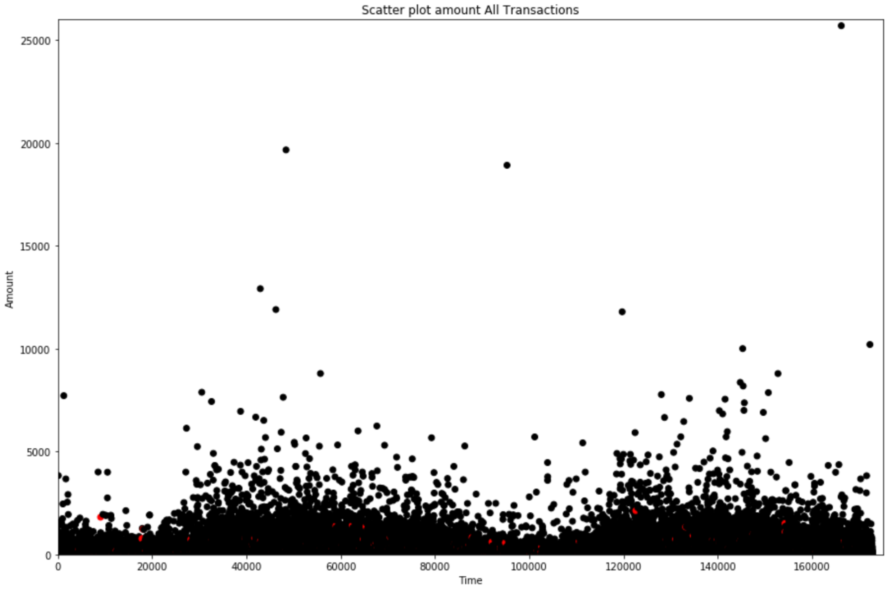
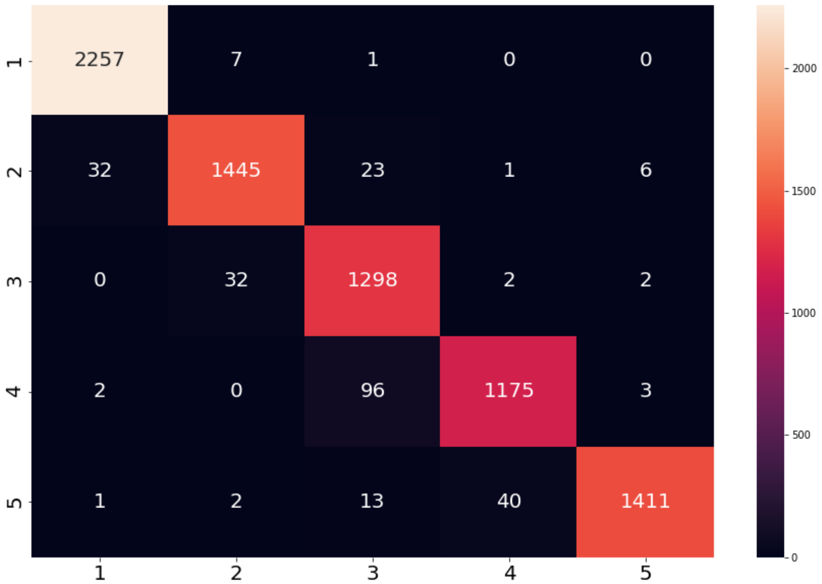
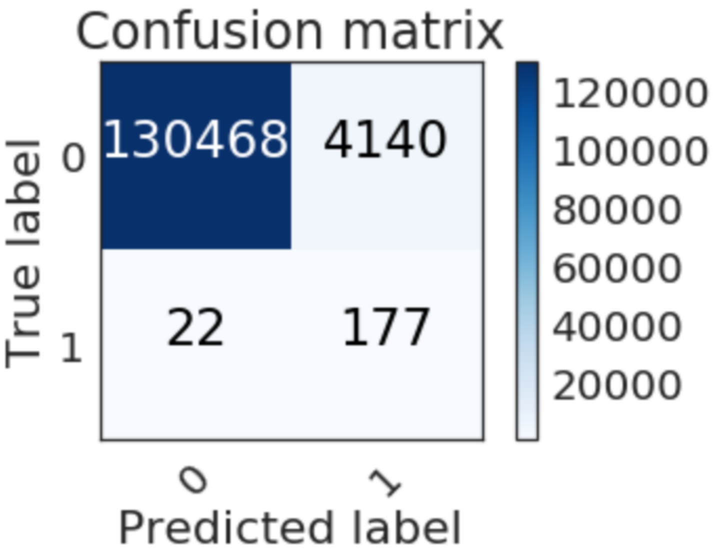


Fig 3b: All Charges (Amount over Time)



	1	2	3	4	5
1	2257	7	1	0	0
2	32	1445	23	1	6
3	0	32	1298	2	2
4	2	0	96	1175	3
5	1	2	13	40	1411

Fig 3: Confusion Matrix of RBF SVM



Confusion matrix

True label	0	1
0	130468	4140
1	22	177

Fig 4: Confusion Matrix of Fraudulent Charges (0,1 are Not Fraud and Fraud, respectively)