Udacity Machine Learning Engineer Nanodegree
Capstone Proposal

Jakaria Rabbi
October 27, 2020

Domain Background

Human Activity Recognition is the problem to classify human day to day activity using smartphone sensors data. Data continuously generated from the accelerometer and gyroscope and these data are very useful to predict our activities such as walking or standing. There are lots of datasets and ongoing research on this topic. In a survey paper [1], I have seen some works with wearable sensor data and prediction with machine learning techniques. Wearable devices can predict a large range of activities by using data from various sensors. Deep learning models are being used to predict various human activities [2]. Nowadays people use smartphones almost all the time and use many wearable devices. Through these devices, physical and mental health can be monitored by predicting human activity, and nowadays it is an efficient, cheap, and safe way to do this as the covid19 pandemic is ongoing.

Problem Statement

I have selected a dataset from the UCI machine learning repository [3] to calculate the accuracy of three machine learning models and to perform some statistical significance tests. I am planning to use a Support Vector Machine, Logistic Regression, and Neural Network with a hidden layer to predict the activity of humans from mobile data. Here the human activities are classified into six categories: walking, walking upstairs, walking downstairs, sitting, standing and laying. I want to see the performance of the models on the basis of classification results and try to get the highest possible accuracy from these models. The sensor data in the dataset are collected from two different sensors and can predict the six different activities. In this project, I will try to get the highest possible performance by these algorithms by parameter tuning, cross-validation and finally comparing the result with two statistical significance tests to get the winner algorithm.

Dataset and Input

The dataset is taken from the UCI machine learning repository [3]. The dataset contains information from 30 volunteers within the age range: 19-48. Each volunteer performs six activities:

- Walking
- Walking upstairs
- Walking downstairs
- Sitting
- Standing
- Laying

Hence, the dataset has six labels to predict. Dataset consists of 561 feature vectors with time and frequency domain variables. These features come from the following data:

- Gravitational acceleration for x, y, and z axes
- Body acceleration data files for x, y, and z axes
- Body gyroscope data files for x, y, and z axes

I also have the ID of an individual volunteer for each record. There are a total of 10299 records and training and test dataset splits are 70%-30%.

Solution Statement

Support Vector Machine [4], Logistic Regression, and Neural Network [5] with a single hidden layer will be used to predict the activity of humans from mobile data that are collected from the UCI machine learning repository [3]. The highest possible performance of these algorithms by parameter tuning, cross-validation and finally comparing the result with two statistical significance tests to get the winner algorithm. Here, the performance means the accuracy of correctly classifying one of the six activities from the mobile sensors data.

Benchmark Models

I will use Logistic Regression (LR) as a benchmark model and it focuses on maximizing the probability of the data. This is a basic model without much complexity like neural networks. The farther the data lies from the separating hyperplane (on the correct side), the happier LR is [6]. I will use the kernel trick and 'lbfgs' optimizer to compare the performance. Some previous research has some good results with that combination of LR [7]. Hence, I will use this algorithm as LR and its variants will be a good choice for benchmarking.

Evaluation Metrics

The following evaluation metrics will be used:

Accuracy classification score

Accuracy Classification score is defined as the number of correct predictions divided by the total number of predictions. These metrics are very important as we want to know whether a human is walking or sitting correctly. This is the main metric to distinguish between our models.

Project Design

Dataset has a training and test portion. Train and test data will be stored in the data frame initially. I will use all the features for training and testing.

The learning will be divided into two parts. Initially, the best parameters for all the algorithms have to be learned. Then, in the second step, the algorithms will be compared. The dataset will be evaluated by a K-fold cross-validation set. I will use k=5 for cross-validation. The data splitting will be done by stratified sampling.

Parameter

For finding best hyperparameters, the cross-validation dataset will be used for an exhaustive grid search. This cross-validation will learn a model for all combinations of the listed parameters. Table 1 (for SVC), Table 2 (for ANN) and Table 3 (for LR) have the parameter list that I will use to get the best combination of parameters. The parameters are chosen from a range of values. Some parameters are taken as a good selection because they work well in many problems. Accuracy will be used as a scoring method.

Parameters	Values	Description
Kernel	[Linear, RBF, Sigmoid]	Specifies the kernel type to be used in the algorithm
C	[0.1, 0.5, 1, 2, 5, 10, 100]	Penalty parameter C of the error term.

Table 1. Parameters for SVC

Parameters	Values	Description
Hidden layer sizes	[(10,), (50,), (100,)]	Number of hidden nodes
Alpha	[1e-4, 1e-3, 1e-2]	L2 regularization parameter
Learning rate	[1e-3, 1e-2, 1e-1]	Step size
Beta1	[0.1, 0.5, 0.9]	Adam specific parameter
Beta2	[0.1, 0.5, 0.9]	Adam specific parameter

Table 2. Parameters for ANN

Parameters	Values	Description
Kernel	[Linear, Hamming]	Specifies the kernel type to be used in the algorithm
Regularizer	[L1, L2]	Specifies the type of regularization
C	[0.1, 0.5, 1, 2, 5, 10, 100]	Penalty parameter C of the error term.

Table 3. Parameters for LR

Evaluation

After finding the best parameters, I will evaluate the three algorithms on test data. Test data will be divided into 10 random sets and every set will consist of 50% of the data. Hence, I can have 10 runs for every algorithm and calculate the mean and variance for these runs. The learning process to find the best parameters will use only training data. Hence, I will get an unbiased estimation for cross algorithm comparisons.

Statistics

I will use hypothesis testing for this project. A hypothesis is tested by measuring and examining a random sample of the population being analyzed. A random population sample is used to test two different hypotheses: the null hypothesis and the alternative hypothesis. The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population mean return is equal to zero. The alternative hypothesis is effectively the opposite of a null hypothesis; e.g., the population mean return is not equal to zero. Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true. All hypotheses are tested using a four-step process:

1. The first step is for the analyst to state the two hypotheses so that only one can be right.
2. The next step is to formulate an analysis plan, which outlines how the data will be evaluated.
3. The third step is to carry out the plan and physically analyze the sample data.
4. The fourth and final step is to analyze the results and either reject the null hypothesis, or state that the null hypothesis is plausible, given the data [8].

The algorithms will be compared using Welch's t-test [9] method because the variances of the algorithms are different. I can find output from t-test that shows whether the algorithms have significant changes in their performance measurements. The t test tells us how significant the differences between groups are; in other words it lets us know if those differences (measured in means) could have happened by chance. The two-tailed t-test will be done pairwise for every algorithm with H_0: $\mu_0 = \mu_1$ and $\alpha = 0.05$ (significance level) [10] to produce a ranking between the three algorithms and for finding the winner algorithm. Here, H_0 is the null hypothesis and μ_0

and μ_1 are the means of two groups of population. Also, the algorithms will be compared using 5 times 2-fold cross-validated paired t-test as suggested by researchers [11].

References

[1] O. D. Lara and M. A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," in *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192-1209, Third Quarter 2013. doi: 10.1109/SURV.2012.110112.00192

[2] T. Plötz and Y. Guan, "Deep Learning for Human Activity Recognition in Mobile Computing," in *Computer*, vol. 51, no. 5, pp. 50-59, May 2018. doi: 10.1109/MC.2018.2381112

[3] https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones

[4] https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[5] D. P. Kingma and J. Ba, \Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[6] http://www.cs.toronto.edu/~kswersky/wp-content/uploads/svm_vs_lr.pdf

[7] G. C. Cawley and N. L. C. Talbot, Efficient approximate leave-one-out cross-validation for kernel logistic regression, Machine Learning, vol, 71, no. 2-3, pp. 243--264, June 2008.

[8]https://www.investopedia.com/terms/h/hypothesistesting.asp[9] https://en.wikipedia.org/wiki/Welch%27s_t-test

[10] https://en.wikipedia.org/wiki/Null_hypothesis

[11] Dietterich TG (1998) Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput* 10:1895–1923.