Machine Learning Engineer Nanodegree

Capstone Proposal:

Chin Min Tee
January 19, 2018

Early Diagnosis of Parkinson’s Disease via keystrokes

Domain Background

Parkinson’s disease sickened over 6 million individuals worldwide. The impact extended to their immediate families and friends, societies, medical communities and many more resources are not negligible. This neurodegenerative disease is incurable but early discovery indisputably benefit the wellbeing of the patient and all involved. Experiment has been conducted to use keyboard stroke as a mean to allow non-clinical personnel to make a guided diagnosis. This help captures the pre-motor phase of the disease, which could be years, or decades before the degeneration and the tell-tale symptoms of the failing motor control.

Inspired by the study 'High-accuracy detection of early Parkinson's Disease using multiple characteristics of finger movement while typing' recorded in physionet.org, I am eager to put machine learning to work to build a great classification model so that precious opportunity to start treating or researching the disease is not lost.

Problem Statement

The goal is to take in data sets provided by physionet.org for the study 'High-accuracy detection of early Parkinson's Disease using multiple characteristics of finger movement while typing' and build a satisfactory binary classification model out of it. The model will be able to predict by using patient’s keystroke data input if he or she has Parkinspn’s disease. The test data set will come from another project also hosted by physio.net: neuroQWERTY MIT-CSXPD. Although different in structure and names, the data sets contain similar data that can be extracted to form the same features as were used in training set and then apply as test data set.

Datasets and Inputs

There are 2 datasets from two different sources that we will combine to form a larger datasets. The two sources are both from physionet.org, one is 'Tappy Keystroke Data' (Tappy hereafter) and the other is 'neuroQWERTY MIT-CSXPD Dataset' (NQ hereafter).

In the Tappy dataset, there are two files that need to be zipped together to form pandas DataFrame for modeling. There are ArchivedUsers.zip (User files) and ArchivedData.zip (Data files). The User file has 10 character code as filename that represent an individual user, that 10 character code is also present in corresponding data file for that user.

The User file contains the following:

- Birth Year: Year of birth
- Gender: Male/Female
- Parkinsons: Whether they have Parkinson's Disease [True/False]
- Tremors: Whether they have tremors [True/False]
- Diagnosis Year: If they have Parkinson's, when was it first diagnosed
- Whether there is sidedness of movement [Left/Right/None] (self reported)
- UPDRS: The UPDRS score (if known) [1 to 5]
- Impact: The Parkinsons disease severity or impact on their daily life [Mild/Medium/Severe] (self reported)
- Levadopa: Whether they are using Sinemet and the like [Yes/No]
- DA: Whether they are using a dopamine agonist [Yes/No]
- MAOB: Whether they are using an MAO-B inhibitor [Yes/No]
- Other: Whether they are taking another Parkinson's medication [Yes/No]

The Data file:

- UserKey: 10 character code for that user
- Date: YYMMDD
- Timestamp: HH:MM:SS.SSS
- Hand: L or R key pressed
- Hold time: Time between press and release for current key mmmm.m milliseconds
- Direction: Previous to current LL, LR, RL, RR (and S for a space key)
- Latency time: Time between pressing the previous key and pressing current key. Milliseconds
- Flight time: Time between release of previous key and press of current key. Milliseconds

The field Parkinsons is the label for the dataset, and the rest would eventually be features. The medicine mitigating the symptoms may have an effect on the keyboarding activity. They may be discarded if deemed irrelevant. The direction of the movement over keyboard will be examined and represented with appropriate categories.

The data file logged all 8 fields in a row and there could be as many as 150,000 rows per user depending on the length of the keyboarding activity. There are latencies that exceeds a certain threshold and the row will be treated as irrelevant. The threshold will be determined with reference to some medical study.

As for the NQ dataset, the data consists of the following:

- The key pressed.
- The hold duration in seconds.
- The key release time in seconds from time 0.
- The key press time in seconds from time 0.

The 'hold duration' is the same type of information with the 'Hold time'. The 'Flight time' will need to be calculated in this test file with 'key release time' and 'key press time'. To get the 'Direction' from this test set, it needs some calculation as well from the 'key pressed' from two adjacent rows.

The Tappy dataset has 227 users, but only 217 users has actual activities. And Among those, 48 are healthy and 169 have Parkinson's disease.

For the NQ dataset, there are two sets of experiments, PD_MIT-CS1PD and PD_MIT-CS2PD. When combined they have 85 users, 43 are healthy and 42 have Parkinson's disease.

The Tappy dataset is imbalanced, a ratio of 169/48, or 3.5. The NQ is balanced. Therefore in order to mitigate this I am combining the two datasets. The total would be 302 users, with 91 healthy and 211 have Parkinson's disease. With the 80/20 training and test set split, I could have a training set of 240 users with 73 healthy and 168 have Parkinson's disease. The PD over healthy ration would be 168/73, or 2.3.

Solution Statement

A binary classification model with satisfactory performance will be produced. Features will be extracted from raw data to form suitable pandas DataFrame for training and testing. Cross validation technique and hyper parameter optimizing technique will be applied. SVM, DecisionTree, RandomForest and gradient boosting and maybe deep learning will be used to build a model that can perform well. Test metric will include confusion matrix and F1 score and ROC curve.

Benchmark Model

I would use a model that either predicts all positive PD or negative PD, whichever is greater. Then a random estimator will be run for few iterations and the average used as a benchmark if it is better than the ‘All positives/negatives’ model.

Evaluation Metrics

Confusion matrix and F1 score will be used to evaluate models instead.

		Predicted	
		1	0
Actual	1	True Positive (**TP**)	False Negative (**FN**)
	0	False Positive (**FP**)	True Negative (**TN**)

Table 1. Confusion Matrix

Accuracy is intuitive and easy to understand. It is the correct predictions over the total predictions. However it **would not be suitable** here since we have a somewhat imbalanced dataset. It could lead to seemingly high performance while the only thing model will do is predicting everyone to have PD, which is majority of the two classes.

$$\boldsymbol{Accuracy} = \frac{\mathbf{TP + TN}}{(\boldsymbol{TP + TN + FP + FN})}$$

Precision is the correct positive predictions over all positive predictions. It answers question: How correct a model is when it predicts positive?

$$\boldsymbol{Precision} = \frac{\mathbf{TP}}{(\boldsymbol{TP + FP})}$$

Recall (also known as Sensitivity) explains that out of all actual positive cases how many are correctly predicted as positive.

$$\boldsymbol{Recall} = \frac{\mathbf{TP}}{(\boldsymbol{TP + FN})}$$

Specificity explains that out of all actual negative cases how many are correctly predicted as negative.

$$\boldsymbol{Specificity} = \frac{\mathbf{TN}}{(\boldsymbol{TN} + \boldsymbol{FP})}$$

F1 score is a weighted average of precision (P) and recall (R).

$$\boldsymbol{F1} = \frac{\mathbf{2 * PR}}{(\boldsymbol{P} + \boldsymbol{R})}$$

Project Design

Tasks outlined:

- Read in data sets for both training and test and preprocess them to a useful and suitable state for pandas DataFrame.
  The raw data consists of series of keystrokes and characteristics of movement over time. It needs to be analyzed to extract much needed data and reject long pauses since it means a break from the keyboarding activity.
- Data exploration to gain insight. There will be statistics and visual graphs.
  This will eventually allow meaningful features to be developed and still applicable to test set.
- Proper transformation to data to facilitate modeling. Modeling involves mathematic calculation and thus name and string for example will need to be transform to numbers to fit in modeling. Sklearn preprocessing OneHotEncoder may be applied to categorical features. StadardScaler may be applied as well to normalize features.
- These features should include:
  Gender
  Age
  Hold time (mean and standard deviation)
  Going right flight time (mean and standard deviation)
  Going left flight time (mean and standard deviation)
  Same side same key flight time (mean and standard deviation)
- Build classification model
  A few models will be attempted. This may include Decision Tree, Random Forest, gradient boosting, deep neural network and may even attempt convolution neural network if possible.
- Cross validation technique from scikit-learn model_selection ShuffleSplit and k-fold will be used to provide more robust training.
- Scikit-learn model_selection hyper-parameter optimizer RandomizedSearchCV will be applied to hunt for best parameters of models.

- Establish test and performance metrics. Confusion matrix will be reported. Scikit-learn classification scoring function like f1_score can be used to evaluate models.
- Make available in github.com so that anyone can potentially benefit from it or help improve further.

References

1. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **101**(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215]; 2000 (June 13).

2. L. Giancardo, A. Sánchez-Ferro, T. Arroyo-Gallego, I. Butterworth, C. S. Mendoza, P. Montero, M. Matarazzo, J. A. Obeso, M. L. Gray, R. San José Estépar. Computer keyboard interaction as an indicator of early Parkinson's disease. Scientific Reports 6, 34468; doi: 10.1038/srep34468 (2016)

3. http://news.mit.edu/2015/typing-patterns-diagnose-early-onset-parkinsons-0401
   Anne Trafton, MIT News Office
   April 1, 2015

4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5708704/
   Holger Fröhlich
   2017