Two Factor Authentication

September 16, 2022

1 Datasets

Datasets we have looked into, and the following is what we learnt about them:

1. MIT-BIH Arrhythmia Database Directory
2. ECG-ID Dataset
3. MIT-BIH P wave Database

1.1 MIT-BIH Arrhythmia Database Directory:

About: The source of the ECGs included in the MIT-BIH Arrhythmia Database is a set of over 4000 long-term Holter recordings that were obtained by the Beth Israel Hospital Arrhythmia Laboratory between 1975 and 1979. Approximately 60 percent of these recordings were obtained from inpatients. The database contains 23 records (numbered from 100 to 124 inclusive with some numbers missing) chosen at random from this set, and 25 records (numbered from 200 to 234 inclusive, again with some numbers missing) selected from the same set to include a variety of rare but clinically important phenomena that would not be well-represented by a small random sample of Holter recordings. Each of the 48 records is slightly over 30 minutes long.

Features

In most records, the upper signal is a modified limb lead II (MLII), obtained by placing the electrodes on the chest. The lower signal is usually a modified lead V1 (occasionally V2 or V5, and in one instance V4); as for the upper signal, the electrodes are also placed on the chest. This configuration is routinely used by the BIH Arrhythmia Laboratory. Normal QRS complexes are usually prominent in the upper signal. The lead axis for the lower signal may be nearly orthogonal to the mean cardiac electrical axis.

1.2 ECG-ID Dataset:

About: The database contains 310 ECG recordings, obtained from 90 persons. Each recording contains:

ECG lead I, recorded for 20 seconds, digitized at 500 Hz with 12-bit resolution over a nominal ±10 mV range; 10 annotated beats (unaudited R- and T-wave peaks annotations from an automated detector); information (in the .hea file for the record) containing age, gender and recording date. The records were obtained from volunteers (44 men and 46 women aged from 13 to 75 years who were students, colleagues, and friends of the author). The number of records for each person varies from 2 (collected during one day) to 20 (collected periodically over 6 months).

The raw ECG signals are rather noisy and contain both high and low frequency noise components. Each record includes both raw and filtered signals:

Signal 0: ECG I (raw signal) Signal 1: ECG I filtered (filtered signal)

1.3 MIT-BIH P wave Database:

About: This database contains reference P-wave annotations for twelve signals from the MIT-BIH Arrhythmia Database. Arrhythmic dataset contains 48 recordings and Each recording is recorded for 30 minutess. Everyone has 23 recordings. Raw files are in the .atr, .dat, .hea. These files are recorded by using MATLAB and these signals are recorded signals. The Signal recorded is from two resources MLII and V5 which are placed on the chest. MLII is the upper signal and V5 is considered as Lower Signal.

CONCLUSION: The database we will be using for project is MIT-BIH Arrhythmia Database Directory as it has two features which we can use to train and identify individuals. The ECG-ID Dataset has only one feature which is not ideal when compared to the MIT-BIT Arrhythmia which has more features. We have tried converting the raw files from the MIT-BIH P wave Database but the method we have used to convert them gave us a .csv file from .dat file but the data is not in form of way we can use to move forward in our project.

2 Tool

The tool we have used for collecting the data collected from the heartbeat of people which is in the form of .csv file is Physiobank ATM. Physio Bank's Automated Teller Machine is a self-service facility for exploring Physio Bank using your web browser. Currently, its toolbox includes software that can display annotated waveforms, RR interval time series and histograms, convert WFDB signal files to text, CSV, EDF, or .mat files.

2.1 Process of Using Physio Bank Atm:

• select an Input (a Physio Bank record),

• set any relevant Output options,

• choose a tool from the Toolbox, and

• move around within the record you have chosen with the Navigation buttons.

2.3 Inputs Of Tool:

Physio Bank's collections are organized into more than 50 databases, each containing a number of records, and each record containing information collected from a single subject.

(ref: https://archive.physionet.org/physiobank/physiobank-intro.shtml)

Some databases have record sets. If you have selected one of these, two record menus appear. Choose a record set from the menu on the left, then a record within the set using the menu on the right.

2.4 Output of Physio Bank:

Choices in this section affect the output produced by some of the tools, but not all of them.

1.Length: The duration of the observation window within the input record like 10 seconds, one minute , one hour, etc. (The Navigation buttons at the right define the location of the window within the record.)

2.Time format: Time format is made up of time/date, elapsed time, hours, minutes, seconds and samples.

3.Data format: How sample values are given as output like standard, high precision and raw ADC units.

3 Training and Testing

We will train and test the dataset using Pandas. Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for

manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance productivity for users.

We split the dataset into two sets:

1. Training Set: Used to train the model (70-80 percentage of original dataset)

2. Testing Set: Used to get an unbiased estim ate of the model performance (20-30 percentage of original dataset)

In Python, there are two common ways to split a pandas DataFrame into a training set and testing set:

Method 1: Use train_test_split() from sklearn
Reference Code:
from sklearn.model_selection import train_test_split
train, test = train_test_split(df, test_size=0.2, random_state=0)

Method 2: Use sample() from pandas
Reference Code:
train = df.sample(frac=0.8,random_state=0)
test = df.drop(train.index)

Packages
1) pandas
2) numpy
Creation of Dataframe
create DataFrame with size = 1,000 rows and 3 columns
df = pd.DataFrame('x1': np.random.randint(30, size=1000),'x2': np.random.randint(12, size=1000), 'y': np.random.randint(2, size=1000))

4.1 Support-vector machine

The SVM algorithm is an algorithm which is used for splitting of data by using a hyperplane ie simply a line diving the plane into two parts equally.The hyperplane is used for classification regression etc or other tasks like outliker detection. A good separation is allowed by hyperplane that does the largest distance to the nearest training data point.such as the line is named as functional margin.

Wtx-pb=0

Features

An SVM training algorithm creates a model that categorizes fresh samples to one of two categories based on a collection of training examples, making it a

non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).SVM assigns training samples to spatial coordinates in order to maximize the distance between the two categories. Then, based on which side of the gap they fall, new samples are projected into that same area and predicted to belong to a category.

SVMs may effectively do non-linear classification in addition to linear classification by implicitly translating their inputs into high-dimensional feature spaces. This technique is known as the kernel trick.

5.1 Decision Tree algorithm

About:A decision tree algorithm is a two step process which includes learning step and prediction step.The Learning step is defined by developed based upon the giving training data.The training step model is used to predict the given data. IG=E(parent)-sigma(wi(E(chi)) The main goal of decision tree is to create a training model can use to predict the class or value of target variable by learning simple decision rules.

Features

There are three different nodes in the decision tree algorithm root Node ,splitting Node,Decision Node