Description of the Selected Classification Problem

**Classification problem**: evaluate a patient’s medical history in order to predict the likelihood the patient currently has or is likely to contract heart disease.

Heart disease is the most common cause of death in the world. The ability to forecast the potential of an individual have or contracting heart disease could be a valuable resource for physicians, hospitals and patients.

This is classification forecasting problem. The final result should be able to accurately classify a patient as either having (or at risk of heart disease) or not having (or at risk of) heart disease base on a give set of independent cardiac related medical attributes.

Description of Available Data

The data set being used to build a forecasting algorithm was provided in a Kaggle competition posting [https://www.kaggle.com/fedesoriano/heart-failure-prediction]. The dataset is made of 920 individual patient medical information. Each patient in the data set is described by two background attributes, nine medical attributes with some relationship to heart failure and a single classification attribute denoting whether or not the patient has heart disease. All data is being used to complete this assignment.

**Background Attributes**

The background attributes for this data set are age and sex. These attributes are considered background attributes because they do not relate directly to medical conditions but are clinically relevant when determining the likelihood a patient has heart disease.

**Medical Attributes**

Medical attributes are attributes which describe medical conditions which have been shown to have a relationship with heart disease in patients. These attributes include chest pain type, resting blood pressure, cholesterol levels, fasting blood pressure, resting ECG, maximum heart rate, exercise angina, exercise relative to rest (old peak) and exercise induced increments in heart rate (ST slope). These attributes are used in the dataset because individually each of them is often found in patients with heart disease, however patients present with unique variations and levels of each of these conditions.

**Numerical Attributes**

Resting blood pressure, cholesterol levels, fasting blood pressure, maximum heart rate and old peak attributes are given as numerical values. After examining each attribute no large outliers or excessively large or small values were found. Therefore the data did not need to be normalized prior to being used in the three forecasting algorithms.

**Ordinal Attributes**

Sex, chest pain type, resting ECG, exercise angina and ST slope are all given in the data set as ordinal attributes. Prior to being used in a forecasting algorithm the data was vectorized. The table below details the results of the vectorizing process.

Table 1. Vectorized Data from Given Ordinal Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sex | Vectorized | Chest Pain Type | Vectorized | Resting ECG | Vectorized | ST Slope | Vectorized | Exercise Angina | Vectorized |
| M | 0 | ATA | 0 | Normal | 0 | Down | 0 | N | 0 |
| F | 1 | NAP | 1 | ST | 1 | Flat | 1 | Y | 1 |
|  |  | ASY | 2 | LVH | 2 | Up | 2 |  |  |
|  |  | TA | 3 |  |  |  |  |  |  |

**Classification Attribute**

The final attribute is a heart disease attribute. This attribute represents the dependent variable in this forecasting problem. In the given data set the attribute is given as a binary value. The value is either zero, the patient does not have heart disease or one, the patient has heart disease.

Overview of Network Architecture

* Selected network
  + Feed forward neural network
* Other networks and why they were not selected
  + CNN – not images

A convolutional neural network is a deep learning algorithm derived from the feedforward neural network. However instead of being fully connected the CNN uses a convolution filter and only connected data points that are in close proximity to one another. The network functions by repeating the pattern of convolution filtering, ReLU and pooling.

A CNN was not selected as the deep learning architecture for this assignment because the data set selected is not given in an image format.

* + RNN – not time series

A recurrent neural network is a deep learning algorithm base on the feed forward neural network that adds an extra dimension (time) into the network. The connection between neutrons span adjacent time steps.

A RNN was not selected as the deep learning architecture for this assignment because the data set is independent of time.

* + Autoencoder – not large enough sample size

Autoencoders are fully connected deep learning algorithms that have a unique hidden layer structure. With each successive layer of an autoencoder in the first half of the network the number of neurons per layer decrease. The mid point of the hidden layers in the network has the lowest numbers of neurons. This point is, the bottleneck, is the point where the inputs is compressed. The layers after the bottle neck mirrors the number of neurons in the first half. The final output of the network is the same as the input values.

An auto encoder was not selected as the deep learning algorithm used in this assignment because autoencoders should be trained with normal data and data with anomalies will be detected as part of the test set. The given data in roughly an even split of patients with heart disease and without and would not be enough data to train a deep learning network.

* Important parameters
* Tuning parameters

Process

* Preprocessing / Feature generation

The only preprocessing required prior to training the feed forward neural network was to convert all data types represented as ordinal data into numerical data, as described in the *Description of Available Data* section. The categories of sex and exercise angina have binary value options therefore they were vectorized as 1 or 0. The categories of chest pain type and resting ECG have independent value options, therefore they were vectorized from 0-3 and 0-2 respective, in the order they first appeared in the data set. The category ST slope contains three value options that are associated to each other by different amounts, i.e. an increasing (up) ST slope is associated closer to a flat ST slope then a decreasing (down) ST slope. Similar to how when vectorizing sales for a business, the month April and May are more closely related than April and October. Due to the association the ST slope was vectorized in the specific sequence, down, flat , up and labeled 0,1,2 respectively.

No feature generation was done to the data set due to the fact that this data set is not a time series, no new pertinent attributes could be extracted.

* Model training/testing
* Evaluation

Results/Tuning