Heart Disease Prediction using Machine Learning

Daniel Mendoza, Wei-Hsu Chao and Tsun-Han Huang
Stanford University

Problem Definition and Motivation

Heart disease is an umbrella of medical conditions including blood vessel diseases, coronary artery disease, heart rhythm problems, and heart defects [1]. Proper treatment for this disease is imperative as the condition is detrimental to one's health. With recent strong effort to develop machine learning techniques to make inferences on data, researchers have been attempting to apply state of the art techniques for medical diagnosis [2, 3]. The Center for Machine Learning and Intelligent Systems at the University of California, Irvine has made public a credible data set containing labeled medical data for the presence of heart disease [4]. This project aims to use this dataset to build machine learning models and test the prediction accuracy of heart disease.

Challenges

We map the problem of predicting the presence of heart disease to a classic problem in machine learning called binary classification. With high dimension of inputs, we cannot identify an obvious decision boundary by inspection of the data.
The training data for our supervised models is limited to a small dataset of 303 examples. Fitting machine learning models to small datasets are likely to perform well on the small dataset, but their performance may not perform well in a more general setting

Approach

This project aims to develop a prediction model for heart disease with high accuracy.
Dataset contains 303 examples each with 14 features:

- Age
- Sex
- Chest pain type
- Resting blood pressure
- Serum cholestoral
- Fasting blood sugar
- Resting electrocardiographic results
- Maximum heart rate achieved
- Exercise induced angina
- ST depression induced by exercise relative to rest
- Slope of the peak exercise ST segment
- Number of major vessels
- Thalium
- Presence of heart disease

We use the presence of heart disease as the ground truth label in supervised learning algorithms.

Machine Learning Models:
Baseline: Perceptron
- 0-1 loss, learning rate=0.1
- Limited to linear decision boundary

K Nearest Neighbors (KNN)
- K = 11
- Non-linear and non-parametric
- Limited by training data
- Highly susceptible to poor performance for small training datasets

Kernel SVM
- RBF kernel, C = 1, gamma = 0.01
- Implicit regularized loss and convex optimization
- Non-linear and non-parametric
- Limited to decision boundary produced by kernel

Deep Neural Networks (DNN), illustrated in Figure 1
- 6 different architectures -
  - Two hidden layers: (13-13), (32-16), (64-32) neurons
  - Three hidden layers: (13-13-13), (32-16-8), (64-32-16) neurons
- Output Activation: Sigmoid, Hidden Layer Activation: Relu
- Learning Rate: 0.001, Batch Size: 64, Loss: Binary Cross entropy
- Non-linear and parametric
- Prone to local optima

Each model trained on 70% of the data and tested on 30%

	Perceptron	KNN	Kernel SVM
Accuracy	80%	80%	82%
Precision	77%	76%	77%
Recall	87%	89%	94%
F1_Score	82%	82%	85%

Chart 1. Results of Perceptron, KNN, and SVM

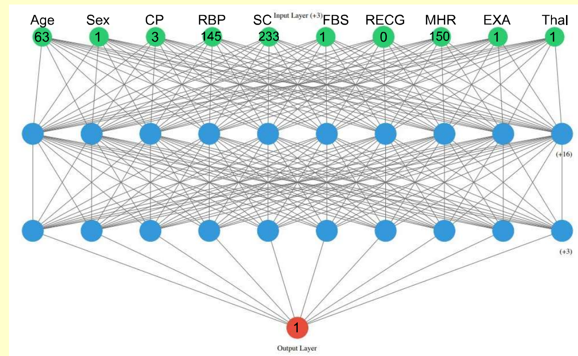

Figure 1. Neural Network with Input Layer (Green), Two Hidden Layers (Blue), and Output Layer (Red). Exemplifies test example and outputs '1', which indicates the predictor has diagnosed heart disease.

	Accuracy	Precision	Recall	F1 score
13 / 13	83.61%	83.33%	88.24%	85.71%
32 / 16	85.25%	83.78%	91.18%	87.32%
64 / 32	88.52%	88.57%	91.18%	89.86%

Chart 2. Results of DNN with Two Hidden Layers

	Accuracy	Precision	Recall	F1 score
13 / 13 / 13	83.61%	81.58%	91.18%	86.11%
32 / 16 / 8	86.89%	86.11%	91.18%	88.57%
64 / 32 / 16	86.89%	84.21%	94.12%	88.88%

Chart 3. Results of DNN with Three Hidden Layers

Analysis

Initial Data Analysis Prior to Experimentation:
- Comparatively small dataset [2]
- Unbalanced amount of female and male examples
  - 207 males and 96 females
  - Dataset may not be representative of all heart disease patients

An Interpretation of Precision, Recall:
- Precision can be interpreted as the likelihood of heart disease given that the predictor has diagnosed heart disease.
- Recall can be interpreted as the likelihood that the predictor correctly identifies a patient with heart disease given he/she actually has heart disease.

Baseline Perceptron Results Analysis:
- Chart 1 shows 80% accuracy, data is not perfectly linearly separable
- More complex decision boundaries are likely needed to improve prediction accuracy.

KNN Results Analysis:
- Chart 1 shows similar performance to perceptron despite non-linear decision boundary
- KNN often cannot perform well in small datasets

Kernel SVM Results Analysis:
- Chart 1 shows improved accuracy and recall against Perceptron and SVM model
- Suggests that non-linear decision boundaries are promising

DNN Results Analysis:
- Charts 2 and 3 show improvement in all metrics with DNN against Perceptron, KNN, and Kernel SVM.
- DNN with two hidden layers of (64-32) neurons exemplifies peak accuracy of 88.52% and F1 score of 89.86%
- Simultaneously achieving highest accuracy and F1 score exemplifies most robust predictions out of all the results

Concluding Remarks:
- DNNs promising avenue for expressive and robust heart disease prediction models
- Results show that each model exemplifies high recall (> ~80%), especially for models with high accuracy. High recall is significant for diagnosis so that patients who have heart disease get the treatment they need.

References

1. "Heart disease," Mayo Clinic, 22-Mar-2018. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118. [Accessed: 30-Nov-2019].
2. S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," Plos One, vol. 12, no. 4, Apr. 2017.
3. A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, " Heart Disease Data Set," UCI Machine Learning Repository.