Machine Learning Final Project Report

Team member:

Yibin Zheng(yz6118)

Matianqi Song(ms11679)

Project introduction

This machine learning project is about evaluating and predicting the heart disease condition of a patient. The team retrieved the data set from UCI machine learning repository, and the website URL is https://archive.ics.uci.edu/ml/datasets/Heart+Disease. After evaluating these attributes mentioned after this paragraph, the team can get the result of the person's heart disease level in terms of its angiographic status. The result would be 0 if the angiographic diameter is narrowing less that 50%, and the result would be 1 if the angiographic diameter is narrowing greater than 50%. The team compared different machine learning techniques while realizing the goal of their project such as Clustering, PCA, Logistic Regression and Neural Network.

Attribute Information:

- age
- sex
- cp: chest pain type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic
- trestbps: resting blood pressure (in mm Hg)
- chol: serum cholestoral (in mg/dl)
- fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- restecg: resting electrocardiographic results
    - Value 0: normal
    - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach: maximum heart rate achieved
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment
    - Value 1: upsloping
    - Value 2: flat
    - Value 3: downsloping
- ca: number of major vessels (0-3) colored by flourosopy
- thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Initialization:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.000000	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.000000	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.000000	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.000000	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.000000	3.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	45.0	1.0	1.0	110.0	264.0	0.0	0.0	132.0	0.0	1.2	2.0	0.000000	7.0	1
299	68.0	1.0	4.0	144.0	193.0	1.0	0.0	141.0	0.0	3.4	2.0	2.000000	7.0	2
300	57.0	1.0	4.0	130.0	131.0	0.0	0.0	115.0	1.0	1.2	2.0	1.000000	7.0	3
301	57.0	0.0	2.0	130.0	236.0	0.0	2.0	174.0	0.0	0.0	2.0	1.000000	3.0	1
302	38.0	1.0	3.0	138.0	175.0	0.0	0.0	173.0	0.0	0.0	1.0	0.672241	3.0	0

303 rows × 14 columns

Figure 1. Patients Data

After reading the data set from UCI machine learning repository, the data of 303 patients with 14 attributes are imported. The “num” column represents the heart disease level of the patient.

```
[0. 1. 1. 0. 0. 0. 1. 0. 1. 1. 0. 0. 1. 0. 0. 0. 1. 0. 0. 0. 0. 0. 1. 1.
 1. 0. 0. 0. 0. 1. 0. 1. 1. 0. 0. 0. 1. 1. 1. 0. 1. 0. 0. 0. 1. 1. 0. 1.
 0. 0. 0. 0. 1. 0. 1. 1. 1. 1. 0. 0. 1. 0. 1. 0. 1. 1. 1. 0. 1. 1. 0. 1.
 1. 1. 1. 0. 1. 0. 0. 1. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 1.
 1. 1. 0. 0. 0. 0. 0. 0. 1. 0. 1. 1. 1. 1. 1. 1. 0. 1. 1. 0. 0. 0. 1. 1.
 1. 1. 0. 1. 1. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 0. 0. 1. 0. 1.
 0. 1. 1. 0. 0. 0. 0. 0. 0. 1. 1. 1. 1. 1. 1. 0. 0. 1. 0. 0. 0. 0. 0. 0.
 1. 0. 1. 0. 1. 0. 1. 1. 0. 1. 0. 0. 1. 1. 0. 0. 1. 0. 0. 1. 1. 1. 0. 1.
 1. 1. 0. 1. 0. 0. 0. 1. 0. 0. 0. 0. 0. 1. 1. 1. 0. 1. 0. 1. 0. 1. 1. 0.
 0. 0. 0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 1. 1. 0. 1. 1. 0. 0. 1. 1. 1. 0. 0.
 0. 0. 0. 1. 0. 1. 1. 1. 1. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1. 0. 0.
 1. 1. 1. 1. 1. 0. 1. 0. 1. 0. 1. 0. 0. 0. 1. 0. 1. 0. 1. 0. 1. 1. 1. 0.
 0. 0. 1. 0. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 0.]
```

Figure 2. Patient Heart Disease Level

While evaluating the patient data, the team can get the patient heart disease level of each patient. “0” means that the angiographic diameter is narrowing less that 50%, and “1” means that the angiographic diameter is narrowing greater than 50%.

Method 1: Clustering

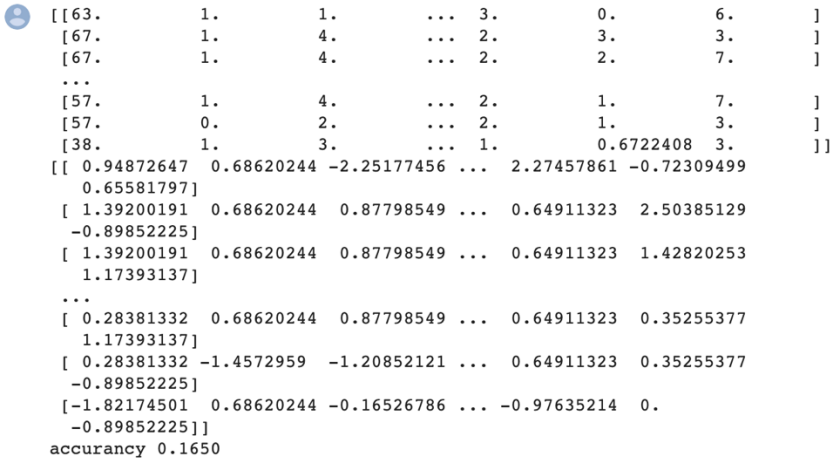

Figure 3. Result of Clustering

The result of clustering shows that the accuracy is only 0.165, so that clustering method is not suitable for this project.

Method 2: PCA

The team construct the PCA object and then fit and transform the data, and finally get the plotting graph.

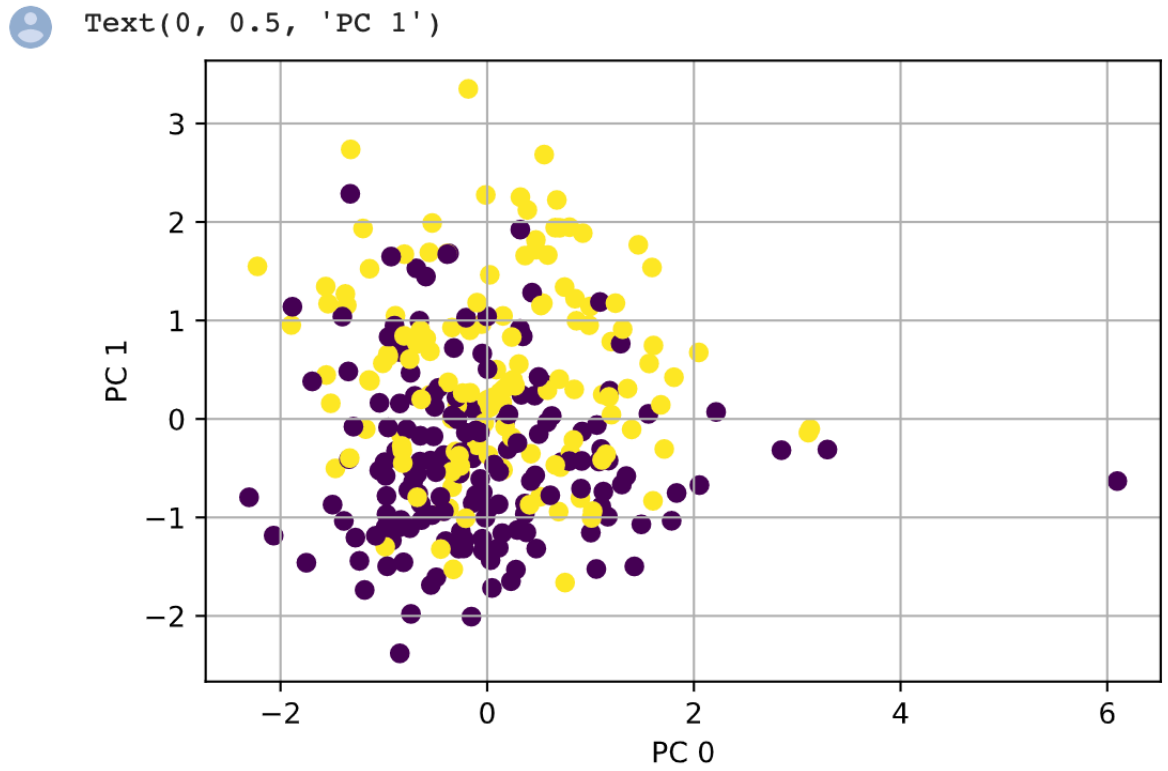

Figure 4. PCA graph

The yellow points represent the disease level 0 and purple points represents the disease level 1. While analyzing the graph, we can clearly see that yellow points and purple points cannot be separated by two groups but mix together, which shows that clustering method is not suitable for this project, and SVM is not suitable for this project either.

Method 3: Logistic Regression

The team uses 10-fold cross validation to get the more accurate result.

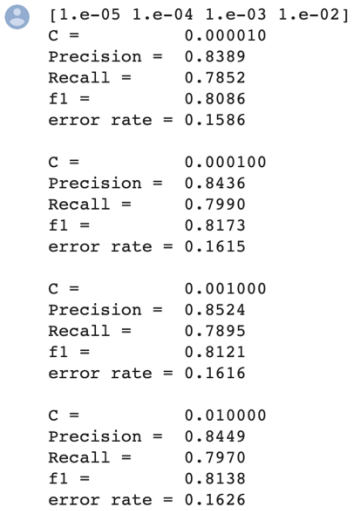

Figure 5. Logistic Regression Result

After analyzing the result, the team found out that all model's precision rates are around 85%, which means, logistic regression does have a better performance than cluster. The model of this project can be trained well with logistic regression technique.

Method 4: Neural Network

For neural network, the team creates the model and optimizer, builds a deeper network and fits the model for 40 epochs. The team tries three learning rate 0.02, 0.01 and 0.005. The accuracy of each epoch is calculated, and the team find that the accuracy rates for 40 epochs are around 80%, which is not accurate enough for neural network. The team also generates the loss graph and accuracy graph for neural network while training the model.

```
Epoch 12/40
272/272 [==============================] - 0s 33us/sample - loss: 0.3590 - acc: 0.8456 - val_loss: 0.4173 - val_acc: 0.8387
Epoch 13/40
272/272 [==============================] - 0s 26us/sample - loss: 0.3681 - acc: 0.8419 - val_loss: 0.4511 - val_acc: 0.7742
Epoch 14/40
272/272 [==============================] - 0s 29us/sample - loss: 0.3586 - acc: 0.8566 - val_loss: 0.4082 - val_acc: 0.8387
Epoch 15/40
272/272 [==============================] - 0s 33us/sample - loss: 0.3499 - acc: 0.8493 - val_loss: 0.3954 - val_acc: 0.8387
Epoch 16/40
272/272 [==============================] - 0s 29us/sample - loss: 0.3524 - acc: 0.8493 - val_loss: 0.4036 - val_acc: 0.8387
Epoch 17/40
272/272 [==============================] - 0s 29us/sample - loss: 0.3478 - acc: 0.8566 - val_loss: 0.4207 - val_acc: 0.8387
Epoch 18/40
272/272 [==============================] - 0s 33us/sample - loss: 0.3480 - acc: 0.8566 - val_loss: 0.4319 - val_acc: 0.8065
Epoch 19/40
272/272 [==============================] - 0s 33us/sample - loss: 0.3458 - acc: 0.8566 - val_loss: 0.4171 - val_acc: 0.8387
Epoch 20/40
272/272 [==============================] - 0s 26us/sample - loss: 0.3452 - acc: 0.8529 - val_loss: 0.4141 - val_acc: 0.8387
Epoch 21/40
272/272 [==============================] - 0s 33us/sample - loss: 0.3448 - acc: 0.8603 - val_loss: 0.4236 - val_acc: 0.8065
Epoch 22/40
272/272 [==============================] - 0s 26us/sample - loss: 0.3443 - acc: 0.8493 - val_loss: 0.4323 - val_acc: 0.8065
```

Figure 6. Accuracy of Neural Network model

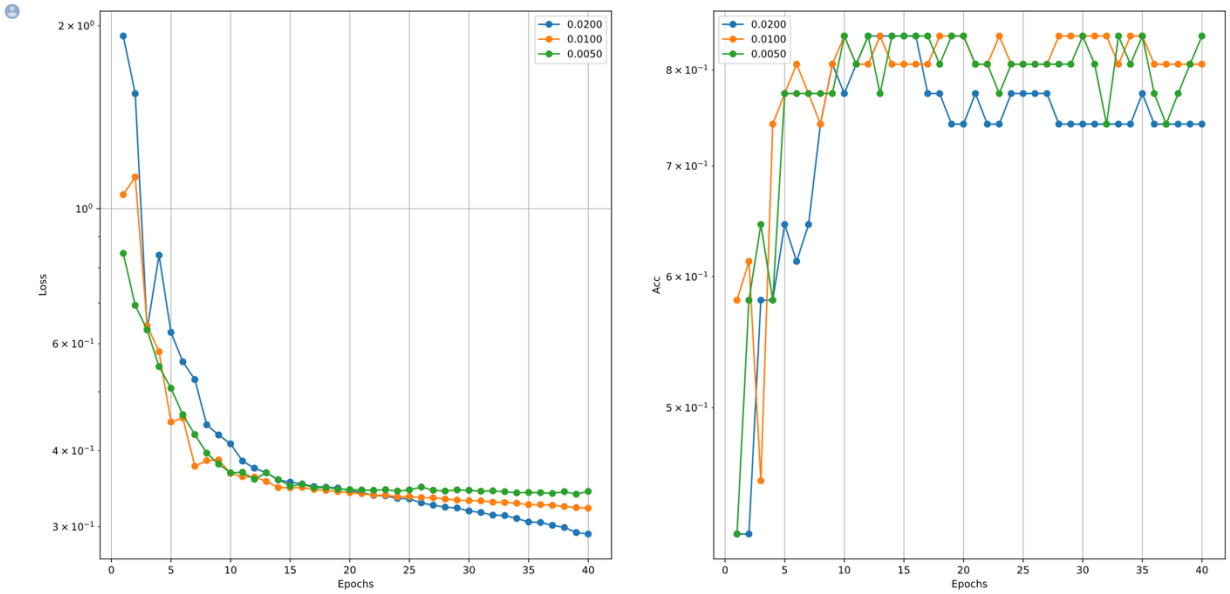

Figure 7. Loss Graph and Accuracy Graph for Neural Network

Data Source:

http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

http://archive.ics.uci.edu/ml/datasets/Heart+Disease Source:

Creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D.,

Ph.D. Donor: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779