DS 210 Final Project Writeup

Collaborators: None
Sources: Rust textbook, class lectures, clustering - Rust (docs.rs), Complete Guide To Testing Code In Rust | Zero To Mastery, Reading CSV Files in Rust. In this article we will explore how to… | by Emanuelmechie | Medium, Averages in Rust (benjaminbrandt.com)

Git Hub Link: aakashd25/project1 (github.com)

Project Overview

The project is based on a "Heart Disease" data set which had piqued my interest as a pre medical student. Here is the URL to the data set:Heart disease (kaggle.com). The dataset is rated a 10/10 on usability, and is liscened by MIT. Along with that this dataset is updated quarterly so that data seems to be up to date. The dataset contains 14 attributes including symptoms of the patients who are diagnosed with a heart disease. The goal for this project was to create a clustering algorithm to partion the data. I wanted to take the network of patients and identify which symptoms were the most meaninful by comparing the impact each had on the diagnosis the patients. After that, I wanted to find the k best representitves of the clusters and see how it compares to the common knowledge about symptoms related to heart disease.

Project Code

The final project consists of 6 main parts: Organizing and reading the csv file, calculating metrics, clustering the patients, selecting the best representitives from each cluster, outputing the code, and creating a test module.

1. Reading and organizing the CSV file

This was the first step of the project, and it consisted of creating a struct which is a breakdown of each patient which holds all of the symptom characteristcs and the diagnosis (the last coloumn). A function was created to read the data set and parse through each of the symptoms in order to put each patient into a vector type. Then another funciton was created in order to split the data into two groups: the first is patients that are diagnosed with heart disease and other other is patients are not diagnosed with the disease. This was done by creating a loop which analyzes each patient's diagnosis and finds the binary condition of heart disease.

2. Calculating Metrics

Now that we have two groups of patients, I wanted to get some background metrics on this data. A good metric for this data are the median values for all the symptoms in each group. This was calculated by sorting the values and iterating through each symptom and then calculating the median by finding the middle value for each coloumn. A different method was used to calculate depending on whether there were an odd or even number of elements.

3. Clustering algorithm - In Cluster Module

The next part involved creating a clustering algorithm for the data set. First the clusters were intialized with random centroids within the range of data values in the dataset. Then after that was collected, the program iterates until convergence or the maximum number of iterations are reached. Each patient is then assigned to the nearest cluster and the cluster centroids are updated. The centroids are then chosen by iterating through the clusters with the maximum number of members. The algorithm for eassigning each patient to a cluster involves measuring a distance metric in which a euclidean distance function was used.

4. Find Best Representatives

The program then contains another function which finds the k best representatives of each of the clusters. In order to do this, a new vector for each of the best representatives is created and then there is a for loop that goes through each point in each cluster and finds the point (patient) that has the shortest distance from the centroid.

5. The Main Function

To put the project together, the main function contains all of the code that produces an output. First the heart_disease.csv file is loaded and prepared, and the primary analysis is completed by splitting the data into two groups. One of the groups corresponds to patients with Heart Disease and the other corresponds to patients without Heart Disease. The diagnosis of heart disease is given as the last value in each patient vector as a binary condition: 1 = has the disease and 0 = does not have disease. Next the median symptom values for each group was printed. Then the clustering settings were created and set to k=2 cluster in order to see if we can target the differences in symptom values that differentiate patient diagnosis. After that, the centroid values are printed for each cluster. Next, the best representative patients of each cluster with its respective diagnosis value.

6. Test Code

Lastly a test module was created in order to test the functions of the program. A small test data set was created and each of the functions used in the main program were tested. Assert_eq was used to confirm that the test ran correctly for all of the functions.

Running the code:

In order to run the program, first you must run the code by clicking the run code button. Then type cargo run in the terminal. This should provide you with an output. In order to test the code type cargo test in the terminal. This will provide the output from the test module.

Program Output:

Column Headings:
rest_bp, chest_pain, thalassemia, age, fasting_bs, max_hr, exercise_angina, gender, st_slope, cholesterol, st_depression, rest_ecg, num_vessels, diagnosis

Median Symptoms for Patients Diagnosed with Heart Disease: [130.0, 3.0, 2.0, 58.0, 0.0, 142.0, 1.0, 1.0, 1.0, 253.0, 1.4, 2.0, 1.0]
Median Symptoms for Patients Not Diagnosed with Heart Disease: [130.0, 2.0, 0.0, 52.0, 0.0, 161.0, 0.0, 1.0, 0.0, 235.5, 0.2, 0.0, 0.0]

Centroid 1: [129.64516129032256, 2.1075268817204287, 0.8010752688172038, 53.263440860215, 0.1505376344086021, 151.7634408602151, 0.28494623655913964, 0.7365591397849457, 0.5967741935483867, 216.54838709677418, 1.0021505376344082, 0.8924731182795693, 0.5860215053763437]
Centroid 2: [135.12612612612617, 2.243243243243241, 0.8918918918918926, 56.68468468468468, 0.13513513513513514, 145.97297297297294, 0.3963963963963966, 0.5765765765765771, 0.6126126126126131, 298.96396396396403, 1.1450450450450447, 1.1711711711711723, 0.8288288288288289]

Best Representatives:
Cluster 1: ([128.0, 2.0, 2.0, 57.0, 0.0, 150.0, 0.0, 1.0, 1.0, 229.0, 0.4, 2.0, 1.0], 1)
Cluster 2: ([140.0, 1.0, 0.0, 56.0, 0.0, 153.0, 0.0, 0.0, 1.0, 294.0, 1.3, 2.0, 0.0], 0)

The output seemed interesting to me as based on the centroids from the clusters the resting blood pressure is greater in those who have a higher risk of heart disease which matches my previous knowledge. Other variables that make sense with previous understanding of the heart condition is the increased chest pain, and cholesterol. But it was also odd that when the best representative patients were taken from each cluster the opposite effects were seen which could indicate an outlier patient which contains symptoms of a usual patient with heart disease but is diagnosed with heart disease.