Using K-Nearest Neighbors to Predict Heart Disease, Sex

CS 5830 Project 4 Kaden Hart & Chase Mortensen

Introduction

The purpose of our analysis is to use health data, such as resting blood pressure and cholesterol to predict whether or not a patient has heart disease. Additionally, we were tasked with implementing a K-nearest neighbors (knn) model and "[determining] optimal values of k and an optimal set of attributes to use to maximize predictive power."

In addition to implementing our own knn model, we retrieved data from ics.uci.edu and implemented a model to predict an individual's sex based on age, education, and hours worked per week.

Presentation Slides:
https://docs.google.com/presentation/d/1f-7o2MJtkJ304B8kVqGl0zeaD0cV-jHpffQIOKLV4LE/edit?usp=sharing

Datasets

The dataset used to make predictions for heart disease was provided on the class canvas page and was retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease. The dataset consists of 14 attributes, including age, sex, chest pain, blood pressure, cholesterol, blood sugar, electrocardiographic results, heart rate, and exercise-induced chest pain. There are 303 data points in the dataset. The health dataset was in a CSV format and included a few missing values.

The dataset used to predict male or female was found at https://archive.ics.uci.edu/ml/datasets/Adult. This dataset included 14 attributes: age, workclass, fnlwgt, education, education number, marital status, occupation, race, sex, capital gain, capital loss, hours per week, and native country. The dataset included 48,842 adults, after removing any row with a question mark or nan, there were 32,561 adults left. The attributes used to predict sex were education number, age, and hours per week. This data was from 1996.

Analysis Techniques

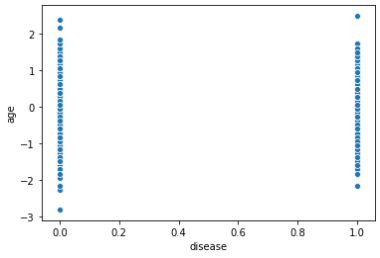
For part one, we used scatterplots and Pearson correlations to explore the data and then, as stated in the introduction, built a knn classifier. We used the `GridSearchCV` tool from sklearn.model_selection to perform 10-fold cross-validation on our classifiers and analyzed the results using `precision_recall_fscore_support` and `classification_report` from sklearn.metrics.

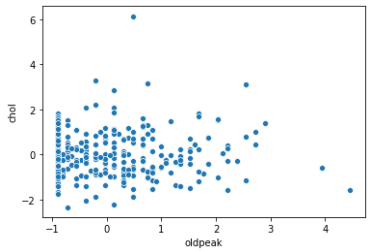
Pearson correlations made sense in this case because we were trying to predict whether or not an individual had heart disease. We were able to generate the table below, which allowed us to find the data features that were most highly correlated with ‘disease’ and modify our classifiers based on that information.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	disease
age	1.000000	-0.09754	0.104139	0.284946	0.208950	0.118530	0.148868	-0.39380	0.091661	0.203805	0.161770	0.223120
sex	-0.09754	1.000000	0.010084	-0.06445	-0.19991	0.047862	0.021647	-0.04866	0.146201	0.102173	0.037533	0.276816
cp	0.104139	0.010084	1.000000	-0.03607	0.072319	-0.03997	0.067505	-0.33442	0.384060	0.202277	0.152050	0.414446
trestbps	0.284946	-0.06445	-0.03607	1.000000	0.130120	0.175340	0.146560	-0.04535	0.064762	0.189171	0.117382	0.150825
chol	0.208950	-0.19991	0.072319	0.130120	1.000000	0.009841	0.171043	-0.00343	0.061310	0.046564	-0.00406	0.085164
fbs	0.118530	0.047862	-0.03997	0.175340	0.009841	1.000000	0.069564	-0.00785	0.025665	0.005747	0.059894	0.025264
restecg	0.148868	0.021647	0.067505	0.146560	0.171043	0.069564	1.000000	-0.08338	0.084867	0.114133	0.133946	0.169202
thalach	-0.39380	-0.04866	-0.33442	-0.04535	-0.00343	-0.00785	-0.08338	1.000000	-0.37810	-0.34308	-0.38560	-0.41716
exang	0.091661	0.146201	0.384060	0.064762	0.061310	0.025665	0.084867	-0.37810	1.000000	0.288223	0.257748	0.431894
oldpeak	0.203805	0.102173	0.202277	0.189171	0.046564	0.005747	0.114133	-0.34308	0.288223	1.000000	0.577537	0.424510
slope	0.161770	0.037533	0.152050	0.117382	-0.00406	0.059894	0.133946	-0.38560	0.257748	0.577537	1.000000	0.339213
disease	0.223120	0.276816	0.414446	0.150825	0.085164	0.025264	0.169202	-0.41716	0.431894	0.424510	0.339213	1.000000

Additionally, we used scatterplots in a similar way to familiarize ourselves with the data and to see trends. The Pearson correlations between disease and other features were generally more useful than the scatterplots of the same data. Below to the left is an example of a generated scatterplot between ‘disease’ and ‘age.’ Clearly, the correlation coefficient was more useful.

.





However, other plots, such as the plot between cholesterol and ‘oldpeak’ (ST depression induced by exercise relative to rest) was more interesting, although less useful for our goal of predicting heart disease.

Additionally, we standardized our data so that no feature would heavily outweigh the others in our classifier when calculating the distance between points.

`GridSearchCV` was useful for testing our data because it provided an easy interface for cross-validating our models.

For part two, we tested combinations of the three attributes and varied the value of k on a very small set of 488 people. We then took the most consistent region and increased the sample size to 4,884 people, which is where we found our highest f score.

Results

Part 1

Initially, our knn classifier had accuracy in the mid 80s (it varied each run, but was 0.88 in the latest run). Based on the Pearson correlations, we removed cholesterol and fasting blood sugar features from our classifier since these features were the least correlated to disease and were likely adding noise and worsening the curse of dimensionality. This version of our classifier improved to approximately 0.90. Removing two additional columns, ‘trestbps’ and ‘restecg’ added slight improvement to about 0.91.

We tested different values of *k* and found interesting results. While our accuracies in CV-10 generally improved with *k* values around 11-15, our highest F-score for new data was often with *k*=1. This suggests that our model with *k*=1 generalizes better than other models. We suspect that this may have to do with data sparsity and that a higher *k* value might work better provided more data.

Part 2

The results for part 2 show the process of trying different attributes and showing how changing k values affects the f score. When the two attributes with the most consistency were found to be working hours and age, the test group was increased in size, and k values were tested over a smaller range. This change resulted in higher consistency between k values so results error could be reasonably estimated. The highest F score for females was 0.72. The highest F score for males was 0.54.