Final Report: COVID-19 Diagnoser

Oliver Stein
Department of Computer Science
Virginia Tech

Diego Espinoza
Department of Computer Science
Virginia Tech

William Pryor
Department of Computer Science
Virginia Tech

Nebiyu Elias
Department of Computer Science
Virginia Tech

Abstract

Our project is attempting to recreate and improve a paper[1] that uses machine learning to determine whether a person with a certain set of symptoms and pre-existing conditions has COVID-19. The paper we are attempting to recreate uses logistic regression and xgboost models, but our model will use weighted k-nearest neighbor (KNN) to classify data points, utilizing scikit learn's knn algorithm along with our own adapted weighting procedure.

1 Methodology

We are utilizing Jupyter Notebook in Python to collect, sort, manipulate, and graph data found from a CDC report in July. The initial steps in the Jupyter Notebook are similar to HW 2 regarding setting up the plotting mechanism and graphs. We are importing discrete data values including 5 symptoms and 5 pre-existing conditions - from which we calculate a score for each (symptom score and pre-existing condition score) and plot them on 2D graph, and use the Euclidean distance formula to operate our KNN classifier[Fig. 2]. From there, we can move on to the actual weighted KNN algorithm, which gives more emphasis (or "weight") to the features more highly correlated with a specific COVID-19 outcome.

We split our imported data up into 2 sections: 80% of the data will go to "training" (just plotting the labeled points on the graph in order to operate with KNN), and 20% for validation to determine how well our algorithm is classifying. Validation is used for our error calculation in this case because there is no traditional loss function for KNN.

Because our values that we are importing from our dataset are entirely discrete, we need to compute a score for plotting that belongs to a higher-dimensional range. We plotted our values on a graph labeled "symptom score" on the x-axis and "pre-existing condition score" on the y-axis, each of which is a non-discrete value calculated from 5 discrete symptom/pre-existing condition data points in conjunction with the feature's respective weight.

The weight calculation was modelled based on a common DNA parsing scoring matrix algorithm which computes weights for nucleotides based on their frequency of correlation with other similar sequences.

2 Algorithm Choice

We are going to use a weighted KNN algorithm to ensure that the prediction of whether the patient has COVID-19 is biased towards more consistent groupings of data points. This is important because it will prevent a data point's classification from being affected as much by erratic data points, which are caused by people with odd combinations of symptoms, as well as accounting for bad or inconsistent

data collection by hospital workers and researchers. It also emphasizes features that are more strongly correlated with the diagnosis. We can create the weights using a scoring matrix similar to the bioinformatic BLASTP scoring algorithm which calculates the log-based frequency of any given occurrence in the data [Fig 3].

3 Applications

Knowing whether a person is likely to have COVID-19 or not is a powerful tool at this time. We can use different data representations to signify how likely the virus is to exist based on different features such as each individual symptom, a common combination of symptoms, or different time periods where some symptoms were more prevalent than others. We can use this to track the virus' development over time, and potentially pinpoint what evolution the virus is at.

This type of symptom and condition classifying has been attempted by websites like WebMD using far more rudimentary techniques - by applying legitimate learning algorithms and scoring mechanisms we can much more accurately deliver diagnostic scores to people at home suffering from diseases/conditions that exceed COVID-19. Many people still refer to WebMD to try to diagnose their illnesses, highlighting the demand for this type of technology.

4 Results

The COVID-19 diagnoser has an accuracy of 54.6%. The calculation for accuracy occurs in the last block of the Jupyter Notebook. The classifier in the paper we attempted to recreate was claimed to be approximately 99%. Our diagnoser used 11 input variables: breathing problems, fever, dry cough, sore throat, running nose, asthma, chronic lung disease, heart disease, diabetes, and hyper tension [Fig. 4]. The classifier in the notebook used 19 input variables [Fig. 1].

5 Conclusion

We created a COVID-19 diagnosing tool that was less accurate than the tool in the paper that we were recreating. However, our tool improved upon the one in the paper in that it did not require as many input variables[Fig 1]. Inputs that were left out included gastrointestinal, contact with COVID patient, abroad travel, and many others. Removing these data points will allow people to get results without being as invasive or asking about things that the user may not know how to answer [Fig. 4]. Compared with the model provided in the paper we recreated, our model traded some accuracy for ease of use.

A potential change that could be made in the future that might increase accuracy is using an algorithm that works better with higher-dimensional data, such as a Naive Bayes classifier. We could also do more research and experimentation to determine which symptoms and conditions have the greatest impact on the diagnosis. This would allow us to maximize the accuracy while minimizing the number of required inputs. Although our agent's accuracy was not as good as the agent we were attempting to recreate, our group is confident that we are able to use machine learning to process real world data and provide valuable analysis of that data.

6 Contributions

Will Pryor: Determined how the data we collected could work with KNN, research ways to weight data points for KNN, help determine which features should be included in our model.

Nebiyu Elias: Developing basic methodology, initial kaggle analysis and developed different ways to tweak the algorithm. Discussed different ways to represent data and expansion possibilities at a later date.

Diego Espinoza: Researched alternative approaches to Covid-19 diagnoser and found the paper (Kaggle notebook) directly related to the implementation in the project. We enhanced the algorithm used in the notebook and made sure our approach produced more accurate results.

Oliver Stein: Determined scoring procedure and implemented weights calculation. Additionally, implemented the testing and individual user algorithm interaction.

Breathing	Fever	Dry Cough	Sore throa	Running N	Asthma	Chronic Lu	Headache	Heart Dise	Diabetes	Hyper Ten	Fatigue	Gastrointe	Abroad tra	Contact w	Attended	Visited Pu	Family wo	Wearing N	Sanitizatic	COVID-19
Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	No	Yes
Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No	Yes	Yes	No	No	No	Yes
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	No	No	No	No	No	Yes
Yes	Yes	Yes	No	No	Yes	No	No	Yes	Yes	No	No	No	Yes	No	Yes	Yes	No	No	No	Yes
Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	No	Yes
Yes	Yes	Yes	No	No	No	No	No	Yes	No	Yes	No	No	No	No	No	No	No	No	No	Yes
Yes	Yes	Yes	No	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No	Yes
Yes	Yes	Yes	No	Yes	Yes	No	No	No	Yes	Yes	No	Yes	Yes	No	No	Yes	No	No	No	Yes
Yes	Yes	Yes	No	Yes	No	Yes	No	No	Yes	No	Yes	No	Yes	Yes	Yes	No	No	No	No	Yes
Yes	Yes	Yes	No	No	Yes	No	No	No	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	Yes
Yes	Yes	Yes	No	No	No	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	No	No	No	Yes
Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	No	No	Yes
Yes	Yes	Yes	Yes	No	Yes	No	No	No	No	Yes	No	Yes	Yes	No	No	Yes	No	No	No	Yes
Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	No	No	Yes
Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes
Yes	Yes	Yes	No	Yes	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	Yes
Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	No	No	No	No	No	No	No	No	Yes
Yes	Yes	Yes	No	Yes	No	No	Yes	Yes	No	No	No	Yes	No	Yes	No	No	Yes	No	No	Yes
Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	No	No	Yes	No	No	No	Yes
Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes	No	No	No	No	No	Yes	No	No	Yes
Yes	Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No	Yes
Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	Yes	Yes	No	Yes	No	No	No	No	No	No	Yes
Yes	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No	No	No	Yes	Yes	No	No	No	No	No	Yes
Yes	Yes	Yes	No	Yes	No	No	Yes	No	No	Yes	No	No	No	No	No	No	Yes	No	No	Yes
Yes	Yes	Yes	No	No	Yes	No	No	Yes	No	No	No	No	Yes	No	No	No	No	No	No	Yes
Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	No	Yes	No	No	Yes
Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	No	No	Yes	No	No	Yes
Yes	Yes	Yes	Yes	No	No	No	Yes	No	Yes	No	No	Yes	No	Yes	No	Yes	No	No	No	Yes
Yes	Yes	Yes	No	No	No	No	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes
Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	No	Yes	No	No	No	Yes	Yes	Yes	No	No	Yes
Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	No	No	Yes	Yes	Yes	No	No	Yes
Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	Yes	No	No	Yes	No	Yes	No	No	No	Yes
Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes	Yes	No	No	No	Yes
Yes	Yes	Yes	Yes	No	No	No	Yes	No	No	No	Yes	Yes	No	No	Yes	No	Yes	No	No	Yes
Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	No	No	Yes

Table 1: Data points used by our model

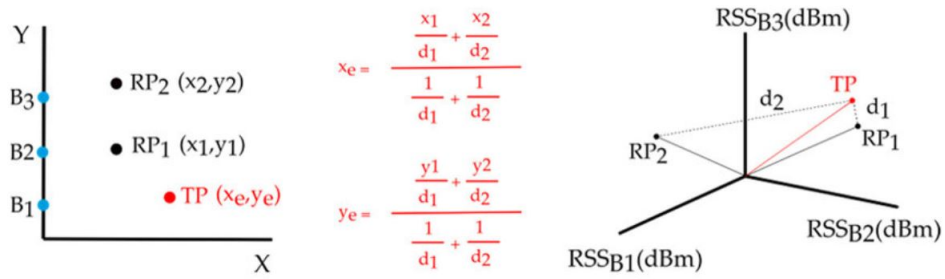

Figure 1: KNN Illustration

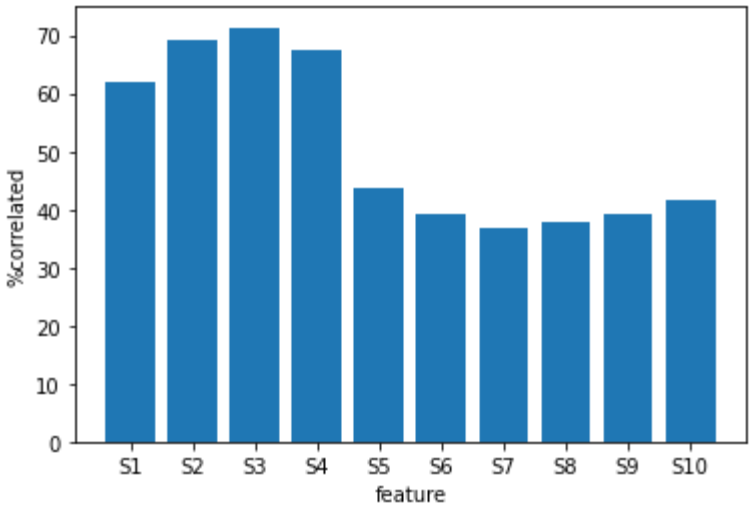

Figure 2: Results of the weighting of the features

```
We will now ask you 5 questions about your symptoms and 5 questions about pre-existing conditions
In the last 2 weeks have you experienced the following symptom: Breathing Problem?yes
In the last 2 weeks have you experienced the following symptom: Fever?yes
In the last 2 weeks have you experienced the following symptom: Dry Cough?no
In the last 2 weeks have you experienced the following symptom: Sore throat?no
In the last 2 weeks have you experienced the following symptom: Running Nose?yes
Does this condition apply to you: Asthma?no
Does this condition apply to you: Chronic Lung Disease?no
Does this condition apply to you: Heart Disease?no
Does this condition apply to you: Diabetes?no
Does this condition apply to you: Hyper Tension?no

Output of your score:
 score:  1.0336418658018323
 You are Covid-negative
```

Figure 3: Questions used to determine diagnosis

References

[1] Results we are trying to recreate: https://www.kaggle.com/meesalasaidhanush/symptoms-and-COVID-presence-99-acc

[2] https://www.researchgate.net/publication/334435471/figure/fig4/AS:780009182621696@1562980089805/Example-of-application-of-the-weighted-k-nearest-neighbor-WKNN-algorithm-for-two.jpg