Probability and Random Variables Project Report

1st Muhammad Irtiza
Dept. of Computer Engineering
Air University
180382@students.au.edu.pk

2nd Malik Ehtisham Ali
Dept. of Computer Engineering
Air University
180397@students.au.edu.pk

3rd Raza Ullah
Dept. of Computer Engineering
Air University
180385@students.au.edu.pk

***Abstract*—This document is about an expert systems, or decision support systems that have been trained with real cases to perform complicated tasks such as predicting occurring or non-occurring of COVID-19 in a human body based on some past and observed data.**

***Index Terms*—Naive Bayes Classifiers, Expert System, Decision Trees,**

I. Introduction

In probability theory and statistics, Bayes theorem (alternatively Bayes' law or Bayes' rule), named after Reverend Thomas Bayes, describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Those events could be our observed data. For example, if the risk of developing health problems is known to increase with age, Bayes theorem allows the risk to an individual of a known age to be assessed more accurately (by conditioning it on their age) than simply assuming that the individual is typical of the population as a whole. like that, we can also find the occurrence of a virus or disease in a body by knowing the symptoms.

One of the many applications of Bayes' theorem is Bayesian inference, a particular approach to statistical inference. When applied, the probabilities involved in the theorem may have different probability interpretations. With Bayesian probability interpretation, the theorem expresses how a degree of belief, expressed as a probability, should rationally change to account for the availability of related evidence. Bayesian inference is fundamental to Bayesian statistics.

An expert system (ES) is a knowledge-based system that employs knowledge about its application domain and uses an procedure to solve problems that would otherwise require human competence or expertise. The power of expert systems stems primarily from the specific knowledge about a narrow domain stored in the expert system's knowledge base.

It is important to stress to students that expert systems are assistants to decision makers and not substitutes for them. Expert systems do not have human capabilities. They use a knowledge base of a particular domain and bring that knowledge to bear on the facts of the particular situation at hand.The strength of an Expert System derives from its knowledge base, an organized collection of facts and heuristics about the system's domain.

II. Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

In decision, we will divide our data on the basis of different questions. Like in this data set, we have people infected with COVID-19 and not infected with COVID-19. So, we will divide our data in two sets of people infected with virus and not infected with virus. Also decision trees help us to make decisions for our algorithm.

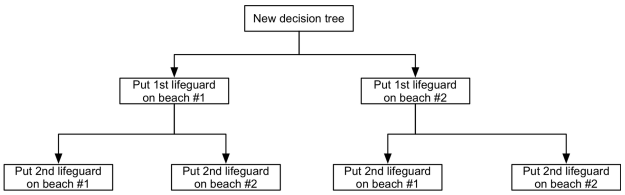

Fig. 1. Decision Tree

III. Bayesian Inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference is an important technique in statistics, and especially in mathematical statistics.

Bayesian inference derives the posterior probability as a consequence of two antecedents: a prior probability and a "likelihood function" derived from a statistical model for the observed data. Bayesian inference computes the posterior probability according to Bayes' theorem:

$$P(C|S) = \frac{P(C \cap S)}{P(S)} \tag{1}$$

$$P(C|S) = \frac{P(C)P(S \cap C)}{P(S)} \tag{2}$$

where C represents that your are infected with the COVID-19 and S represents the symptoms due to which you are infected with COVID-19. The Detailed explanation of the equation is given in section 5.

IV. Methodology

The algorithm we used to predict the occurrence of COVID-19 is mainly consist of decision tree and Bayesian Inference. Here, the Bayesian Inference is the core to find the output from the expert system.

A. Pseudo Code of COVID-19 Prediction

Algorithm 1: COVID-19 Prediction

```
Read prior symptom data;
Read test data;
Split prior data into categories (infected/Not infected);
Calculate conditional probability of each test symptom
  for each category;
Calculate total total probability of test data for each
  category;
Normalize Probabilities by dividing then with the sum
  of probabilities (total Probability);
if If the Probability for infected category is greater
  then
    The patient is infected;
```

B. Flowchart of COVID-19 Prediction

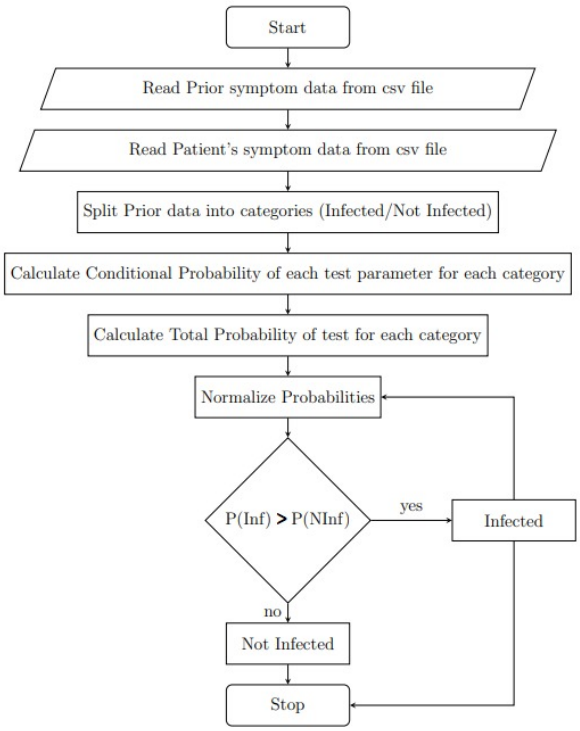

Fig. 2. Flowchart for City Lockdown Recommendation

Algorithm 2: City Lockdown Recommendation

```
Read City data from csv file;
Calculate score for each city;
Normalize score by dividing by max score;
if Cityś score is greater than mean score then
    Recommend lockdown;
```

C. Pseudo Code of City Lockdown Recommendation

D. Flowchart of City Lockdown Recommendation

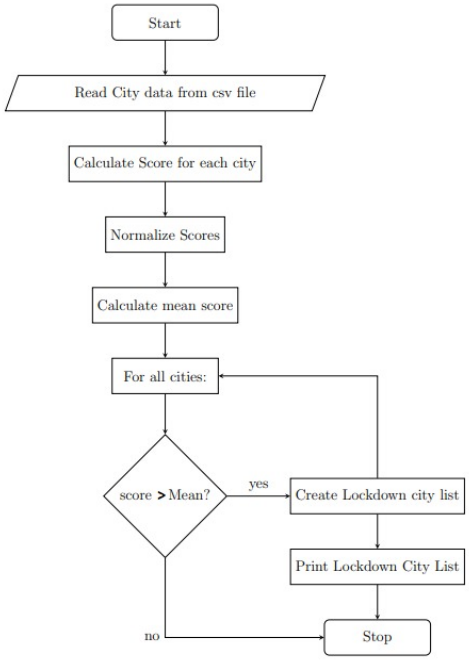

Fig. 3. Flowchart for City Lockdown Recommendation

V. System Design

As we have learnt about the Bayesian Inference in section 2, let us now see how we calculated the probabilities. Let, C represents that you are infected with COVID-19 and S represents the symptoms because of which you get infected. So, the Bayesian Inference will be:

$$P(C|S) = \frac{P(C)P(S \cap C)}{P(S)} \tag{3}$$

There are multiple symptoms of COVID-19, so the S in our equation can be written as:

$$S = S_1, S_2, S_3, ...S_n \tag{4}$$

So, our Bayesian Inference equation will be,

$$P(C|S_1 \cap S_2..) = \frac{P(C)P(S_1 \cap C)P(S_2 \cap C)..}{P(S_1 \cap S_2..)} \tag{5}$$

and

$$P(C^{'}|S_1 \cap S_2..) = \frac{P(C^{'})P(S_1 \cap C^{'})P(S_2 \cap C^{'})..}{P(S_1 \cap S_2..)} \quad (6)$$

where $C^{'}$ represents that you are not infected with COVID-19. As we are going to compare these two equations 4 and 5, so the value we have in denominator $P(S_1 \cap S_2..)$ will not affect our result. So we can ignore these values which will make it easy for us to calculate the result.

Now we will calculating these equations and find probabilities. We can find the value of $P(C)$ and $P(C^{'})$ as:

$$P(C) = \frac{No.of patients infected with Covid19}{total number of patients} \quad (7)$$

$$P(C^{'}) = \frac{No.of patients not infected with Covid19}{total number of patients} \quad (8)$$

And you we can find the value of $P(S_1 \cap C)$ by dividing the counts of that symptom in corona infected patientś data set with total number of corona infected patients. Now we will continue with this process until we find the probability of every symptom such that corona exist and such that corona doesnt́ exist. As now we have find the values of $P(C), P(S_1 \cap C), P(S_2 \cap C) and P(C^{'}), P(S_1 \cap C^{'}), P(S_2 \cap C^{'})$ now we will multiply these values to find the probability of you are infected with corona virus and you are not infected with corona virus. Now, if the probability of that you are infected with COVID-19 is greater than that of you are not infected than we will say that the patient is not infected with COVID-19.

The current Lockdown Recommendation program does not use Bayesian inference. It uses the past data of cities/regions (incl. Population, Population Density, Health conditions, Health Provider Standards, Infection rate, Demographics) to calculate an intuitive score for each city. Cities above a score threshold, e.g mean score, are recommended to be put on lockdown since they are deemed the least prepared for infection control.

VI. Advantages/Disadvantages of Naive Bayesian Classifier constructed from the probability model

A. Advantage

- Simple Implementation.

B. Disadvantages

- Accuracy greatly depends on the prior data.
- Only works on discrete data already present in prior information. If a test value is not present in prior information for that outcome (e.g. age, temperature), its conditional probability will be 0. As a result, the probability of being infected will be wrongly 0. The lack of support for continuous information limits the information that can be used to train the model

C. Solution

- A roundabout was of making up for the drawback is to use range/groups for continuous data. But depending on the dataset, it is possible for an entire group to be absent from the training dataset. Though, a data range being mussing from the training data is much less likely than a missing value.
- Use of Parameter estimation and event models like Gaussian naive Bayes, Multinomial Naive Bayes,Bernoulli naive Bayes can remove the limitation of using discrete data.
- Gaussian probability function uses the mean and standard deviation of the column of data , along with the test value for that column in its formula to calculate the conditional probability of that variable. Unlike the current system, it will be able to predict the results given test data whose elements are not found anywhere in the training dataset.

VII. Results

- The Covid19 detection program was provided a small dataset containing symptoms and Covid19 test results, and the test symptoms for prediction.
- The system predicted the outcome using the patientś symptoms.
- The test data i.e symptoms were given to the program and here is the result,

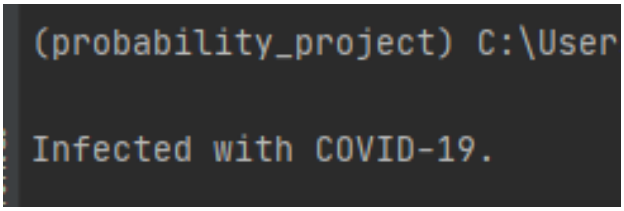

Fig. 4. Result of COVID Prediction

As you can see, we gave some test symptoms to our program and he gave us the output whether the patients is infected with COVID-19 or not. As you can see, we gave some test

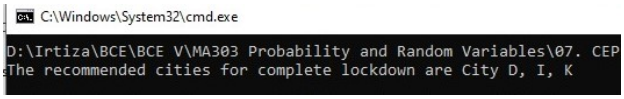

Fig. 5. Result for City Lockdown Recommendation

symptoms to our program and he gave us the recommendation about the cities.

VIII. Future Work

While symptom checking tools are not a substitute for professional medical advice, they help patients keep track of their health.

It would be desirable to develop a personalized diagnostic system to predict the diseases a person may be suffering from by using general symptoms and the individual's medical history.

IX. Project Code

The complete Project code and input data files can be found in this GitHub repository, and can be accessed using the hyperlink or the url below:
https://github.com/ehtishamalik/Pobability-Project.git

X. Disclaimer

This project was carried out as a learning activity.The Content is not intended to be a substitute for professional medical advice, diagnosis, or treatment.

References

[1] https://www.eurekaselect.com/82021/article/medical-expert-systems
[2] https://www.geeksforgeeks.org/naive-bayes-classifiers/