DEVELOPMENT OF A FORMULA TO DETERMINE THE RISK OF PATIENTS WITH CORONAVIRUS

Álvaro de la Fuente, Alfonso Gordon, Pablo Rodríguez and Daniel Anchuela

Abstract

In this document we will explain the fundamentals from which we started to build the formula that we will use in determining the risk score of patients in our application Covidist, as well as the steps that we use in its development making use of all the statistical data obtained and calculated by the application.

In order to obtain a mathematical modeling as realistic as possible, we will use the percentages of the data used, we will use weights when assigning a patient to their case, we will increase the data score to different numbers depending on the case in order to reward more the score that requires it, and we will multiply it by the value of the importance that type of value has in that moment. Because data can be provided from patients who have just be infected or from patients that have died. In this last case, the data will be more significant.

Explanation

The most important element our application needs to calculate the risk score of a patient in each moment is all the data that our application has in that moment.

Having said that, we start to model our formula starting from the percentages of each feature that we have already received from the class that calculate all the statistics.

In order to start to start to model the formula, we are going to use the feature *sex.* It is one of the simplest ones because it only has two possible cases: men and women. A possible solution could be giving more score to the highest percentages.

However, this idea, sometimes is not realistic. For example, in this case the application will give a higher score to a 50% of men, and a lower score to a 40% of patients that all of them suffer the same disease. It does not make sense, so the best idea would be calculating the difference between the highest and lowest percentage. This way, we obtain:

$$Pmax_i - Pmin_i$$

Where *Pmax* determines the percentage of the most common case and *Pmin* determines the percentage of the less frequent case among all the possible ones in each characteristic (represented by the subindex *i).*

Once we have obtained the difference between the percentages, we need to associate that difference with the case the patient that we are treating belongs to. The way we have reasoned would be multiplying that difference by the percentage as per one that the patient belongs to.

$$(Pmax_i - Pmin_i) \cdot Pp_{p,i}$$

Where $Pp_{p,i}$ is the percentage as per one the patient *p* belongs to in each characteristic.

However, this formula is giving a score to a patient just based on the percentages of that characteristic. By the moment, we are obviating the fact that some features have more importance than others depending on the number of cases that it could be.

Going back to the previous example, with the characteristic *sex*, we could say that it is important that 70% of men are infected, for example. However, we consider that, for example, the fact that 35% of all patients suffer from cancer, with all the possibilities that it could be is something that we have to consider. Therefore, we are going to increase that score to "1." and followed by the number of possible cases of the characteristic. We increase the score to that number because it grows exponentially, so a higher difference will generate a higher final score in the features with more cases. And we didn't decide to increase it to the power of

the number of total cases in order to avoid huge numbers. So, by the moment, our formula would be like that:

$$\left[(Pmax_i - Pmin_i) \cdot Pp_{p,i}\right]^{\left(1+\frac{Nopt_i}{10}\right)}$$

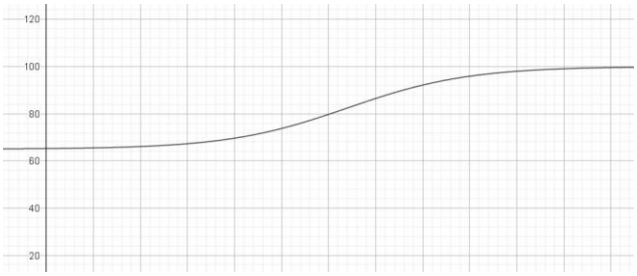
Where $Nopt_i$ is the number of cases the characteristic *i* has.

At this point, we have a formula that takes into account the difference in percentages, associates this difference to the case to which the patient in question belongs and assesses the whole range of possibilities of each data to give a higher score to those features that have more cases since the percentage is more distributed. But at the moment we are not taking into account that our data can be of two types:

1. Data from patients who have already died.
2. Data from patients who have been infected with coronavirus.

Since what we want is to prevent the people from dying, we are going to give more weight to the data from dead patients. However, this weight must be variable over time because it has to depend on the number of dead patients we have in recent days. If, for example, the data says that, in recent days, the number of coronavirus patients who have died rises exponentially, we will have to give it more weight than it normally has. For this reason, to obtain the formula that determines the importance of the data on dead patients, we will receive the data on these patients from the class that calculates the statistics, as we did previously.
In this case, we will receive the average rate of increase or decrease in deaths of the last 7 days. We do this to prevent the possibility of sharp rises and falls in deaths. We cannot just look at the rate of increase of the previous day because it may have risen very little, and we would be ignoring the fact that the day before that it could have risen a lot. In this way, starting from this last data, we proceed to model the formula for the importance of the data on dead patients.

First, we will establish an importance limit, which in this case will be 65% and 35%. What we are doing now is to say that, although there is a very low number of deaths, we want that the minimum value of the importance of the information received from the data of the dead patients is 65% compared to the other 35% that would be the importance that the data of the just-infected patients would receive, always depending on the other one. In addition, we will determine a limit of 200% growth of deaths, which, if exceeded, the importance of dead patient data would tend to 100%. That said, the plot of the formula that we obtain would be the following:



In this way, the function that this graph represents and that depends on our variable a_d that represents the average of the rate of increase of the last 7 days would be:

$$f(a_d) = \frac{35}{1 + e^{-0.039 \cdot a_d + 5}} + 65$$

Although for our case we will divide everything by 100 because that is how we have been doing it until now.
For this reason, the formula that determines the value of the data of the infected patients would be the following:

$$1 - \frac{\frac{35}{1 + e^{-0.039 \cdot a_d + 5}} + 65}{100}$$

Finally, as we have explained before, we put everything together in the same formula to give the result of the score obtained from a single data, depending on whether it belongs to infected patients or to died patients. In addition, to obtain

the total score of a patient considering all the data obtained we will use a summation that will go from the first data to *n* which represents the total number of data we have. In this way, our final formula would be like this:

$$\underline{Score\ of\ a\ patient\ \mathrm{p}\ in\ a\ day\ \mathrm{d}}$$

$$score(p,d) = \sum_{i=1}^{n} \left\{ \left[(Pmax_i - Pmin_i) \cdot Pp_{p,i} \right]^{\left(1+\frac{Nopt_i}{10}\right)} \cdot \left[\underbrace{\frac{35}{1+e^{-0.039 \cdot a_d + 5}} + 65}_{\substack{if\ characteristic \\ of\ dead\ people}} , \underbrace{\left(1 - \frac{35}{1+e^{-0.039 \cdot a_d + 5}} + 65 \right)}_{\substack{if\ characteristic \\ of\ infected\ people}} \right] \right\}$$