Pear Inc. Question

April 1, 2022

DEÜ Bilgisayar Bilimleri ve Yapay Zeka Topluluğu

---

1 Welcome to Pear Inc.

Hi there! My name is Robert! You can call me Bob . I'm the communications officer (fancy title ha!) in our glorious company . My job is to help facilitate product development and market penetration . I spent endless hours talking to engineers, product managers, and customers

Since we are a 20 people start-up (All of us have fancy names), I also do some recruiting from time to time We are looking for **brave souls who are not afraid of a challenge and will help us** with our new product line of smart t-shirts!

(Our CEO believes that smart t-shirts are the right direction for some reason I guess if you make something nobody needs, you won't have to sell it)

Let me tell you a little bit more about our problem that you can help us with: We are creating a life changing smart t-shirt which has bluetooth and connects to your phone . They will be customizable outfits through downloaded applications. Our smart t-shirt will be developed with Google Wear OS which is a version of Google's Android operating system designed for smartwatches and other wearables. So users will be able to install custom programs through Google Play Store . And we will sell them for 999.9$ a piece But our engineers wanted to ensure that only Pear Inc. approved

programs can be installed on our t-shirts because market analysis showed that potential customers are afraid of ransomware that will break their "*premium*" t-shirts . So we need an antivirus for approving apps on the fly! However, we don't want to install an off the shelf antivirus to our t-shirts , because BIG profit margins matter !

Enough chit-chat! Let's get down to the business of why I contacted you: Our bright engineers came up with an algorithm that creates compressed signatures for the apps in the Google Play Store. It is called '*manifold averaging generally intelligent compressor*' or as we call it 'MAGIC'. The engineers told us that the outputs of MAGIC reflect the statistical properties of the uncompressed apps (whatever that may mean!). MAGIC takes a Google Play Store app as an input and outputs a 4 dimensional numerical signature (they called it a vector but calling it a vector is not fancy enough for marketing!).

Now, since these signatures are just numbers, an off the shelf antivirus can't work with them (even if it could, we can't install an off the shelf antivirus into our t-shirts – too much computing power and space is needed). Therefore **we need a light weight proof of concept that takes these signatures as inputs and outputs labels (virus or not) for them.** We eventually want to install your program into our smart t-shirts, where it will scan a Google Play Store app (its signature to be precise!) and stop the app's execution if it thinks the app is a virus! But we are not going so far just yet so you only need to create the pipeline that take the signatures, and output labels for them. Don't worry about the rest, it is just a proof of concept at the end . We are providing the dataset for you to develop your model.

In a nutshell: - There 4 dimensional (4 feature) numerical inputs (signatures) with labels! - We need a simple model that takes these inputs and labels them (Virus, Not a Virus) - We also need you to evaluate your model. Choose any metric you want, but don't forget to explain why, since I don't know much about this field (that is why we need your help!)

Things to keep in mind: - There are less 'Virus' in the dataset than 'Not a Virus'. (Naturally!) - While we call it MAGIC, it still sometimes doesn't work well , so there are signatures with missing features (missing values). - I don't know much about these things so please show your work, your thinking process and please make it as clear as possible, otherwise I get confused . (Visualizations of the data and comments in your code would be great!)

Let me describe the dataset, and you are ready to get to work! It is a CSV file. Each row represents a signature for an app. First 4 columns from left to right represent dimensions (features) and the last column is the label (isVirus: True or False).

- Visualize the data (so that people like me can understand!)
- Clean up the data (balance it out, impute missing values and so on... depending on the method you are going to use!)
- Visualize the cleaned data (so that people like me can understand the effect of cleaning process!)
- Create a simple model that performs reasonably well. (If it doesn't perform well, comment on why and how to improve it!)
- Evaluate the model with a testset you will create from the dataset. (Pretty plots make things easier to understand)

- Upload your code to a private github repo you can share with us, and invite us https://github.com/deubbt as collaborators so only we can see your super-secret project.

And you are done! (Don't forget to comment, and show your work please)

1.0.1 SOLUTION :

[1]:
```
# Your code here!
```