PRINCETON COS CON



# Problem 6: Cyberchondria (20 points) [Email Submission]

By Pedro Paredes

The Yuuzhan Vong species is an advanced sentient life form that lives in a far away galaxy. One of the leading causes of death among the Yuuzhan Vong is a disease known as *Cyberchondria*, a fatal condition that affects the brain and nervous system. Doctors have been studying this disease for a while and collecting data on it, and it turns out the most important thing to contain this disease is to diagnose it as early as possible.

With that in mind, the doctors of Destrillion University have collected a data set of brains of patients, some of which eventually developed Cyberchondria. Each one of these samples was labeled depending on whether or not the patient eventually got the disease. Each brain is represented according to its lobe structure, so it is defined as a collection of lobes (i.e. sections of the brain) and connections between pairs of lobes, which represent neural communication paths formed by synapses. Each connection is bidirectional (i.e. communication can happen in both directions along one neural path) and it is possible that a pair of lobes has multiple connections between them.

You are given this labeled data and also an unlabeled data set of brains of patients. Your task is to predict which of the new patients are at risk of developing Cyberchondria.

*This is a binary classification task. You will be given the labeled and unlabeled data sets and are free to apply whatever methods or tools you wish to use to come up with a label prediction for the unlabeled set. Once you are happy with your predictions, you should email in the proper format (specified below) them to the COSCON submissions email* `coscon.submit@gmail.com`*. The subject of you email should be "Problem6Submission" and you should include your team members' names on the body of the email. Your final score will be revealed after the contest is over.*

*The following sections describe the data format, scoring function and how the data was obtained.*




### Data format

The labeled data set collected by the doctors of Destrillion University is stored in file `train.txt`. This file contains $4,000$ lines, each one corresponding to one labeled patient brain instance. Each line starts with three space separated integers: $L$ $n$ $m$ — $L$ is either `0` or `1`, and it represents whether or not this is a sample from a healthy brain (`0`) or a brain from a patient that developed Cyberchondria (`1`); $n$ is an integer between 80 and 120, representing the number of lobes of this brain; $m$ is an integer between 0 and $4,950$, representing the number of connections between lobes. Then follow $m$ pairs of space separated integers, each pair $u$ $v$ consists of two distinct integers between 1 and $n$ that indicate that lobe $u$ has a connection to lobe $v$.

Here is an example of fake data set following the above format:

```
0 3 4 1 2 1 3 2 3 2 1
1 3 2 3 1 2 1
```

In this example there are two patient brains, the first brain is a healthy brain containing 3 lobes and 4 lobe connections such that lobes 1 and 2 are connected twice, lobes 1 and 3 are connected, and lobes 2 and 3 are connected. The second brain is a diseased brain containing 3 lobes and 2 connections such that lobes 1 and 3 are connected, and lobes 1 and 2 are connected.

The unlabeled data set you should predict labels for is stored in file `test.txt`. This file contains $1,000$ lines, each one corresponding to one unlabeled patient brain instance. Each line follows the exact same format as before, but it does not contain the initial integer label.

Here is an example of fake data set following the above format:

```
3 2 3 1 2 1
```

In this example there is one patient brain, which is exactly the same as the second example from example fake labeled data set.




### Output format

To make your prediction you should produce a `predict.txt` file with exactly 1,000 lines. Each line should contain either a `0` or a `1`, indicating to your prediction for the corresponding patient brain from the `test.txt` data set.

If your output format doesn't follow this format (e.g. it has more/less than 1,000 lines, contains any symbol other than a `0` or a `1`, etc) it will score 0 points.

### Scoring function

For a certain prediction, let $H$ be the number of correctly identified healthy patients and $D$ the number of correctly identified diseased patients. Then this prediction would score $H + 10D$. This means that correctly identified diseased patients are worth more, since we would rather be overly cautious than miss a possible Cyberchondria case. You goal is to maximize this score.

*Note that this score is not the same as the number of points you'll get in the problem. Once the contest is over we will find the scores of all the submissions and the highest score will get 20 points. All the other submissions will get a number of points which will be a linear interpolation between the highest obtained score and the score obtained by random guessing. So random guessing (picking* `0` *or* `1` *with probability .5 each) will get 0 points and anything between that and the highest score will get a proportional number of points.*

### Data description

*We tried obtained real data of brains from Yuuzhan Vong patients, unfortunately we were unable to obtain it due to the fact that this species is made up. So instead we used synthetic data and here we describe some of the process we used to obtain that data. A lot of information will be missing, so you will have to explore the data in order to learn more about it and be successful in this problem.*




Each healthy patient brain data, from both the labeled and unlabeled sets, was obtained by sampling a network from a random graph distribution, which we will call the *healthy distribution*. Note that this distribution is *not* necessarily uniform, but each healthy brain data was sampled independently from the same healthy distribution.

Each diseased patient brain data, from both the labeled and unlabeled sets, was obtained by first sampling an element from the healthy distribution, and then altered by adding or removing a small number (less than 20% in expectation, so the expected value of edges added or removed is at most $.2m$, for a base brain with $m$ connections) of connections at random (not necessarily uniformly), according to some random process which we will call the *Cyberchondria random process*. Each diseased brain data was sampled independently from the same healthy distribution and then independently altered according to the same *Cyberchondria random process*.

To construct the two data sets the following procedure was followed: it is known that the rate of incidence of Cyberchondria is some fixed number $p$ less than 0.5. So around $5,000p$ diseased brain instances where created using the above method and around $5,000(1-p)$ healthy brain instances. This data was then uniformly partitioned at random into a set of $4,000$ elements for the labeled data set and $1,000$ for the unlabeled data set.

Additionally, we know that there was some small corruption of the labeled data, and around 5% of the labels were flipped (`0`s turned to `1`s and vice-versa). This corruption was uniform, meaning that for each element of the labeled data set a biased coin of probability .05 was tossed to decide whether a label was flipped or not, independently of the original label. Note that the unlabeled data set suffered no corruption, they are all true samples from the above distributions.