Capstone Proposal

Urinary system diseases classification

For this project I decided to work with the dataset “Urinary system diseases classification” found on Kaggle at https://www.kaggle.com/yamqwe/acute-inflammationse. Improving diagnoses can help patients to have an early treatment, faster recovery and less spending with medical routines, surgeries and medications. This is a great incentive for me to work on this data as even I am a software developer I can help others with serious problems in their lives.

Each row in the dataset consists of symptoms about a potential patient and there two possible diseases to be diagnosed, Acute Inflammations of Urinary Bladder or Acute Nephritises. The Kaggle dataset webpage has some descriptions about each disease. I am going to quote them as I am not a medical expert:

- Acute inflammation of urinary bladder

Acute inflammation of urinary bladder is characterized by sudden occurrence of pains in the abdomen region and the urination in form of constant urine pushing, micturition pains and sometimes lack of urine keeping. Temperature of the body is rising, however most often not above 38C. The excreted urine is turbid and sometimes bloody. At proper treatment, symptoms decay usually within several days. However, there is inclination to returns. At persons with acute inflammation of urinary bladder, we should expect that the illness will turn into protracted form.

- Acute nephritis of renal pelvis

Acute nephritis of renal pelvis origin occurs considerably more often at women than at men. It begins with sudden fever, which reaches, and sometimes exceeds 40C. The fever is accompanied by shivers and one- or both-side lumbar pains, which are sometimes very strong. Symptoms of acute inflammation of urinary bladder appear very often. Quite not infrequently there are nausea and vomiting and spread pains of whole abdomen.

The objective is to create a machine learning model that will receive the symptoms as input and inform if the patient has one of the diseases, both or none. The model will train in a larger sample of the given dataset and evaluate on the smaller sample from the same dataset. I am also going to create a webpage in AWS where new data can be provided to the model and it will return the expected diagnose.

The given dataset has 6 symptoms and two diagnoses, but both diagnoses can be “yes” at the same time, which means the patient has both diseases, or “no” for both, which means the patient doesn’t have any of these diseases (it is not clear from the dataset if the patient is healthy, but it is not any of these two diseases in this project).

These are the symptoms, given in order in the dataset:

1. Temperature of patient – number from 35.5 to 41.5 (Celsius degree)
2. Occurrence of nausea – yes or no
3. Lumbar pain – yes or no
4. Urine pushing (continuous need for urination) – yes or no
5. Micturition pains – yes or no

6. Burning of urethra, itch, swelling of urethra outlet – yes or no

The dataset has two columns, one for each diagnose, they can also be both yes or both no

1. Inflammation of urinary bladder – yes or no
2. Nephritis of renal pelvis origin – yes or no

The dataset file is a TAB separated file, for each row, each symptom is separated by a tab character \t.

The solution for this problem is a classification model with multiple categories. We have seen in the nanodegree course some examples of classification models, but they were only binary, as for example the Fraud Detection, which the result could be only two possibilities, 1 or 0 (fraud or no-fraud). Here we have 4 possibilities: no diseases, only Inflammation, only Nephritis or both. In machine learning is common to have multiple classification algorithm, I would have to make some adjustments from what I have learned during the course. Once the model is ready, it can be reproduced and applied in a production environment. I'll do it by setting up a lambda function, a API gateway and a webpage for new patients with new symptoms.

The paper in the citation concludes with a strict rule for determining the diagnose. My proposal is different as I am going to create a Neural Network model using PyTorch.

```
Rn1: IF(c2=n)&(c4=n)&(c5=n)&(c6=n)                    THEN(d1=n)&(d2=n)
Rn2: IF(c1=g)&(c2=n)&(c3=t)&(c4=t)&(c5=n)&(c6=t)  THEN(d1=n)&(d2=t)
Rn3: IF(c1=w)&(c3=t)                                  THEN(d1=n)&(d2=t)
Rn4: IF(c1=p)&(c2=n)&(c3=n)&(c4=t)                    THEN(d1=t)&(d2=n)
Rn5: IF(c1=n)&(c2=n)&(c3=n)&(c4=t)&(c5=t)&(c6=t)  THEN(d1=t)&(d2=n)
Rn6: IF(c1=w)&(c2=t)&(c3=t)&(c4=t)&(c5=t)             THEN(d1=t)&(d2=t)
```

Figure 1: Deterministic rule from citation paper

My Neural Network model can be evaluated using accuracy metric, if the final diagnose is the same as the given label.

The dataset requires to give citation and credit:

Citation Request:
J.Czerniak, H.Zarzycki, Application of rough sets in the presumptive diagnosis of urinary system diseases, Artifical Inteligence and Security in Computing Systems, ACS'2002 9th International Conference Proceedings, Kluwer Academic Publishers,2003, pp. 41-51

Source: http://archive.ics.uci.edu/ml/datasets/Acute+Inflammations