proposal

June 8, 2019

1 Machine Learning Engineer Nanodegree

1.1 Capstone Proposal

Eoin Kelly
June 4th, 2019

1.2 Proposal

This is a study into a mechanisms based approachto diagnosing Lower Back Pain (LBP)

1.2.1 Domain Background

This project is based upon a study by *Smart et al.* in 2011 [1] in St. Vincents Hospital in Dublin. It is a study into three different types of musculoskeletal back pains. These are "nociceptive," "peripheral neuropathic," and "central sensitization" pain in patients with lower back pain disorders. We are looking to find a way to diagnose which group these patients belong to based on what symptoms they are feeling using both clustering and classification methods.

1.2.2 Problem Statement

The problem with back pain is that it can be difficult to diagnose exactly what the root cause of the problem is. If we are able to analyse a number of symptoms that our patients show, we can try to cluster them into groups and certain groupings we hope to show are part of different types of lower back pain. This would help doctors to diagnose the best treatments for their patients. We have the response variable for each of these patients so we will be able to measure how well our models have performed using some format of training, validation and testing models so as not to bias our research.

Once we have these clusters of symptoms, we will be using the classification method of learning to attempt to diagnose which type of LBP the patient is showing. The symptoms/inputs that we will be analysing are different symptoms resulting from back pain. We will be putting these through our classifier algorithm and this will give us an output which is the type of LBP exhibited by our patient. We will then compare our classifications against the real symptoms which are also recorded in the dataset. This is therefore a supervised learning multi-class classification problem.

1.2.3 Datasets and Inputs

The dataset to be considered for this project was available on my past university's website and it had been touched on during one of our lectures which is where I got the idea to use this as my final project for this course. There is no sensitive data within our dataset.

The dataset itself contains data acquired by doctors examining patients with lower back pain. There are 464 subjects in this study. It contains the group of back pain the patient belongs to, which is our response variable and can be any of the categorical responses:

1. Nociceptive
2. Peripheral Neuropathy
3. Central Neuropathy

Each patient is also asked whether or not they have each of 36 different symptoms, described below, and all of these answers are recorded. We will be using this data to try and predict the group of back pain that the patient hasbased on what his/her symptoms are. The symptoms analysed leave us a binary yes/no set of data points regarding the following 36 LBP symptoms:

1. Pain of recent onset.
2. Pain associated with trauma, pathology, and movement.
3. History of nerve injury, trauma.
4. Pain disproportionate to injury, pathology.
5. Intermittent and sharp or constant dull ache.
6. More constant, unremitting.
7. Burning, shooting, sharp, electric-shock like.
8. Localised to area of injury/dysfunction.
9. Referred in dermatomal, cutaneous distribution.
10. Widespread, non-anatomical distribution.
11. Mechanical nature to aggs + eases.
12. Mechanical pattern associated with movement, loading, and compression of neural tissue.
13. Disproportionate, non-mechanical pattern to aggs + eases.
14. Spontaneous, paroxysmal pain.
15. Pain with dyesthesias.
16. Pain of high severity and irritability.
17. Pain with neurological symptoms.
18. Night pain, disturbed sleep.
19. Response to simple analgesia, NSAIDS.
20. Rapidly resolving, resolving with expected tissue healing, pathology recovery times.
21. Pain persisting beyond expected tissue healing, pathology recovery times.
22. History of failed interventions.
23. Strong association with maladaptive psychological factors.
24. Pain with high levels of functional disability.
25. Antalgic postures, movement patterns.
26. Consistent, proportionate pain reproduction on mechanical testing.
27. Pain, symptom provocation with mechanical tests that move, load, compress neural tissue.
28. Disproportionate, non-mechanical pattern of pain provocation on mechanical testing.
29. Positive neurological findings.
30. Localised pain on palpation.
31. Diffuse, non-anatomic areas of pain on palpation.

32. Positive findings of allodynia.
33. Positive findings of hyperalgesia.
34. Positive findings of hyperpathia.
35. Pain, symptom provocation on palpation of neural tissues.
36. Positive identification of psychosocial factors

1.2.4 Solution Statement

I would like to first begin by looking at clustering techniques to see if we can find any defined patterns in the different pains that our patients are feeling to see if we can someway group them together. Following this, the plan is to begin classifying the patients to any of the three groups of lower back pain. We will then use a number of different measures of accuracy to see how well our model has worked. We will measure simple accuracy, sensitivity and also specificity calculations to see how well we are able to diagnose patients. We do not want to misdiagnose patients so therefore we will likely use the recall method of sensitivity when analysing our results.

K-means clustering is the method we will be looking at in order to group our aymptoms together. We will be checking the silhouette scores of these clusters for different values of k to find the correct value and to see how well grouped together our clusters are. A mixture of decision trees, random forests, adaboost modelling is what we will be using to find a measure as to how well our model actually works and to see how well we can use this mechanisms-based approach to diagnose patients displaying signs of lower back pain.

1.2.5 Benchmark Model

For a full up to date description of the issues underlying the diagnosis and definition of neuropathic pain see *Finnerup et al. (2016)*. [2] This explains that diagnosing each of these LBP conditions is a difficult task to get right. For neuropathic pain alone, clinical practice guidelines typically suggest that the prevalence of neuropathic pain in LBP is approximately 5%; however, some reports suggest that as many as 16–55% of patients with chronic LBP have possible neuropathic pain components *(R. Baron et al. 2016)*. [3]

We will use a benchmark model of logistic regression in analysing these patients and we will use this as our benchmark to see how well our model actually works.

1.2.6 Evaluation Metrics

The probability that given a random patient in the dataset, lies in each group is 0.228. 0.543 and 0.219 respectively meaning this is an unbalanced dataset. As the dataset is unbalanced, the focus should be on improving the sensitivity and specificity scoring. These scores allow us to analyse models that prioritize either false positives or false negatives. As we are diagnosing patients here, we would prefer to have as few false negatives as possible, meaning we recall patients for further testing. With our unbalanced dataset, I think looking at the sensitivity and specificity should be sufficient in measuring our algorithms success rate, however if needs be we could also look at the AUC ROC curve as an extra measurement.

1.2.7 Project Design

Exploratory Analysis Starting from the top, we will examine the data and clean it up. There are a small number of datapoints (~5%) with some data mising, we will omit these data points from the group and work with our observations which have all of their symptoms evaluated.

Clustering After our initial exploration of the data, we will begin by using some clustering algorithms to see if we can find patterns in where certain symptoms are found. Are some symptoms paired and show up together with other symptoms. Are we able to create clusters in this without using our response variables and then compare our answers versus the real solutions to check to see if our clusterings match to the real diagnosis. We will use k-means clustering to find our groupings and we will check the silhouette scores to see how well defined these clusters are. If we need to find out the most important factors/symptoms our patients feel, we can use PCA to find the biggest deciding factors. A lot of exploratory analysis is needed on this part.

Classification We will then dive into some classification models to see how well our decision trees, random forests and Adaboost models get on. We will try each of these models, after splitting the data into training and testing sets. If a validation set is required, we will add this too. This will identify which are the most important factors/symptoms at deciding which pain the symptom belongs too and use this to choose our chronic back pain group.

Conclusion We will discuss all of our results, how our new supervised classifier beats the benchmark evaluations of logistic regression and also how it compares against a clustering based approach. This section will highlight the best model to use and exactly how accurate it is at diagnosing Lower Back Pain in patients.

1.2.8 References

[1] Keith M. Smart, PhD,* Catherine Blake, PhD,w Anthony Staines, PhD,z and Catherine Doody, PhDw - https://www.elleboogkliniek.nl/wp-content/uploads/Smart-et-al-The-Discriminative-Validity-of-Nociceptive-Peripheral-Neuropathic-and-Central-Sensitization-kopie.pdf

[2] Finnerup N.B., Haroutounian S., Kamerman P., Baron R., Bennett D.L., Bouhassira D., Cruccu G., Freeman R., Hansson P., Nurmikko T. Neuropathic pain: an updated grading system for research and clinical practice. Pain. 2016;157:1599–1606.
[PMC free article] - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4949003/

[3] R. Baron, Binder, N. Attal, R. Casale, A.H. Dickenson and RD. Treede - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5069616/