Project proposal Machine Learning

Thomas van Ess (s2601109), Christiaan Steenkist (s2744244)
Daan Sijbring (s2410133), Leonard Praetorius (s3779394)

1 Introduction

We were tasked with designing a project, in particular, concerning a dataset from a reputable source. This is a machine learning project aimed at learning more about prediction, regression, classification and/or clustering. Our project covers the mixture of regression, classification and prediction of variables related to patient reviews of drugs and their effectiveness.

2 Dataset

We have chosen to use the Drug Review Dataset (`Drugs.com`) [4] from the UCI Machine Learning Repository [1]. This dataset contains reviews of various drugs that were given to patients for a number of ailments. In addition, a second Drug Review Dataset [3] from the same authors was published, that contains similar fields. This data here is instead taken from `Druglib.com`.

2.1 `Drugs.com` dataset

The fields from the `Drugs.com` dataset are as follows:

- Drug name (categorical)
- Medical condition (categorical)
- Patient review (text)
- 10 star patient rating (numerical)
- Review date (date)
- Number of people that found the review useful (numerical)

2.2 Druglib dataset

The fields from the Druglib dataset are as follows:

- Drug name (categorical)

- Medical condition (categorical)
- Benefits review (text)
- Side effects review (text)
- Other comments (text)
- 10 star patient rating (numerical)
- 5 step side effect rating (categorical)
- 5 step effectiveness rating (categorical)

2.3 The final decision

We decided to use `Drugs.com` as our main focus over the Druglib dataset. The Druglib dataset has 8 fields instead of 6, three of which are text fields. However, it only contains a meager 4000 data points while the `Drugs.com` dataset contains over 200 thousand. In the interest of having enough training data for our future model(s) we went with the larger dataset.

3 Goals

The original use [2] of these datasets was to use sentiment analyses of the various types of reviews along with looking into the transferability of the models. Just like the original study, we would like to mainly analyse and use the patient reviews.

3.1 Text analysis

We want to try to predict the rating that was given by the patient via the text review that the patient submitted. Additionally we would also like to see if predicting the usefulness of the review is possible in this way. Text sentiment may have a strong correlation to the patient rating, but other models are possible as well. A Bayesian model may be able to predict these variables when they are put into categorical bins or even in a direct way. A neural network is suited to both the categorical approach and training to predict the value outright.

3.2 Extra

We are also interested in looking into the relations between other fields, especially the date. It would be interesting to see if there are any seasonal changes in medical conditions, ratings and text sentiment. This may fall outside the scope of the project if we are to focus on our main goal and take too much time. If there is spare time we may also attempt to transfer our models from the `Drugs.com` dataset to the Drugslib dataset. This is not our main goal, however.

4 Bi-weekly planning

week 3-4: Search for relevant literature, similar research, existing software and start developing the machine learning algorithm. week 5-6: Complete the first version of our algorithm, additional tweaks may follow based on the results week 7-8: Optimising and tweaking the hyperparameters of our implementation based on the results from the data. Preparing the presentation. week 9-10: Summarising the results and finishing the report. Maybe implement additional features depending on time and necessity We hope to have a functioning algorithm ideally before December 21st, but at the latest after the holidays on January 7th. That way, we can focus on running experiments and optimizing the hyperparameters for the rest of the course.

References

[1] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.

[2] Felix Grä, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 International Conference on Digital Health*, DH '18, pages 121–125, New York, NY, USA, 2018. ACM.

[3] Surya Kallumadi and Felix Gräßer. Drug review dataset (druglib.com) data set, 10 2018.

[4] Surya Kallumadi and Felix Gräßer. Drug review dataset (drugs.com) data set, 10 2018.