Group Project Proposal

CS7CS4/CSU44061 Machine Learning

Andrej Liadov: 15324730

Isobel Mahon: 17331358

Caroline Liu: 16327123

Motivation: What problem are you tackling?

The main problem we will tackle will be to provide structure to unstructured text data. If text data is organised into categories, it may help the efficiency and accuracy of search engines. It might also provide a framework for automatically labeling datasets for further machine learning research.

Dataset: What data will you use and how will you collect it?

We intend to use the WikiMedia API as our dataset. WikiMedia has a number of categories that can be used as labels, as well as text that can be used to create feature vectors for our model. The WikiMedia API provides us with relatively easy access to this data, which we will then need to clean. The WikiMedia API provides an endpoint that will give us the 500 most recent articles in a category, which we can use to get labeled data. If 500 turns out to be too small, we can call the endpoint more than once.

Method: What machine learning techniques are you planning to apply or improve upon?

We aim to use support vector machines (SVM) with one-hot encoding and term frequency (tf) from the Python sklearn library on our data for the project.

Term frequency is counting up the number of times a specific word comes up in a sentence, allowing vectorisation of the sentence. One-hot encoding is a method which is used to preprocess categorical data to be more expressive and allow them to be processed by machine learning algorithms for better prediction accuracy. This is done by splitting the columns of the categorical data into multiple columns where the numbers are replaced by 1s or 0s depending on the value of the column, where only 1 bit would be 1 (hot) for each instance. Then by using SVMs, we would be able to label the processed text to a specific category.

Intended experiments: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

We plan to evaluate how accurately our classifiers will predict unseen data into predefined categories. Wikipedia contains millions of articles in English that are already categorised. We will split the data we scraped into training and test data. The split ratio will be 80:20 respectively.

We will compare our trained model against a baseline model. We will try to compare our trained model with an existing model in established literature. We will evaluate how our trained classifier deals with text data from websites other than Wikipedia. A comparison will be made between one-hot encoding and term frequency vectorisation.