**Machine Learning - CS 7641 – Assignment 1**

Jeremy Martinez

Computer Science Dept., Georgia Institute of Technology

Abstract

Two common applications of machine learning, among many, are the ability to process natural language and improve health care. Natural Language Processing (NLP) and Medical data are two subjects this paper will explore. The data sets will examine yelp restaurant reviews and biopsies of cervical cancer, each of which have binary classifications. The data will be preprocessed and then run through a series of common machine learning algorithms to compare and contrast their performance on classifying these types of datasets. From this, we hope to gain an understanding of how these algorithms behave when presented with datasets that possess certain characteristics

**The Data**

**Restaurant Reviews – Natural Language Processing**

Natural Language Processing is a field that analyzes text in an attempt to classify it according to some set of hypotheses. This subsection of machine learning has received an immense amount of attention and is therefore an area where one can find an extensive amount of research that has already been conducted. It has a practical application for many businesses and as tools around NLP continue to grow, more companies will begin to explore it for their own use cases. It’s applicability in the real world makes it an interesting topic to explore for machine learning and when combined with other tools like voice recognition software, can accomplish impressive engineering feats.

NLP takes a large body of text and quantifies it into a “bag” of words. Analysis of this bag of words can be approached in different ways. The most common of which are either classifying each individual word as a feature in and of itself. Another would be to group words together in segments as to account for word order. This can play a significant role in instances where negation needs to be taken into account. Consider the restaurant review: “this meal was not very good.” This sentence, if quantifying each word individually, can be interpreted as a positive review if the word “not” is not weighted properly. However, if grouping these words in such a way that we label “not very good” as a standalone feature, it is clear that this is a negative review. There are several ways to approach preprocessing data in NLP, however, these are two of the most straightforward.

Our data for this project consists of yelp reviews that we have retrieved from a yelp database. Our data is in plain English, is purely categorical, has a binary classification, evenly distributed (50% positive classification, 50% negative), a maximum feature set of 1500, and has 1000 rows of data. When considering all of these features, this is quite an interesting dataset and one we can generate some powerful models from

It is important for your data to be raw enough to intrigue, though not too intimidating to extract some classification from. The fact that all reviews are in a consistent language makes classifying this data not too overwhelming. The data isn’t trying to do too much, which is really what we are going for here. It is purely categorical, which again, helps narrow the focus of the problem at hand (it also contrasts our next dataset in this respect). It is evenly distributed which is important to consider when analyzing the accuracy, precision and recall. It has a wide range of features, which is where it contrasts our other dataset the most. With this, we will get to see how certain algorithms handle large feature sets. This is a major contributing factor to the success rate of some machine learning algorithms and will offer a nice juxtaposition when comparing alongside the performance of our other dataset. The only downside to our dataset is that it has only 1000 rows. This is rather low, especially when considering how many features we have to compare to. It would be ideal to have a dataset exactly like this one, which more entries.

**Risk Factors of Cervical Cancer – Categorical and Discrete Data**

Medical care is field that, relatively speaking, is in earlier stages of artificial intelligence. This is in large part due to that fact that the medical industry, technologically speaking, moves slower than most. It could also be due to the sensitivity of medical data due to HIPAA laws. As artificial intelligence grows, it will transition from a technology revered as experimental, to a standard. The nature of medical care and the effect it has one people’s lives alone is what makes this topic so interesting. Few go into software engineering with the thought that they would have the ability to save someone’s life through health care. Now with how far ML/AI has come, that is a very real possibility.

Medical data can come in all shapes and sizes. Monitoring heart patterns is a real-world application for continuous classifiers. While predictive modeling of future health patterns based on previous health records almost always consists of the combination of categorical and continuous features. With our dataset in this project, we will be dealing with the latter.

Our data consists of health records, containing both discrete and categorical data features. The data determines, based on a limited health record, if a patient is a candidate for a positive biopsy of cervical cancer. Therefore, it is a binary classification, it is not evenly distributed (93.5% negative classification, 6.5% position), has a feature set of 46, and has 858 rows of data. This largely differs from our other dataset and will shed light on interesting strengths across algorithms.

Immediately, the distribution of this dataset stands out. While most would consider this distribution poor, it introduces some interesting challenges. It is also important to note that in this medical industry, it is not uncommon to get data with this kind of distribution. It is reflective of the real population. For every 10 patients diagnosed with cervical cancer that you document health records of, you will undoubtedly have had to go through 200 to find. Due to the nature of this problem, we are forced to improvise. Also, the data has features that are both numeric, as well as categorical. This will be important to consider when analyzing the performance of each algorithm. The feature set is much smaller than our NLP problem (1454 less). It is also in a manageable range with respect to rows of data.

**Comparing Data Sets**

These two-dataset approach two very different issues, which will be intriguing if you are either interested in medical care or natural language processing. However, the implementation and performance of our data preprocessing and fitting our classifiers will vary greatly as well. This will be interesting to compare and contrast and will shed light on when and where to use certain algorithms to solve certain problems. One things that would be nice to have in this project is one dataset that is purely continuous. Having all three of these types of datasets, purely categorical, purely continuous, and a mixture of both, would be ideal when making these comparisons.