CS5785/ORIE5750/ECE5414 Project - Proposal

Thomas Berry, Jake Sauter, Nicholas Bartelo, Mingxuan Zhang

TOTAL POINTS

5 / 5

QUESTION 1

1 Motivation 1 / 1

✓ **- 0 pts** **Good**

- 0.5 pts Lacking details

QUESTION 2

2 Method 3 / 3

✓ **- 0 pts** **Good**

- 1 pts Lacking details

- 1 pts Need depth

Sounds good! This may require significant compute, so make sure you have access to it.

QUESTION 3

3 Future work 1 / 1

✓ **- 0 pts** **Good**

- 0.5 pts Lacking details

Great

Raw Data vs Expert Opinion: Predicting Mortality, Hospital Stay Duration, and/or readmission from Mutually Exclusive Numerical Data and Multi-Domain NLP Data

Our team is composed of Nicholas Bartelo (nab273), Jake Sauter (jns4001), Mingxuan Zhang (mz587), Tom Berry (tb564)

Motivation

Predictive medicine is a growing part of the medical field which attempts to predict the probability of the development of a disease for a patient based on a comparison between his/her health record and many previous data records. One of the main goals of predictive medicine is the expansion towards early disease detection, or complete prevention of the disease in a patient. Data driven algorithms using large amounts of patient information have the potential to add insight to some medical problems. Using the results of numerous other patients, it may be possible to use the similarities between pieces of data to create a model which correctly identifies potential health problems of a certain individual. Given both numerical and textual data, we want to create an accurate model which predicts patient outcomes.

Methods

To explore the possible combination of different modalities, we would like to first train without-text models and text-only models separately. For text-only models, we will be using various architectures such as LSTM-RNN, Transformers and Bidirectional transformers(BERT). For without-text models, we will be using regressors for numerical prediction and classifiers for categorical prediction, all with different architectures. For regressors, we will be using a wide range of models including ridge/lasso regression, neural network regression, and supporting vector regression. Similarly, for classifiers we will be using naive bayes classifiers, supporting vector machines, logistic regression, and neural network classifiers. To combine modalities, we will be combining high dimensional embeddings of text and non-text data and applying softmax to obtain the collective

1 Motivation 1 / 1

✓ **- 0 pts** **Good**

- 0.5 pts Lacking details

Raw Data vs Expert Opinion: Predicting Mortality, Hospital Stay Duration, and/or readmission from Mutually Exclusive Numerical Data and Multi-Domain NLP Data

Our team is composed of Nicholas Bartelo (nab273), Jake Sauter (jns4001), Mingxuan Zhang (mz587), Tom Berry (tb564)

Motivation

Predictive medicine is a growing part of the medical field which attempts to predict the probability of the development of a disease for a patient based on a comparison between his/her health record and many previous data records. One of the main goals of predictive medicine is the expansion towards early disease detection, or complete prevention of the disease in a patient. Data driven algorithms using large amounts of patient information have the potential to add insight to some medical problems. Using the results of numerous other patients, it may be possible to use the similarities between pieces of data to create a model which correctly identifies potential health problems of a certain individual. Given both numerical and textual data, we want to create an accurate model which predicts patient outcomes.

Methods

To explore the possible combination of different modalities, we would like to first train without-text models and text-only models separately. For text-only models, we will be using various architectures such as LSTM-RNN, Transformers and Bidirectional transformers(BERT). For without-text models, we will be using regressors for numerical prediction and classifiers for categorical prediction, all with different architectures. For regressors, we will be using a wide range of models including ridge/lasso regression, neural network regression, and supporting vector regression. Similarly, for classifiers we will be using naive bayes classifiers, supporting vector machines, logistic regression, and neural network classifiers. To combine modalities, we will be combining high dimensional embeddings of text and non-text data and applying softmax to obtain the collective

2 Method 3 / 3

✓ **- 0 pts** **Good**

- 1 pts Lacking details

- 1 pts Need depth

Sounds good! This may require significant compute, so make sure you have access to it.

embeddings. A LSTM-RNN or neural network will be trained on these collective embeddings and produce the final prediction.

Future Work

We want to work on multi-modal approaches involving both text and numerical data and compare with current best model validation accuracies to prove that expert opinion provides non-redundant, exploitable information in the task of disease treatment best tactic generalization.

To understand text-only models, we will analyze computational properties of language models such as specificity, precision and long range mutual information scaling rules. We will also explore the effect of different tokenization and segmentation algorithms on different types of text data.

For without-text models, we will analyze prediction performances with metrics such as area under ROC, F1 score, and Matthew's correlation coefficient(MCC). We will also analyze feature importance with techniques such as partial dependence plots(PDP) and confusion matrices.

For mult-modal models, we will be evaluating our prediction performances against the expert opinions given in the dataset. To understand how well the modalities are getting combined, we will be looking at the differences between weighted/unweighted contributions of different modalities. We will also look at different combination methods such as additive combination, multiplicative combination and simple concatenation.

3 Future work 1 / 1

✓ **- 0 pts** **Good**

- 0.5 pts Lacking details

Great