HW6 Proposal

RJ Cass

1. Understanding the Problem

Background:

An important aspect of a company's brand is how well they respond to a client's complaint, both in terms of speed and accuracy. Increasing the speed in which a client's complaint gets directed to the correct department allows for quicker resolution of their issues and a better perception from the client. To this end, we have a dataset of complaints submitted to companies (the raw text of each complaint), as well as which department those complaints ended up being sent to. In this analysis we want to use these data to create a model that will allow us to quickly assign a complaint to the correct department just based on the contents of the complaint.

Goals:

In this analysis, we want to address several key questions:

i. How accurately are you able to classify complaints?

- To answer this we will create a predictive model using the given data. We will then use this model to categorize the available complaints and see how many of them match the actual department they were sent to.

ii. If your chosen method allow for it, what key words or symbols or explanatory variables were useful in classifying complaints?

- To answer this we will, if possible, pass our generated model through a system to determine variable importance.

iii. Are there departments that are commonly confused? That is, do you commonly classify complaints as going to department "A" when it should go to department "B"?

- To answer this we will review the categorization of each complaint and see, for those which were incorrect, if there are any trends about how they get categorized.

iv. For the supplied newer complaints, what department(s) does your model think they should be directed to?

- To answer this we will use our model to produce predicted categorizations of these complaints.

2. Exploratory Data Analysis

THe most important feature of this dataset is that there are no actual explanatory factors: all we are given is the raw text of each complain. As such we will need to create our own 'tokens' (variables composed of features of the text) that attempt to capture all the relevant information from a complaint that is necessary for categorization. Wihtout this, there is no data on which to build a model. Furthermore, the output variable (Department) has multiple (>2) values so any model used will need to be able to assign a complaint into any of those groups, otherwise complaints will never be assigned to certain departments.

3. Desired Attributes

Model attributes required from the research questions

The research questions require a model that is able to categorize complaints (predict). They also require a model that, if possible, is able to identify the most important factors.

Model attributes required from the data

The data require that our model include tokenizations of the available text (converting it to some numerical/categorical representation).

Any other anticipated problems

This model will be highly dependent on the tokenization performed on the data. The variables will only consist of ways we are creative enough to create tokens for the data, and we may end up missing an important feature of the responses.

What goes wrong if the above are not accounted for?

While there are no inherent assumptions, the primary risk comes from not tokenizing the data effectively which could result in a low predictive power.

4. Proposed Method

Appropriate models

Models appropriate for this analysis include a decision trees (BART, etc.) and Neural nets.

Specific model proposal

For this analysis, we propose using a neural net. These models do not have a specific mathematical representation, but can be thought of as a multi-layer regression model where each layer is allowed to identify it's own ideal relationship between input and output factors.

Method strengths/weaknesses

i. How this method accounts for issues in the dataset

This model does not have any model assumptions, and is able to take the tokenized data we provide and produce a categorization.

ii. How this method accomplishes research/analysis goals and yields appropriate estimators

This model predicts and will provide indicators for each data point as to which category it belongs to. It is also possible to get variable importance.

iii. How this method will answer the research questions

This model allows for prediction and will allow us to test how well it categorizes the departments for known complaints, as well as provide categorization for new complaints.

iv. What assumptions are needed to use the model adequately? Are they reasonable to assume and explained well?

This model does not have any inherent assumptions.