Ayush Subedi

asubedi6@gatech.edu

Gatech ID: 903743953

Final Project Proposal

Problem Statement

Mushroom foraging has been a popular recreational activity for quite some time now. Additionally, in several parts of the world, wild mushrooms are part of the diet, mostly out of necessity. Through generational knowledge transfer and documentation on various species of mushrooms, we have been able to categorize poisonous and non-poisonous mushrooms. However, news of hospitalizations and deaths due to poisonous mushroom consumption is commonplace, especially in my home country of Nepal (https://www.fungimag.com/spring-09-articles/12_Nepal.pdf). There is plenty of research that suggests there is no simple guide for mushroom edibility (https://www.wildfooduk.com/articles/how-to-tell-the-difference-between-poisonous-and-edible-mushrooms/). For example, there are counterexamples of non-edible mushrooms in several categories of mushrooms; the categories are created using features that are shared among plenty of edible mushrooms. This makes it pretty difficult to make people aware of what properties might constitute separating poisonous mushrooms and edible mushrooms.

Through this project, I will try to answer the following:

- What is the best model for the detection of the edibility of a mushroom?
- What is the best model with the least number of features when detecting the edibility of a mushroom?
- What is the best model with the least number of features that gives the least false positive rate when detecting the edibility of a mushroom?

Additionally, to approach this problem, I am planning to do the following to keep the spirit of our course:

- Read academic papers on topics relevant to this (Example 1: Does outlier detection when all features are categorical make sense?, What are some approaches? What does the math look like? Example 2: Does PCA when all features are categorical make sense?, What are the alternatives? Why?)
- If the paper provides pseudocode for solving the problem that is not available in an "established" package (like scikit-learn), replicate the pseudo code to python code and implement it to the dataset. (for example, CorEx for dimensionality reduction when all features are categorical)
- Finally, for all models used I also plan on deep-diving into mathematics.

Dataset

The dataset is available at *UCI* Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Mushroom) and the origin of the dataset is the mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981). This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. Therefore, this becomes a binary classification problem moving forward. There are 8124 samples and 22 categorical features.

I plan on expanding the dataset first. The raw dataset has attributes in short form. Example (cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s). Additionally, I plan on doing some research on fungi terminologies to understand the problem from the domain side. Exploratory data analysis, and visualization for the purposes of understanding the dataset will follow. For example, do we have two samples where all features are the same but the outcome variable is different? I will look into missing data, unique value counts in columns, correlations between features, etc. Additionally, I will also research outlier detection in a categorical setting. One of the papers I plan to look into is Outlier Detection in Complex Categorical Data by Modelling the Feature Value Couplings (https://www.ijcai.org/Proceedings/16/Papers/272.pdf)

Methodology

Since one of the objectives of the project is to minimize the features, I will most likely use label encoding instead of one-hot encoding. However, I will also spend some time on literature review to research when one might be more applicable than the others.

Additionally, I will experiment with dimensionality reduction with Multiple Correspondance Analysis and CorEx from Greg van Steeg from the University of Southern California (https://arxiv.org/pdf/1410.7404.pdf). I will perform research on if PCA makes sense for categorical data and seek out other techniques designed specifically when all features are categorical.

Similarly, I also plan on performing unsupervised learning here (Clustering) using K-modes. The goal of clustering (apart from researching clustering in all categorical settings), is also to perform exploration on data. I will also research other models/techniques for clustering all categorical data.

For feature selection with categorical data, I intend to test forward-selection, backward elimination, Chi-squared feature selection, mutual information feature selection among others. I will focus on this part significantly (mathematics and research).

The next part is supervised learning using several classification models (single and ensemble), to understand why some algorithms might work better than the others here. Some examples of classifiers that I will use are K-nearest neighbors, Logistic Regression, Neural Nets, SVM, Random Forest Classifiers, Naive Bayes, etc. Similarly, I will focus on this part significantly (mathematics and research) as well.

Evaluation and Final Results

For evaluation, I will be looking into Accuracy, Precision, Recall AUC ROC, F1 score, Confusion matrix, etc. Additionally, I will also be looking into the bias-variance tradeoff. I will also look at execution time for different models and hyperparameters. Since the goal is to find the best model, the best model with the least number of features, and the best model with the least number of features that gives the least false positive rate, I plan on tabulating the results to get to my conclusion. Additional discussion on why certain models did better than the others will follow.

Additionally, I will also analyze and explore the features that are most important for classification. My current assumption that there is a fixed set of features that might give a considerably excellent prediction (least false positives) might be incorrect as well. If that is the case, I plan on reporting on it as well, and commenting on why that might be the case. The idea of the project is to find out the least number of features and its rules someone has to remember to avoid poisonous mushrooms. I will evaluate if my results allow for this possibility.

Finally, I will incorporate gaps in the model and its limitations. Lastly, I will have a references section.