Source

Using Supervised and Unsupervised Learning for Mushroom Classification

Prepared for: Udacity Nanodegree Program

Prepared by: Džanan Ganić

21 November 2016

EXECUTIVE SUMMARY

Domain Background

“A fungus is any member of the group of eukaryotic organisms that includes unicellular microorganisms such as yeasts and molds, as well as multicellular fungi that produce familiar fruiting forms known as mushrooms.”(Source: Wikipedia). “According to records from 1991, there are 1.5 million fungi on the Earth. Fungal habitats include soil, water, and organisms that may harbour large numbers of understudied fungi, estimated to outnumber plants by at least 6 to 1.”(Source) Mushrooms are the fleshy, spore-bearing fruiting bodies of a fungi, typically produced above ground on soil or on its food source. In this project, we want to focus on mushrooms which can be easily found while taking a stroll in the forests (the most common ones).

There exist many edible, high quality mushrooms which are rich with vitamins and minerals and have great value at market. However, some of the mushrooms are toxic, which can cause different type of health problems if consumed, and a small number of them is even deadly. Throughout this project, we want to classify the mushrooms and find out which ones are edible, and which ones are toxic.

Problem Statement

In this project, we will analyse a dataset containing mushroom information, and train a few different supervised learning algorithms in order to classify the new inputs as poisonous or edible. We will run unsupervised learning techniques in order to see what kind of trait correlation exists between the mushroom, for the sake of improving our knowledge for feature selection and transformation. In the end, we will test and tune our supervised learning algorithms, in order to see which one is giving the best performance.

Datasets and Inputs

For this project, we will be using mushroom dataset from UCI School of Information & Computer Sciences which can be found on this link. Although hypothetical, this is the most comprehensive dataset on mushrooms and their traits, and I find it ideal for this type of problem we will be tackling. The number of instances in the dataset is 8124, whereas every instance has 22 attributes. There do exist missing values which means that we will have to do data preprocessing before we start tackling the problem. It also contains the information whether a certain mushroom is poisonous or not, which is ideal for testing our trained algorithms by using train/test splitting and cross validation.

“This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). 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. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy.” (Data Set Information from the website of the dataset)

Solution Statement

After preprocessing the data, we will try running clustering to understand whether are there any correlations between toxic mushrooms, as well as between edible ones. We will do this in order to improve and strengthen our domain knowledge, and to see which features are unnecessary in order to reduce dimensionality and prepare the dataset as much as possible for supervised learning. Then, we can start modeling our supervised learning algorithms in order to successfully classify the given mushroom as poisonous or edible.

Benchmark Model

Dr. Włodzisław Duch of Nicolas Copernicus University in Poland has been able to perfectly classify this mushroom data by using a decision tree. The method Dr. Duch used is focused on finding and using the single criterion that has the fewest wrong classifications in order to separate the data, and then finding the next most effective one until he did completely classify the data. (Source)

In this project, we will use unsupervised learning to get a deep insight in trait correlation, feature selection and transformation, and then test the different supervised learning algorithms with different parameters in order to successfully classify the data, and then compare the results to the decision tree classifier Dr. Duch made.

Evaluation Metrics

In this project we will be evaluating binary classifier. In order to evaluate it, we will use accuracy score. The reason why accuracy is good choice is that in this problem is that there is not a huge imbalance in the data (there are 51.8% edible and 48.2% poisonous mushrooms). However, sole accuracy will not suffice. Some types of misclassifications can be worse than the others. For example, it is less worse to misclassify edible mushroom as poisonous one, than poisonous as edible (as we can kill someone that way). We will be using cost sensitive learning and confusion matrices in the project as it will help us observe the ratio of false positive, false negative, true positive and true negative classifications.

After running our evaluation score, we will know the success score of our algorithms, and we will be able to tune them in order to get as good as possible results.

Project Design

For this project, we will use Python 2.7 along with scikit-learn, pandas, numpy and matplot libraries. We will load the data and check whether there are any inconsistencies, or errors in recording data. After that, we will approach the data through the statistics by calculating percentages of edible/toxic mushrooms, as well as mean, median and standard deviation for different kind of traits.

Since we have quite a lot attributes, we will check whether are there any unnecessary features that can be removed. We will also do some research and then use the obtained domain knowledge to perform feature selection. By using Principal Component Analysis, we will try to see whether some features can be combined into one, and find the correlation between traits.

Then, we will run clustering algorithm to try and group the different data points in order to see whether are there any correlations between them without knowing which mushrooms are poisonous and which ones are edible.

We will pick up to three supervised learning algorithms and try to train them with the data we have. The performances of those algorithms will be compared to the results we have in the initial dataset and according to that we will see which algorithms perform best. Then we will move on to the algorithm tuning and try to tune algorithms using k fold cross validation and grid search. After training and testing again, we should be able to get the decent results in predictions.