Mushrooms Classification
Machine Learning Engineer Nanodegree

Capstone Project

Reem Khaled
November 13, 2018

Proposal

Domain Background

A mushroom, or toadstool, is the fleshy, spore-bearing fruiting body of a fungus, typically produced above ground on soil or on its food source[1]. Mushrooms are used widely in cooking recipes because they are tasty and because all types of edible mushrooms contain varying degrees of protein and fibre. They also contain B vitamins as well as a powerful antioxidant called selenium, which helps to support the immune system and prevent damage to cells and tissues.

Figure 1: *Recipe That Contains Mushrooms*[2]

The problem is there are many mushroom species that are not edible and can in fact cause stomach pains or vomiting if eaten, and in some cases could be fatal, such as the common death cap mushroom[3]. People identify poisonous mushrooms by their color, shape and other features. There are many rules and guidelines to follow when picking good mushrooms, such as avoiding mushrooms with red on the cap or stem, avoiding mushrooms with white spore print, and most importantly, not consuming any mushrooms unless being 100% sure of what they are.[4,5]

Figure 2: *Agaricus Bisporus - Edible Mushroom*[6]

Figure 3: *Conocybe Filaris - Poisonous Mushroom*[7]

Problem Statement

Mushrooms can be classified as edible (fit to be eaten) or poisonous. Eating poisonous mushrooms may cause stomach pains and other problems, or even may be fatal. Each mushroom should be observed separately to collect certain pieces of information such as the cap color, stalk shape and others. Given such information, we can identify edible mushrooms from poisonous ones.

Objective:

To classify mushrooms into two classes: edible or poisonous, using a machine learning model that can predict the category of each mushroom based on its features (information).

Datasets and Inputs

The dataset used for this project is “Mushroom Data Set” from “UCI Machine Learning Repository”[8]. It was published in 1987 and it has 8124 rows and 23 columns, each column specifies a specific aspect of a mushroom (a piece of information). Each row represents a data point: one mushroom observation. The target variable to be predicted in this problem is the “class” variable. It contains two values: p (poisonous) and e (edible). Other columns represent information about mushroom cap, stalk, gill, veil features, among with others. All the columns contains categorical variables and they will be encoded to numerical values to be able to fed them to different machine learning models, and there are some missing values.

Solution Statement

This project will use different machine learning supervised models to find the best patterns that identify mushrooms classes. The best scoring model while training will be chosen to be enhanced to get better results on predicting the class of each mushroom (whether it is edible or not).

Benchmark Model

The chosen benchmark model for this project is a simple “Logistic Regression” classifier. The mushroom dataset was introduced in Kaggle[9], and by examining different kernels I found that “Logistic Regression” is widely used to compare with. Also, the reason for choosing this specific algorithm is because it is simple and straightforward and doesn’t require any hyper-parameters to specify beforehand, and it is so popular for binary classification problems that almost every machine learning engineer has used before.

Evaluation Metrics

There are two main performance measures that will be used for this dataset: testing accuracy and type II error[10] (false negatives). Testing accuracy will ensure that future observations will be classified correctly. Considering the class “poisonous” as positive and “edible” as negative, type II error will be avoided as much as possible so there are no poisonous mushrooms that will be classified as edible, since this is the worst case of classification that someone eats a poisonous mushroom. If on the other hand a mushroom is classified as poisonous but it is not, this wont cause a lot of problems like the previous situation. The chosen model will be the model with the highest test accuracy and with the least type II error. Other performance metrics will be used as needed.

Project Design

The following steps will be used to find the best model that will accurately predict each data point class:

1. Data analysis: to know the data better: to know the number of categories in each columns, to apply some statistics measures and to find missing values (as it is already mentioned in UCI that the dataset contains some).

2. Data preprocessing: converting categorical values to numerical ones and deciding if any further modifications are needed on the dataset based on the previous step results.

3. Splitting the dataset: the data set will be split into two parts: training and testing.

4. Building the benchmark model: build the chosen benchmark model to be able to compare other models against it.

5. Building models: building different machine learning models by training different models using the training set. Models that will be used: RandomForest, Naive Bayes, SVM.

6. Evaluation: comparing the results by using the testing set to validate each model performance using different performance metrics. Also, comparing against the benchmark model.

7. Validation: enhancing the best scoring model from the previous step by using hyper-parameter tuning.

Note: some steps *may* be added if found necessary during project building.

References

[1] *Mushroom - Wikipedia.* URL: https://en.wikipedia.org/wiki/Mushroom.

[2] *Mushroom & Basil Omelette With Smashed Tomato | BBC Good Food Middle East.* URL: https://www.bbcgoodfoodme.com/recipes/mushroom-basil-omelette-smashed-tomato.

[3] *The health benefits of mushrooms | BBC Good Food.* URL: https://www.bbcgoodfood.com/howto/guide/health-benefits-mushrooms.

[4] *Identify Poisonous Mushrooms: Some Detailed Tips.* URL: https://www.mushroom-appreciation.com/identify-poisonous-mushrooms.html.

[5] *How to Tell the Difference Between Poisonous and Edible Mushrooms.* URL: https://www.wildfooduk.com/articles/how-to-tell-the-difference-between-poisonous-and-edible-mushrooms/.

[6] *All the Types of Edible Mushrooms Explained With Pictures.* URL: https://tastessence.com/types-of-edible-mushrooms.

[7] *7 of the World's Most Poisonous Mushrooms | Britannica.com.* URL: https://www.britannica.com/list/7-of-the-worlds-most-poisonous-mushrooms.

[8] *UCI Machine Learning Repository: Mushroom Data Set.* URL: https://archive.ics.uci.edu/ml/datasets/mushroom.

[9] *Mushroom Classification | Kaggle.* URL: https://www.kaggle.com/uciml/mushroom-classification/.

[10] *Type I and type II errors - Wikipedia.* URL: https://en.wikipedia.org/wiki/Type_I_and_type_II_errors.