Data Selection Proposal

The goal of this project is to build a classification model to tell whether or not a mushroom is edible given certain features of its cap, gill, stalk.

1. Dataset

I choose UCI Machine Learning's mushroom data set[1]. Although its original source is the mushroom records drawn from *The Audubon Society Field Guide to North American Mushrooms* years ago, the data is still valid since there is no dramatic mushroom evolution (:D). The dataset is sufficient for this project with 22 attributes and 8124 samples. The disadvantage of the dataset is, however, that it only contains text data without any images. This could simplify the project but lower its practicability. Therefore, making classification based on mushroom photos could be a possible improvement if I could find an image dataset.

2. Methodology

i. Data Preprocessing

The dataset has two classes: edible=e, poisonous=p, with 22 attributes.

ii. Machine Learning Model

Since it is a categorical problem, Neural Network is a possible algorithm, especially if I want to do image processing. It could handle complex nonlinear situation and give relatively accurate prediction. However, more hidden layers cause longer time to adjust weights.

iii. Final conceptualization

This project will be presented in the form of webapp.

[1] UCI Marchine Learning. *mushroom-classification*. 2017. <https://www.kaggle.com/uciml/mushroom-classification>.