UIC CS 412, Fall 2017
Prof. Xinhua Zhang

Classification of Poisonous Mushrooms

Group Members

The project group consists of Noemi Dolnik, David Liang, Megan Hauck, Anh Tran, Hengbin Li, and Jean-Philippe Douailly-Backman.

Dataset

The proposed dataset consists of mushroom classifications from the UCI repository. This dataset is comprised of 8,124 rows, each containing a separate classification and 22 attributes about the mushroom, such as information about the color, gills, and stalk. The dataset is available at the following link: https://www.kaggle.com/uciml/mushroom-classification/data.

Machine Learning Tasks

The classes for the mushrooms are poisonous and nonpoisonous. The proposed machine learning task is to determine the probability of a mushroom being poisonous given the existence of a variety of attributes.

The main machine learning task for this project is classification. The project's proposed dataset is comprised of discrete variables (detecting poisonous or nonpoisonous mushrooms), and classification is optimal for this task.

Techniques

While more advanced techniques such as logistical regression will be introduced throughout the semester and may also be employed, the earliest stages of the project will use k-fold cross-validation and naïve Bayes models for classification.

K-Fold

The k-fold technique will be used for this project to test estimation. Since the proposed dataset includes a finite number of data points, k-fold would be ideal to ensure that every data point is trained k-1 times and tested once. Different values of k will be used to determine how sensitive the data is with the increase of k.

Naïve Bayes

Another technique used for this project is naïve Bayes to construct classifiers that assume the value of a given feature is independent from any other feature. A naïve Bayes model would be suitable for this project, since classes of mushrooms will be observed using specific attributes such as color, shape, and odor, among others. Since the project will focus on a binary classification problem, that is, whether a given mushroom is poisonous or nonpoisonous, naïve Bayes would be an ideal model to calculate the conditional probability of these class values.