Capstone Project Proposal

Anthony Le

January 31, 2017

Abstract

Featuring a Kaggle Competition, tackling leaf classification has many applications. With over a half million species of plants in the world,c lassification of plants have been a hard task which can lead to duplicate indentifications. This proposal demonstrates the feasibility and dataset of classifying the ”random forest for the leaves”.

1 Introduction

Featuring a Kaggle Competition, tackling leaf classification has many applications. With over a half million species of plants in the world,c lassification of plants have been a hard task which can lead to duplicate indentifications.

1.1 Problem

The objective of this playground competition is to use binary leaf images and extracted features, including shape, margin and texture, to accurately identify 99 species of plants. Leaves, due to their volume, prevalence, and unique characteristics, are an effective means of differentiating plant species. They also provide a fun introduction to applying techniques that involve image-based features.

As a first step, try building a classifier that uses the provided pre-extracted features. Next, try creating a set of your own features. Finally, examine the errors you're making and see what you can do to improve.”

1.2 Dataset

The dataset can be found in the kaggle database section of the competition (link below). There are 1583 images of leaf specimens or 16 samples of 99 different species. Some preprocessing have been made to convert images to grey-scale against a white background. Three features are provided per images - “a shape contiguous descriptor, an interior texture histogram, and a ne-scale margin histogram” https://www.kaggle.com/c/leaf-classification/data

1.3 Solution

The solution can be found using supervised learning methods. With scikit-learn, various methodologies such as KNN, Decision Trees, Supper Vector Machines, and much more can be used for classification of leafs.

1.4 Benchmark Model

A benchmark model can be seen from the previous supervised learning Udacity class in determining house prices in Boston.

1.5 Evaluation Metrics

Evaluation can be directly made via Kaggle by uploading the kernel or results to the website. Various other evaluation metrics can be used such determining accuracy between the training and test sets and determining the R2 value.

1.6 Project Design

Project design will follow metodology from the supervised learning class offered by Udacity. This class uses the scikit-learn library and follows a format of preprocessing data, fitting data, and ultimately classifying data for our use.