Project Proposal:
Leaf Classification

Julian Teh, Bai Xin
Theories and Algorithms in Machine Learning, CS5339, AY2016/2017 Semester 2

Motivation

We plan to work on Kaggle's leaf classification problem. Since both of us are interested in computer vision/image auto-tagging problems, we think this project would be a great fit, especially since the dataset is already provided.

The problem given is the identification of the class of a leaf given an image of it. This is interesting because the identification of leaves will not only require features of the leaf surface, but also of the leaf outline, color and pattern; all of which require different features to be extracted from the leaf image. Some of these features are given, but it is certain that to improve on initial result, better feature engineering will have to be done.

Dataset

The dataset we will use is the one provided by the Kaggle challenge. Other than binary leaf images, three sets of features are also provided per image; a shape contiguous descriptor, an interior texture histogram, and a fine-scale margin histogram. The goal is to predict the category of a new leaf image.

Project Outline

We intend to tackle this problem in two main ways:

Feature Engineering

We will focus on transforming the training images of each class into more condensed features, such as edge detection, SIFT features or PCA. We will also explore how to combine different features to optimize the result.

Classification Model

Different classification models will be tested and evaluated, such as Logistic regression, SVM and Neural Networks. Then we will evaluate the performance of each classifier with respect to the model's classification accuracy.

Expected Outcome

Our submission will be evaluated based on multiclass log error. We expect our outcome could achieve an error within 0.05 and this corresponds to a rank of 531/1435 on the Kaggle challenge, which we will then try our best to improve on.

1. **Review completed by: dcsleews**
[dcsleews]

2. The proposal is fine. You should start with using the features that are provided. But more interesting would be to study the class of images being classified and coming up with features/methods that can improve performance of your initial simpler model using properties of the problem (and explaining why the methods you develop are helpful). *[dcsleews]*