CAPSTONE PROJECT I: PROPOSAL

February 4th, 2019

Erik Larsen
Data Science Career Track
January 6th Cohort

Overview

One problem commonly found in both agriculture and urban farming is the identification of seeds. From sorting small garden seeds when they get mixed up, to quality assurance in packaging correct seeds for sale, and even seed type and quality on large farms, the ability to discern a species of seed when its origin is unknown can be vital. This project uses machine learning to identify seeds from geometric properties alone. Future extensions can lead to plant vs weed identification and better quality crops.

Goal

Explore and evaluate various algorithms that estimate the probability of a seed belonging to a certain species.

Approach

The Seeds dataset is relatively small compared to most datasets in use today. It is comprised of only 210 examples, with three classes of wheat grain. In this project we will experiment with various machine learning algorithms (e.g.,neural networks) and compare their relative performance with respect to metrics to be defined. Cross-validation and bias/variance analyses will be performed to maximize accuracy and performance. The Seeds dataset can be found in the UCI Machine Learning Repository at seeds Data Set.

Description of the Data

Classes: Kama, Rosa, Canadian with 70 examples for each.

Attributes:

Area, perimeter, compactness, length, width, asymmetry-coefficient, and groove length

The compactness is derived from $C = 4\pi \frac{A}{P^2}$ and the origin of the asymmetry coefficient is unexplained.

Deliverables

As required, I will submit all Jupyter notebooks that I develop, a final report, and a presentation slide deck.