Ensemble Learning and Random Forests

TM Quest

Overview

What Will we Learn in This Module?

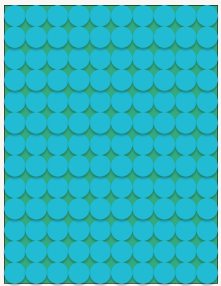
- What is ensemble learning?
- How can ideas like wisdom of the crowd and majority rule help improve our machine learning models?
- What is a random forest?
- What is parallelization and how can this help us with random forests?

Introduction to Ensamble Learning

Motivation: How many marbles are in the glass?

Example

There is a glass full of marbles you are shown for one second. *The objective is to guess how many marbles are in the glass.*

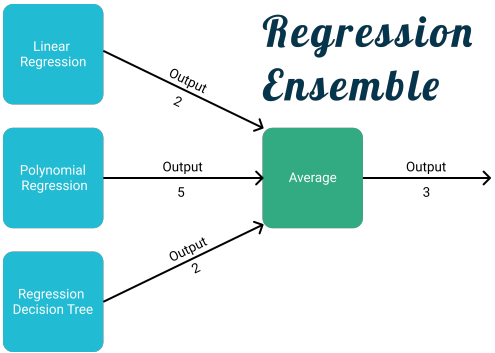


- Person 1: I guess 80!
- Person 2: I guess 150!
- Person 3: I guess 170!
- Average: $\frac{80+150+170}{3} = \frac{400}{3} \approx 133.3$
- True answer: 130. The wisdom of the crowd prevails!

Ensemble Learning

Definition

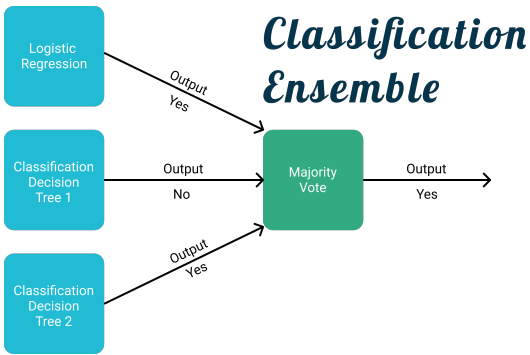
If we train and predict with multiple different models we get different answers. By combining these answers to produce a single answer, we have an ensemble model.



Classification Ensemble

Example

If we predict with several classification algorithms and pick the class with the most total votes, then this ensemble method is called majority vote or majority rule.



Weak Learners and Bagging

Weak Learners

Definition

Models that are only slightly better than a random guess are called weak learners.

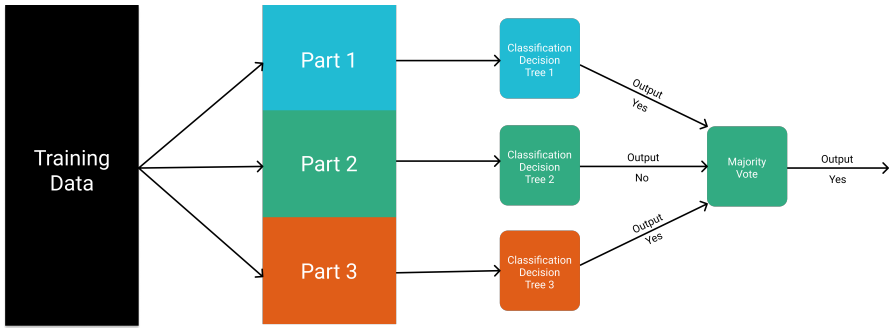
Remarks

- If you have two equal classes, then a classification model with 51% accuracy is a weak learner.
- Weak learners are not very useful by themselves. If we combine many weak learners they might be useful.
- When combining weak learners they need to be relatively independent.

The Idea Behind Bagging

Definition

You can take multiple copies of the same algorithm and train them on different subsets of the data.Combining them afterward is called bagging. So bagging is a special kind of ensemble learning.



Random Forests

Remarks

- In reality, with bagging each data point can be selected for multiple models. This is called selection with replacement. Without replacement is called pasting and is less used.
- In scikit-learn there are classes like BaggingClassifier that you can use for general bagging.
- The resulting model that comes from bagging decision trees is called a random forest. There are special classes in scikit-learn for random forests that we will use.