

Linear Classification and Logistic Regression

In machine learning, classification is a supervised method of segmenting data points into various labels or classes. Unlike regression, the target variable in a classification problem is discrete. Each data point used in training classification models must have a corresponding label in order for the characteristics and patterns in the classes to be learnt appropriately. Classification can either be binary - identifying that a given email is spam or not or, multi-class - classifying a fruit as orange, mango or banana.

All codes used in this lesson can be found [here](https://gist.github.com/HamoyeHQ/94d52ad113d1eac80d073a4affb0a490):- <https://gist.github.com/HamoyeHQ/94d52ad113d1eac80d073a4affb0a490>

Gentle Introduction

Every year people demand more from nature than it can regenerate. Individuals, communities and government leaders use ecological footprint data to better manage limited resources, reduce economic risk, and improve well-being. The Dataset provides Ecological Footprint per capita data for years 1961-2016 in global hectares (gha). Ecological Footprint is a measure of how much area of biologically productive land and water an individual, population, or activity requires to produce all the resources it consumes and to absorb the waste it generates, using prevailing technology and resource management practices. The Ecological Footprint is measured in global hectares. Since trade is global, an individual or country's Footprint tracks area from all over the world.

Apart from predicting numeric values, another important supervised machine learning method is classification and it involves predicting classes (either binary or multinomial classes). In this section, we will cover how to measure performances of class prediction, linear classification methods and non-linear/tree-based methods. We'll also focus on strategies for applying a successful classification model like interpretability-accuracy trade-off, class and imbalance.

The National Footprint and Biocapacity Accounts (NFAs) measure the ecological resource use and resource capacity of nations from 1961 to 2016. The calculations in the National Footprint and Biocapacity Accounts are primarily based on United Nations data sets, including those published by the Food and Agriculture Organization, United Nations Commodity Trade Statistics Database, and the UN Statistics Division, as well as the International Energy Agency. In this project, we will use this data to classify and predict the quality metrics (qascore) of the ecological footprint data for the different countries. This data includes total and per capita national biocapacity, the ecological footprint of consumption, the ecological footprint of production and total area in hectares.

Data Source: <https://data.world/footprint/nfa-2019-edition>