Machine Learning and AI with Python - Notes

HarvardX CS109xa – Alessio Santoro

October 25, 2024

Chapter 1

Decision Trees

1.1 Decision trees for classification

1.1.1 Decision Trees Part 1

Logistic regression is a fundamental statistical method used in machine learning for binary classification tasks. It predicts the probability of an instance belonging to a particular class.

Part A: Classification using trees You may have learned in previous courses that **logistic regression** is most effective for constructing classification boundaries when:

- The classes are well-separated in the feature space
- The classification boundary possesses a simple geometry

The **decision boundary** is determined at the point where the probability of belonging to class 1 is equal to that of class 0.

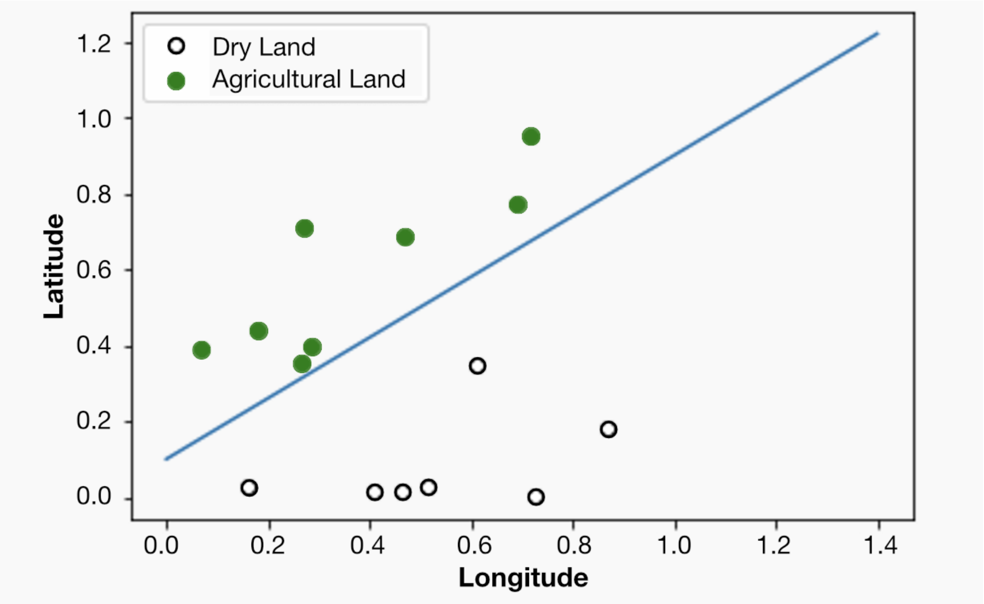
$$P(Y=1) = 1 - P(Y=0)$$

$$\rightarrow P(Y=1) = 0.5$$

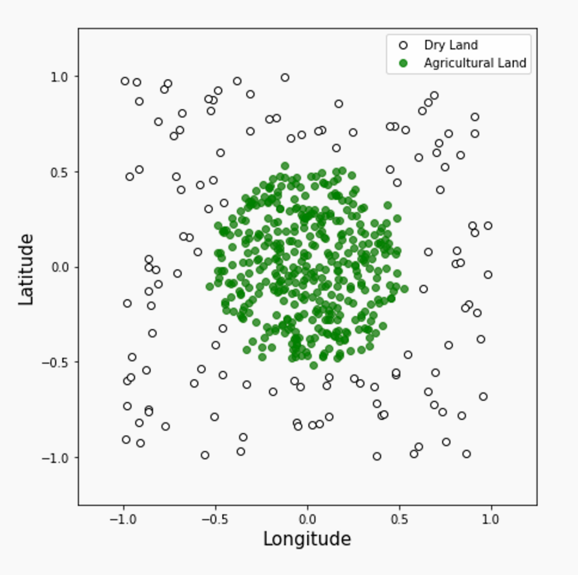
This is equivalent to the scenario where the log-odds are zero. The log-odds are defined as:

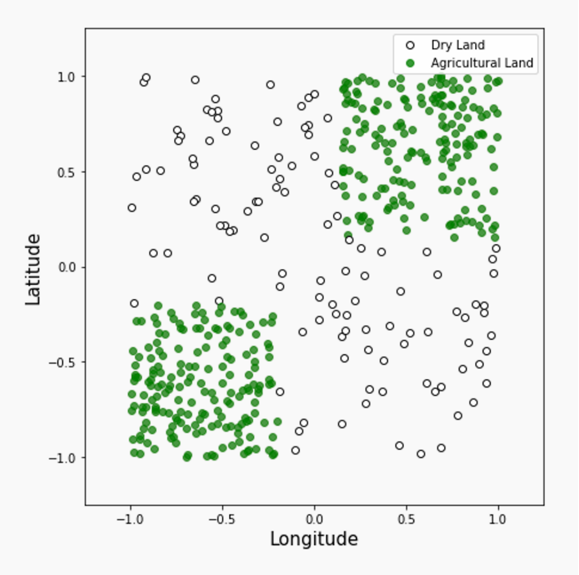
$$\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = X\beta = 0$$

The equation $X\beta = 0$ deine an hyperplane, but can be **generalized** using higher order polynomial terms to specify a non-linear boundary hyperplane.



In this case the blue line can be easily described as ($y = 0.8x + 0.1$) or $-0.8x + y = 0.1$ (assuming that, given their position as coordinates we call *longitude* as x and *latitude* as y).





In the first figure, we see that we can make a circular bounding box whereas in the second figure it is likely that we can make two square bounding boxes. However, as the geometric shapes in the figures become more complicated, determining the appropriate bounding boxes becomes increasingly less straightforward and more complex.