Data Science 101

A predictive model on iris plant using machine learning

Content

1. Background study of the iris plant.
2. Data Visualizations/ statistics on iris.
3. Data Preprocessing and cleaning.
4. Data Transformation/feature engineering.
5. Predictive Modelling.

Background study of the iris plant

Source: here

Background study

Iris is a genus of 260–300 species of flowering plants with showy flowers. In Biology, its a flower with the following:

1. Kingdom - plantae
2. *Clade*: Tracheophytes
3. *Clade*: Angiosperms
4. *Clade*: Monocots
5. Order: Asparagales
6. Family: Iridaceae
7. Subfamily: Iridoideae
8. Tribe: Irideae
9. Genus: *Iris*

Data set

Iris virginica

Iris setosa

Iris versicolor

Data Set

Data Set Characteristics:

- Number of Instances: 150 (50 in each of three classes)
- Number of Attributes: 4 numeric, predictive attributes and the class
- Attribute Information: sepal length in (cm),sepal width in (cm),petal length in (cm),petal width in (cm)

Class: Iris-Setosa,Iris-Versicolour, Iris-Virginica

Statistics

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\text{Median} = l + \frac{\left(\frac{N}{2} - m\right)}{f} \times c$$

mean

$$Mode = l + h\left(\frac{f_m - f_1}{2f_m - f_1 - f_2}\right)$$

Where,

l = Lower Boundary of modal class

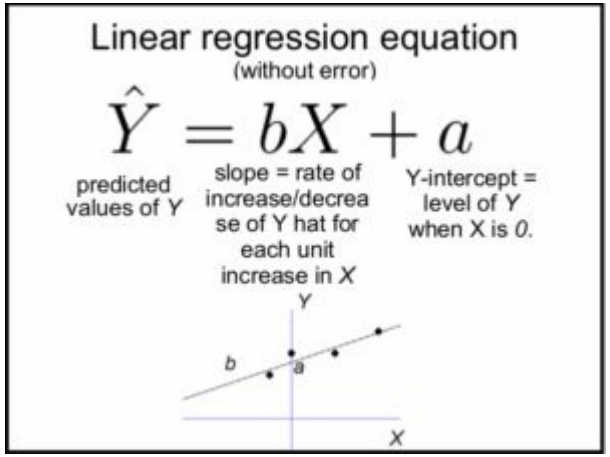
h = *size of model class*

f_m = *Frequency corresponding to modal class*

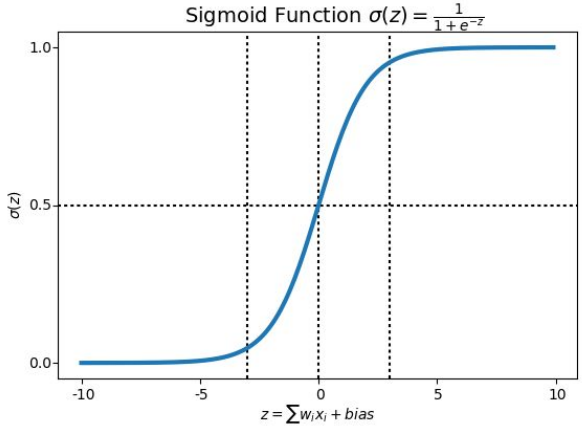
f_1 = *Frequency preceding to modal class*

f_2 = *Frequency proceeding to modal class*

Logistic Regression



$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$



Logistic Regression

Thank you