Assignment DA-3

Title: BigMart Sales Analysis

Problem Statement
BigMart Sales Analysis - For data comprising of transaction records of a sales store. The data has 8253 rows of 12 columns variables. Predict the sales of the store. Sample Test Data set is available in the given link.

Objective
To learn Big sales analysis

Outcome
To be able to analyse sales dataset

Software and hardware requirements
OS: 64-bit Windows 10 / Ubuntu 20
Programming language: Python 3

Theory related concepts

Regression
Used to analyse the relationship between multiple / single independent variables and a dependent variable.
In machine learning, regression is used for formulating a predictor or hypothesis which maps the independent variables to dependent variables.

$h: X \rightarrow Y$ predicted truth values

↓

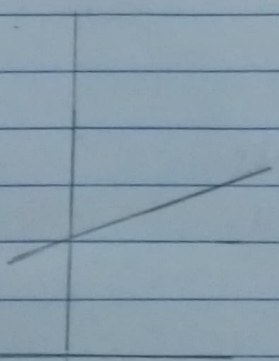
set of independent feature

It is to be noted that the variables involved here belong to real valued numbers

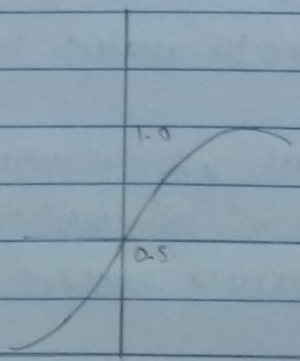
$$H_{reg} = \{x \rightarrow \langle w, x \rangle + b : w \in \mathbb{R}^d \quad b \in \mathbb{R}\}$$

$$H_{sig} = \{x \rightarrow \phi_{sig}(\langle w, x \rangle) : w \in \mathbb{R}^d\}$$

Above equations represents the hypothesis class for linear and logistic regression respectively.



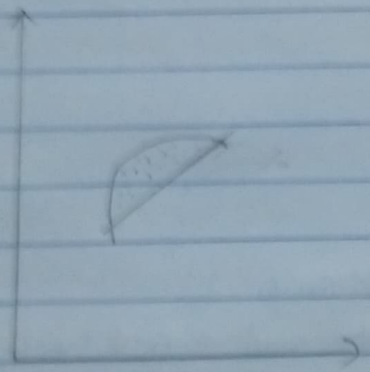
Graph for linear regression



Graph for logistic regression

Regression

- linear
- logistic
- ridge
  - eases collinearity in between predictors
- lasso
  - used to perform both variable selection and regularisation, reduces dimensionality, also called L1 regularisation
- polynomial
  - fit model to non-linear data represented by equation $y = b_0 + b_1 x_1 + b_2 x_2^2 + \ldots$

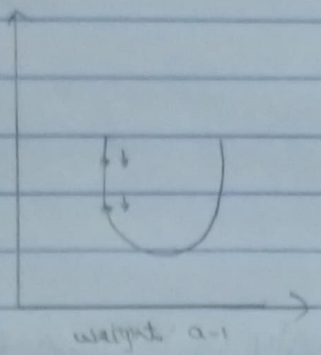


Graph for polynomial regression

The variation weights associated with the model are determined using the cost fn

$$J = \frac{1}{n} \sum_{i=1}^{n} (predi - y_i)^2$$

Optimisation J can be done using Gradient Descent. In GD, we vary the values of the weights in a direction of minimizing cost fn.



Graph of variation of J w.r.t weight a-1

As J is optimised the algorithm is said to converge i.e. the loss is iteratively reduces
For purposes of efficiency and feasibility it is necessary that algorithm converges in a finite no. of discrete steps. At every step, the amount of movement towards minimum loss is determined by learning rate

Algorithm

1. Import libraries
2. Load dataset
3. Perform preprocessing
   - change dtype 'object' to category
   - observe outliers using bar plots and remove such records.
   - identify correlation among numerical features
   - drop features with low correlation
   - handle null values by substituting null values with "mean values" for numerical dtype and "most frequent" values for categorised dtype
   - perform encoding for categorised data using label encoding
4. Fit processed dataset onto model and perform prediction.

Conclusion

Thus, we successfully performed regression analysis and generated hypothesis for prediction of sales.