Linear Regression

Nehal Khosla, Priyanshu Parida

National Institute of Science Education and Research

January 30, 2023

Part I: Linear Regression: Problem & Solution

Part II: Linear Regression: The Inductive Bias

Part III: Linear Regression: Applications and Shortcomings

Part I

Linear Regression: The Problem

Regression

- Regression is a statistical method that attempts to predict the strength and nature of the relationship between a dependent variable, and one or a series of independent variables.
- It does so by finding a curve that minimizes the error in the actual and expected values of the dependent variable of training data-set.
- For proper interpretation of regression, several assumptions (inductive bias) about the data and the model must hold.
- Linear regression is one of the common forms of this method. It establishes a linear relationship between the two variables.

Linear Regression

Introduction

- Linear regression is a **supervised** machine learning algorithm.
- The model tries to find a **best-fit line** to establish a linear relationship between the dependent (y) and independent(x) variable.
- The model then uses this fit to predict the appropriate y-values for unknown x-values.
- The **best-fit line** is achieved minimizing the error between predicted and actual values. This is done by minimizing the **loss function**.

Linear Regression

Regression Line

The line showing the linear relationship between the dependent and independent variable is called a **regression line**. An example of a regression line is shown below:

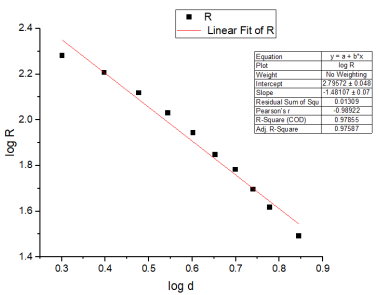

Figure. Regression line for log R (dependent variable) vs log d (independent variable)

The regression line may be positive, wherein the dependent variable increases with increase in independent variable; or it may be negative, wherein the dependent variable decreases with increase in independent variable (as in above figure).

Linear Regression: The Problem

Types

Linear regression may be classified further into the following two types:

- **Simple Linear Regression:** Assumes a linear relationship between a single independent variable and a dependent variable.
- **Multiple Linear Regression:** Assumes a linear relationship between two or more independent variables and a dependent variable.

Linear Regression: The Problem

Mathematical Representation

Once a linear relationship has been determined by the algorithm, the general form of each model may be represented as follows:

- **Simple Linear Regression**
  y = ax + b + u
- **Multiple Linear Regression**
  Y = a_1x + a_2x +a_3x + b + u

where:
y = Dependent variable
x = Independent variable
a = Slope(s) of the variable(s)
b = The y-intercept
u = The regression residual/error term

Linear Regression: The Problem

Loss Function: Mean Squared Error

- The regression line is achieved by minimizing the sum of mean squared error (loss function) for all points in the domain. The loss function is gives as:

$$\boxed{MSE = \frac{1}{N}\Sigma(y - f(x))^2}$$

where f(x) = $a_1 x + a_2 x$ .... +b

Linear Regression: The Solution

Solution

The best fit line may be found in the following two manners:

- **Closed form (Exact form) Solution:**
  - It solves the problem in terms of simple functions and mathematical operators.
  - The closed form solution for linear regression is as follows:

$$\boxed{B = (X'X)^{-1}X'Y}$$

    where B = Matrix of regression parameters
    X = Matrix of X values
    X′ = Transpose of X
    Y = Matrix of Y values
  - Although this method gives an accurate model, it is computationally expensive, especially when there are more than 4 dimensions.
- **Gradient Descent:**
  - It is used to minimize MSE by calculating the gradient of the loss function.
  - It is an iterative optimization algorithm.

Linear Regression: The Solution

Quality of Fit

- The goodness of the fit achieved determines how linearly the variables are correlated.
- The goodness of fit may be calculated using the Pearson correlation coefficient, which is given by:

$$\boxed{r = \frac{\Sigma(x_1 - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \Sigma(y_i - \overline{y})^2}}}$$

- The higher the value of r, the better is the fit.

Part II

Linear Regression: The Inductive Bias

Inductive Bias

A List

Linear regression takes the following assumptions, or inductive biases:

- The assumption that the dependent and independent variables are linearly related.
- Homoscedasticity: The assumption that the error term should be the same for all points.
- The assumption that MSE is the most appropriate loss function for linear regression.

Choice of Loss Function

Let us analyse some loss functions to justify the choice of MSE as an appropriate loss function.

- L_1 **= (y-f(x)):** This loss function gives out both positive and negative values, which cancel out to give near zero error for large data.
- L_2 **= |(y-f(x))|:** Although errors do not cancel out here, the outliers are penalised equally as compared to standard data.
- L_3 **= (y-f(x))^2:** In this case, the errors do not cancel out. Also, outliers are penalised more, giving a more appropriate regression line.

Hence, MSE is an appropriate choice for loss function.

Part III

Linear Regression: Applications and Shortcomings

Applications and Shortcomings

- Linear regression finds its applications in several fields, like market analysis, financial analysis, environmental health, and medicine.
- However, it does leave somethings to desire for. A linear correlation does not indicate causation, i.e. a connection between two variable does not imply that one causes the other.
- Linear regression is prone to noise and overfitting.
- It is prone to multicollinearity, i.e. occurence of correlation between two or more independent variables. This reduces the statistical significance of an independent variable.

References

1. CS460 Machine Learning 2023 Lectures, Subhankar Mishra.
2. Linear Regression in Machine Learning, Javatpoint.
3. ML | Linear Regression, geeksforgeeks.