papergrid

Date / /

LAB 4

Linear Regression

* It computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation with observed data.
* It predicts the continuous output variable based on the independent input variable
* To find $y = mx + c$ (minimize the error b/w predicted value Y and actual value

Algorithm for Linear Regression

* Given a dataset with n samples, each having an input x and output y
  $y = \beta_0 + \beta_1 x + \varepsilon$ that minimizes the error ε
* Initialize Parameters:
  Start with random value for β_0 and β_1 (slope)
* Hypothesis Function:
  Single feature : $\hat{y}_i = \beta_0 + \beta_1 x_i$
  multiple feature: $\hat{y} = \beta_0 + X\beta$
* Calculate mean squared Error

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (\bar{y}_i - y_i)^2$$

* Make Prediction
  Use trained model $- y = \beta_0 + \beta_1 x$ to predict output value for new inputs

$N \rightarrow$ observation
$\bar{Y}_i \rightarrow$ predicted values
$y_i \rightarrow$ actual values

Applications

- Marketing and Advertising
- Business and Economics
- Environmental Science
- Sports Analytics

Pseudocode

```
Function LinearRegression(x, y):
  Step 1: Add column of ones to x for
    the intercept term
  X = AddColumnOfOnes(x)
  Step 2: Compute the coefficients using
    OLS formula
    beta = (X^T * X)^-1 * (X^T * y)
    X_transpose = Transpose(X)
    XTX = Multiply(X_transpose, X)
    XTX_inverse = Inverse(XTX)
    XTY = Multiply(X_transpose, y)
    beta = Multiply(XTX_inverse, XTY)
  Step 3: Return coefficients
    Return beta
```

~~Multi Linear Regression~~

~~involves more than one independent variable~~
~~and one dependent variable.~~

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_n x_n$$

y is dependent variable

$x_1, x_2 \dots x_n$ are independent variables

β_0 is intercept

$\beta_1, \beta_2 \dots \beta_n$ are the slopes

Let datapoints be $(x_1, x_2 \dots x_n, y_i)$

$\forall i \{0 \dots m\}$ where $x_i \, \forall i \in \{0, 0\}$

papergrid

Date: / /

represents independent variables and y values are dependent variables

In the form of matrix

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 + x_{11} + x_{21} + \cdots x_{n1} \\ 1 + x_{12} + \cdots x_{n2} \\ \vdots \\ 1 + x_{1n} + \cdots x_{nn} \end{bmatrix}$$

where $\beta = ((X^T X)^{-1} X^T) Y$

The above value can be used to plot the best fit line and can be used to predict future values.

Logistic Regression

* Logistic regression approach operates on sigmoid curve rather than best fit line. we get a value $\in [0, 1]$ (Binary classification) and then classify into +ve or -ve by comparing with median
let data points be $(x_i, y_i) \forall i \in \{0, n\}$
Finding best fit line through previously mentioned methods

9 21/4

$$v = \frac{1}{1+e^{-(mx+c)}} = \frac{1}{1+e^{-(b_1 m + b_0)}}$$

Classification will be based on the obtained value v

* If $v < 0.5 \rightarrow$ then "no"
* If $v > 0.5 \rightarrow$ then "Yes"