Regression

Here we have the types of classification algorithms in Machine Learning:

Linear Classifiers: Logistic Regression, Naive Bayes Classifier

Nearest Neighbor

Support Vector Machines

Decision Trees

Boosted Trees

Random Forest

Neural Network

Simple linear regression

Simple linear regression

Simple linear regression:

1.It is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

2.The process is pretty straightforward. We will simply try to come up with a straight line that will best fit our data.

|  |  |
| --- | --- |
| Gradient | It is the slop of the line. |
| x | One variable, denoted x, is regarded as the predictor, explanatory, or independent variable |
| y | The other variable, denoted y, is regarded as the response, outcome, or dependent variable |
| Vertical off set | It is the difference between the actual data and the predicted straight line model. |
|  |  |

Our algorithm takes the squared sum of all these vertical offsets, i.e differences, and comes up with the line for which this sum is the minimum. This is the crux of our algorithm.

Assumptions in Linear Regression:

1.Linaer

2.Homoscendastically

3.Multi variant normality

4.Independence of errors

5.Lack of multicollinearity

*Multiple Linear Regression*

|  |  |
| --- | --- |
| Simple | y=+\* |
| Multiple | y=+\*+\*+\*+............+\* |
| Polynomial | y=+\*+\*+\*+............+\* |

The multiple linear regression explains the relationship between one continuous dependent variable (y) and two or more independent variables (x1, x2, x3… etc).

Note that it says CONTINUOUS dependant variable. Since y is the sum of beta, beta1 x1, beta2 x2 etc etc, the resulting y will be a number, a continuous variable, instead of a “yes”, “no” answer (categorical).

or example, with linear regression, I would be trying to find out how much Decibels of noise is being produced, and not if it’s noisy or not (Noisy | Not).

To find categorical variables (e.g “yes” or “no”, “1” or “0”), logistic regression would be used.

Eg:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| y |  | \* | \* | \* | ? |
| Profit |  | R&D spent | Admini | Marketing | State  (Categorical Value) |

Adding Dummy Variable(D):

|  |  |
| --- | --- |
| New York | California |
| 1 | 0 |
| 0 | 1 |
| 0 | 1 |
| 1 | 0 |
| 0 | 1 |

Eg:

0=1-0

1=1-0

y=+\*+\*+\*+

0

1

Dummy variable Trap:

If we have n dummy variables, add (n-1) variables.

*DecisionTree*

CART

Classification Regression

Decision tree:

A decision tree is drawn upside down with its root at the top.

*Random Forest*

It is a type of Ensemble Learning

Steps:

1.Pick at random k data point from the traning set

2.Build the decision tree associated to these k data points

3.Choose the number N tree of trees you want to build and repeat the steps 1 and 2

4.For a new data point ,make each one of your N tree predict the value of Y to the data point in question and assign the new data point .The accurate all of the predicated Y values

*Evaluting Regression Models*

R-Squard

Adjusted R-squared