Data Mining and Data Warehousing 10

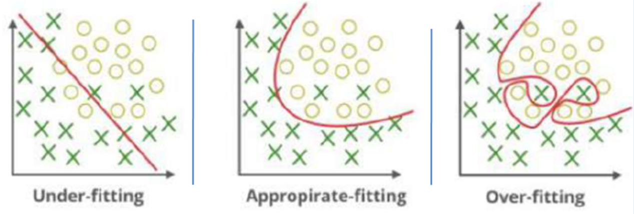
Prediction

Chittaranjan Pradhan
School of Computer Engineering,
KIIT University

Overfitting & Underfitting

Underfitting

- *When a model hasn't learned the patterns in the training data well and is unable to generalize well on the new data, it is known as underfitting*
- An underfitting model has poor performance on the training data and will result in unreliable predictions
- Underfitting occurs due to high bias and low variance
- **Techniques to reduce Underfitting**
  - Increase model complexity
  - Increase the number of features in the dataset
  - Reduce noise in the data
  - Increase the duration of training the data



Overfitting & Underfitting

Overfitting

- *When a model performs very well for training data, but has poor performance with test data, it is known as overfitting*
- Here, the machine learning model learns the details and noise in the training data such that it negatively affects the performance of the model on test data
- Overfitting can happen due to low bias and high variance
- **Techniques to reduce Overfitting**
  - Using K-fold cross-validation
  - Reduce model complexity
  - Training model with sufficient data
  - Remove unnecessary or irrelevant features
  - Early stopping during the training phase
- Overfitting is the result of using an excessively complicated model while underfitting is the result of using an excessively simple model or using few training samples
- A model that is underfit will have high training and high testing error while an overfit model will have extremely low training error but a high testing error

Prediction

Classification vs. Prediction

- Both problems deal with the mapping of the input data to output data but in a different way
- Prediction and classification both are the supervised learning methods, where prediction is trained to predict real number outputs and classification is trained to identify/predict to which category the new values fall into

- Example (classification): Before starting of your 7th sem project, you want to predict whether it is accepted or rejected in final project defense
- Example (Prediction): Before starting 7th sem, if you want to predict how much score you want to obtain, here you use a prediction model by consulting previously obtained marks

Regression

Regression

- Regression analysis can be used to model the relationship between one or more independent (or predictor) variables and a dependent (or response) variable
- The predictor variables are the attributes of interest describing the tuple. Generally, the values of the predictor variables are known
- The response variable is what we want to predict
- **Types of Regression**
  - **Linear Regression**: A linear function of single independent variable x (Ex: $y = a + b.x$)
  - **Multiple Linear Regression**: A linear function of multiple independent variable x1, x2, x3,... (Ex: $y = a + b1.x1 + b2.x2 + b3.x3...$)
  - **Polynomial Regression**: A polynomial function of independent variable x (Ex: $y = a + b1.x + b2.x^2 + b3.x^3...$)
  - **Non-Linear Regression**: A non linear function with one or more parameters (Ex: $y = a.e^{b.x}$)

Correlation Coefficient

Correlation Coefficient

- Correlation is a measure of the extent to which two variables are related
- **Positive correlation**: is a relationship between two variables in which both variables move in same direction. Ex: Height and Weight
- **Negative correlation**: is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. Ex: Climbing mountain and getting colder
- **Zero correlation**: exists when there is no relationship between two variables. Ex: amount of tea taken and intelligence level
- Correlation can be expressed visually through scatterplot. It indicates the strength and direction of the correlation between the co-variables
- Correlation ranges between -1 and +1

Correlation Coefficient...

- The correlation coefficient is calculated as
  $r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2).(n\sum Y^2 - (\sum Y)^2)}}$
  where, n-> number of data points or observations
  $\sum XY$ -> sum of the product of x-value and y-value for each point in the data set
  $\sum X$ -> sum of the x-values in the data set
  $\sum Y$ -> sum of the y-values in the data set
  $\sum X^2$ ->sum of the squares of the x-values in the data set
  $\sum Y^2$ ->sum of the squares of the y-values in the data set

Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

$n = 10$

$\sum X = 80$ & $\sum Y = 255$

$\sum XY = 2289$

$\sum X^2 = 756$ & $\sum Y^2 = 7097$

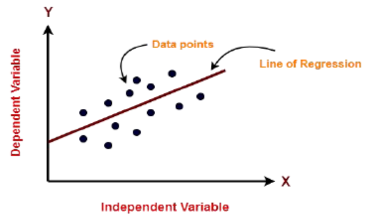
$(\sum X)^2 = 6400$ & $(\sum Y)^2 = 65025$

$\mathbf{r = 0.95}$

Linear Regression

Linear Regression

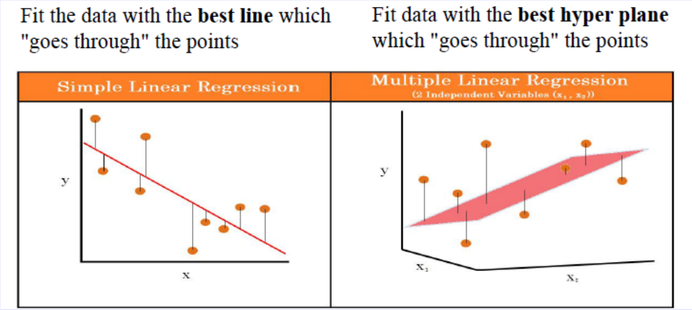
- Linear Regression is a supervised machine learning algorithm. It tries to find the best linear relationship that describes the given data
- Linear regression model represents the linear relationship between a dependent variable and independent variable(s) via a sloped straight line



- Linear regression are of two types:
  - **Simple Linear Regression**: Dependent variable depends only on a single independent variable
  - **Multiple Linear Regression**: Dependent variable depends on more than one independent variables

Linear Regression...

Linear Regression...



Linear Regression...

Linear Regression...

Linear Regression cont…

A linear regression line has an equation of the form **Y = bX +a + e**,

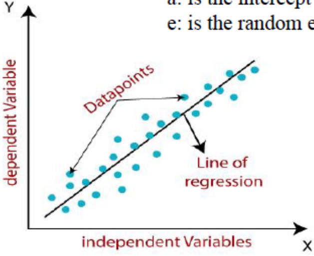
where X: explanatory variable

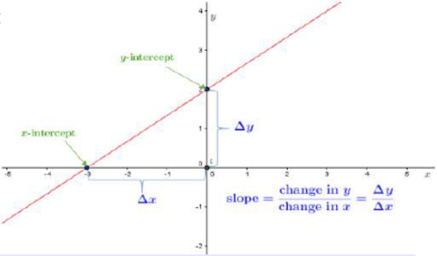
Y: is the dependent variable.

b: regression coefficient or slope of the line

a: is the intercept

e: is the random err

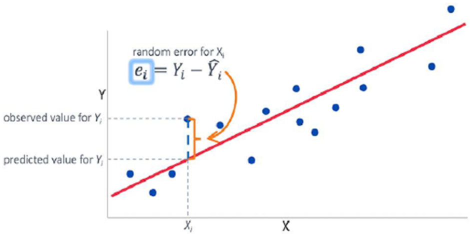




Linear Regression...

The random error in the following linear equation of line:

$\hat{Y}_i$ is predicted values of Y_i.



The calculation of b and a is as follows:

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad a = \frac{\Sigma Y}{n} - b\,\frac{\Sigma X}{n}$$

If b > 0, then x(predictor) and y(target) have a positive relationship. That is increase in x will increase y.

If b < 0, then x(predictor) and y(target) have a negative relationship. That is increase in x will decrease y.

Linear Regression...

Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

$n = 10$

$\sum X = 80$ & $\sum Y = 255$

$\sum XY = 2289$

$\sum X^2 = 756$ & $\sum Y^2 = 7097$

$(\sum X)^2 = 6400$ & $(\sum Y)^2 = 65025$

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{10 \times 2289 - (80 \times 255)}{[10 \times 756 - (80)^2]} = 2.1466;$$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n} = \frac{255}{10} - 2.1466\frac{80}{10} = 8.3272$$

The linear regression will thus be Predicted **Y = 2.1466 X + 8.3272**

The above equation can be used to predict the volume of sales for an insurance company given its agent number. Thus if a company has 1000 agents (10 hundreds) the predicted value of sales will be around ?

Performance Evaluation of Regression

Mean Square Error (MSE)

MSE is defined as mean or average of the square of the difference between actual and estimated values. Mathematically, it is represented as:

$MSE = \frac{1}{n}\sum_{i=1}^{n}(actual\_value(i) - predicted\_value(i))^2$

MSE is useful when large errors are particularly undesirable

Root Mean Square Error (RMSE)

It is just the square root of the mean square error. Mathematically, it is represented as:

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(actual\_value(i) - predicted\_value(i))^2}$

RMSE provides error in the same units as the target variable

Performance Evaluation of Regression

Mean Absolute Percentage Error (MAPE)

The formula to calculate MAPE is as follows:

$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{actual\_value(i) - calculated\_value(i)}{actual\_value(i)} \right|$

MAPE is easy to interpret and easy to explain. The lower the value for MAPE, the better a model is able to predict values

Since it expresses error as a percentage, making it scale-independent

Mean Absolute Error (MAE)

It calculates the average difference between the calculated values and actual values. It is also known as scale-dependent accuracy. The formula to calculate MAE is as follows:

$MAE = \frac{1}{n} \sum_{i=1}^{n} |actual\_value(i) - calculated\_value(i)|$

Performance Evaluation of Regression...

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Value	42	45	49	55	57	60	62	58	54	50	44	40
Predicted Value	44	46	48	50	55	60	64	60	53	48	42	38
Error	-2	-1	1	5	2	0	-2	-2	1	2	2	2
Squared Error	4	1	1	25	4	0	4	4	1	4	4	4

Sum of Square Error=56
So, MSE = 56/12 = 4.6667

RMSE = SQRT(4.6667) = 2.2

MAPE = 3.64%   MAE = 1.833

Ref: J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 3rd edition