Introduction to Supervised Learning

Linear Regression

Intro to supervised learning and Linear Regression – Topics

Machine Learning:

- Intro to machine learning, learning from data.
- Supervised Learning, train - test data.
- Overfitting and Under fitting

Linear Regression:

- Linear relation between two variables, measures of association – correlation and covariance.
- A simple fit, best fit line – measures of a regression fit.
- Simple and Multiple linear regression
- R squared, Adjusted R squared.

Machine Learning

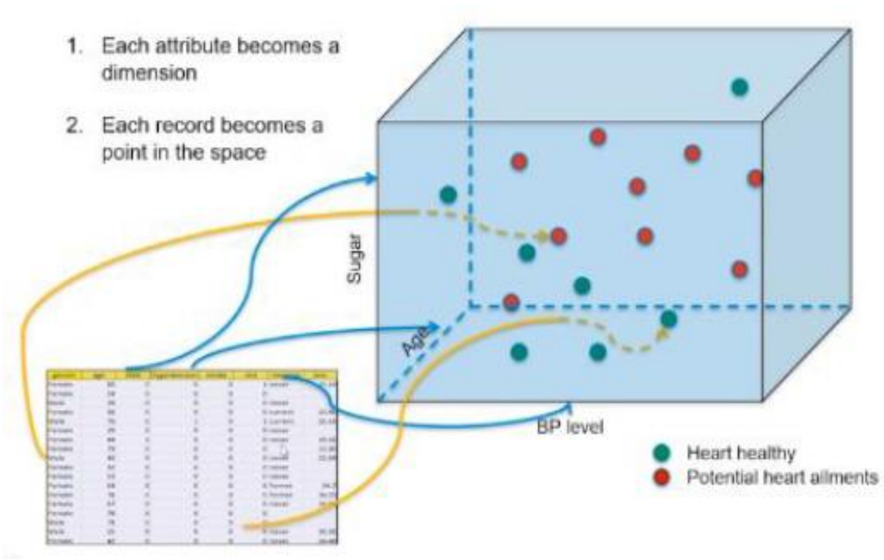
- The ability of a computer to do some task without being explicitly programmed.
- The ability to do the tasks come from the underlying model which is the result of the learning process.
- The model is generated by learning from huge volume of data, huge both in breadth and depth reflecting the real world in which the processes are performed.

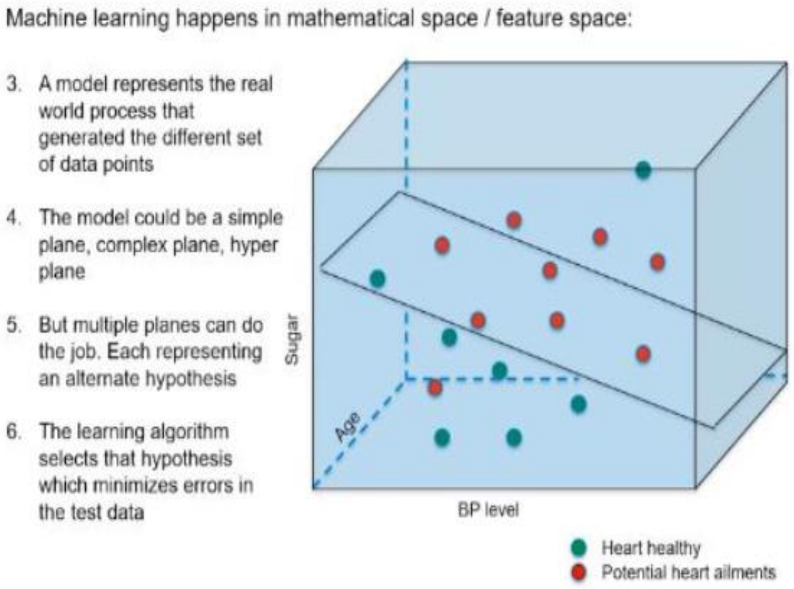
What machine learning algorithms do?

- Search through the data to look for patterns in form of trends, cycles, associations, etc.
- Express these patterns as mathematical structures.

Machine Learning Contd.

- A data point in a real world comprises of different attributes which identify it as an entity. Such data points come together to form a data set to be learned from in a mathematical space.





Supervised Machine Learning

- Class of machine learning that works on externally supplied instances in form of predictor attributes and **associated target values**.
- The target values are the ‘correct answers’ for the predictor model which can either be a regression model or a classification model (classifying data into classes).
- The model learns from the training data using these ‘correct answers/target variables’ as reference variables.
- The model thus generated is used to make predictions about data not seen by the model before.
  - Ex1 : *model to predict the resale value of a car based on its mileage, age, color etc.*
  - *Ex2 : model to determine the type of a tumor.*
- If the model does very well with the training data but fails with test data(unseen data), overfitting is said to have taken place. However, if the data does not capture the features of train data itself, we term it as under fitting.

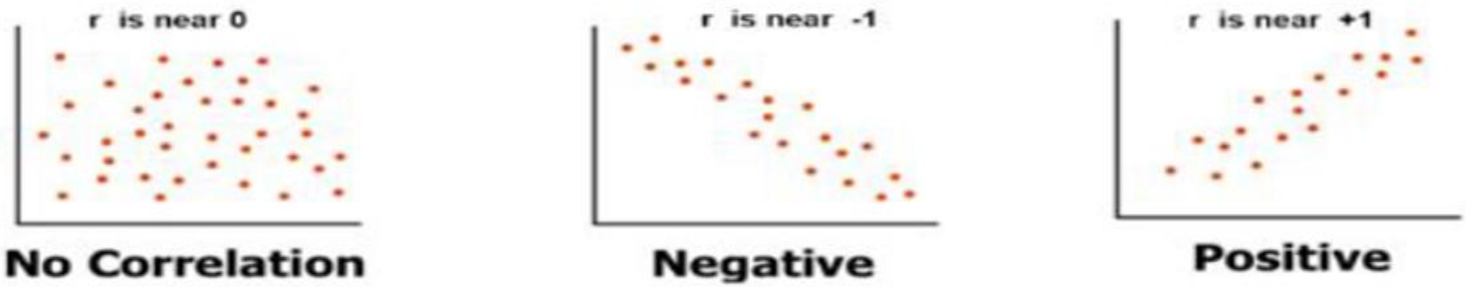
Measures of Association

- **Covariance**
  - Covariance is a measure of association between two variables.
  - It represents association in units of the two variables.
  - It only gives the direction of association i.e. positive or negative. It does not tell us the magnitude of association.
- **Correlation**
  - Correlation is also a measure of association between two variables.
  - Moreover, it is a dimensionless quantity and thus enables comparison beyond units.
  - It measures direction as well as magnitude of the association between two variables.

- Correlation only measures **linear relationship**. Zero correlation means no linear relationship, there **might** be non-linear relationship between variables.
- Coefficient of correlation is also known as Pearson's coefficient

Coefficient of relation - Pearson's coefficient p(x,y) = Cov(x,y) / (stnd Dev (x) X stnd Dev (y))

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



Generating linear model for cases where r is near 0, makes no sense. The model will not be reliable. For a given value of X, there can be many values of Y! Nonlinear models may be better in such cases

Linear Regression

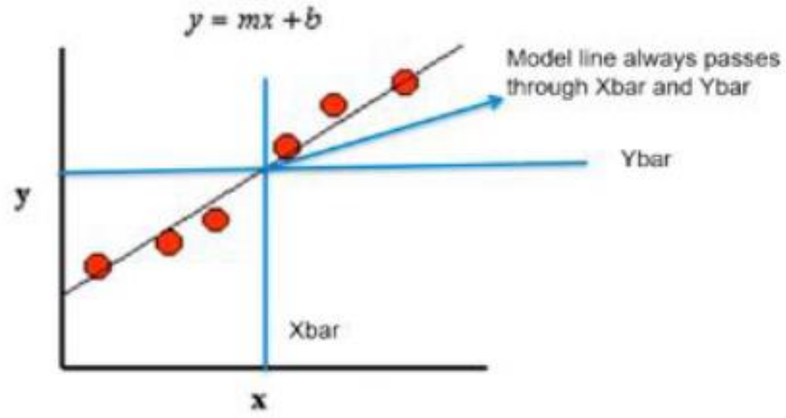
- The term “Regression” generally refers to predicting a target value, which is generally a real number, for a data point based on its attributes.

- The term “linear” in linear regression refers to the fact that the method models data with linear combination of the explanatory variables (attributes).

- In case of linear regression with a single explanatory variable, the linear combination can be expressed as :
  - response = intercept + constant*explanatory variable

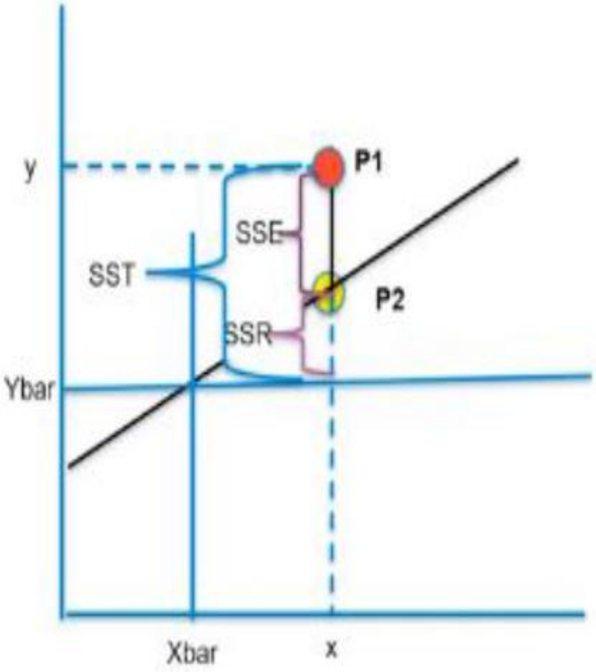
Best fit line

- Learning from the data, the model generates a line that fits the data.
- This line tries to explain the variance in the data.
- Our aim is to find a regression line that best fits the data.
- In the diagram below, we see the regression line. The red dots are the data points which constitute our data set.



Best Fit Line Contd

- The picture shows the different measures of a fit.



1. P1 – Original y data point for given x
2. P2 - Estimated y value for given x
3. Ybar – Average of all Y values in data set
4. SST – Sum of Square error Total (SST) Variance of P1 from Ybar (Y – Ybar)^2
5. SSR - Regression error (p2 – ybar)^2 (portion SST captured by regression model)
6. SSE - Residual error (p1 – p2)^2

R SQUARED VALUE

- The r squared is considered a measure of goodness of a fit. It is also known as **coefficient of determination**.
- It is the portion of the variance in data that is covered by the model. This is given by -

R Squared = (SST-SSE)/SST
= SSR/SST

Multiple regression

- Till now we have seen a simple regression where we have one attribute or independent variable.
- However, in the real world, a data point has various important attributes and they need to be catered to while developing a regression model.
  - Ex: predicting price of a house, we need to consider various attributes related with this house. Such a regression problem is an example of a multiple regression.
  - This can be represented by :

*target = constant1*feature1 + constant2*feature2 + constant3*feature3 + …..+ intercept*

- The model aims to find the constants and intercept such that this line is the best fit.

Pros and Cons of Linear Regression

Advantages

- Simple to implement and easier to interpret the outputs coefficient.
- Gives good results when dependent and independent variables have linear relationship.

Disadvantages

- Assumes a linear relationships between dependent and independent variables.
- Outliers can have huge effects on regression.
- Linear Regression assume independence between attributes.

Case Study

- There have been lot of studies undertaken in the past, on factors affecting life expectancy, considering demographic variables, income composition and mortality rates.
- It was found that affect of immunization and human development index was not taken into account in the past.
- Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this case study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well.
- In this case study, we will use linear regression to see the effect of various factors on Life Expectancy.

Case Study Contd.

- Attribute Information:
- Country: Country
- Year: Year
- Status: Developed or Developing status
- Life expectancy: Life Expectancy in age
- Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- infant deaths: Number of Infant Deaths per 1000 population
- Alcohol: Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)
- percentage expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita(%)
- Hepatitis B: Hepatitis B (Hep B) immunization coverage among 1-year-olds (%)
- Measles: Measles - number of reported cases per 1000 population

Case Study Contd.

- BMI: Average Body Mass Index of entire population
- under-five deaths: Number of under-five deaths per 1000 population
- Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)
- Total expenditure: General government expenditure on health as a percentage of total government expenditure (%)
- Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years)
- GDP: Gross Domestic Product per capita (in USD)
- Population: Population of the country
- thinness 1-19 years: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- thinness 5-9 years: Prevalence of thinness among children for Age 5 to 9(%)
- Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Schooling: Number of years of Schooling(years)

Case Study Contd.

Concepts to Cover:

- Overview of the data
- Data Visualization
- Data Preparation
- Choose Model, Train and Evaluate
- Conclusion
- Add on: Statsmodels
- Business Insights

greatlearning

Power Ahead

Happy Learning !