greatlearning
Learning for Life

Intro to Supervised Learning

Linear Regression

Intro to supervised learning and Linear Regression – Topics

Machine Learning:

- Intro to machine learning, learning from data.
- Supervised and Unsupervised 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 – measure of a regression fit.
- Multiple regression
- R squared.

Machine Learning

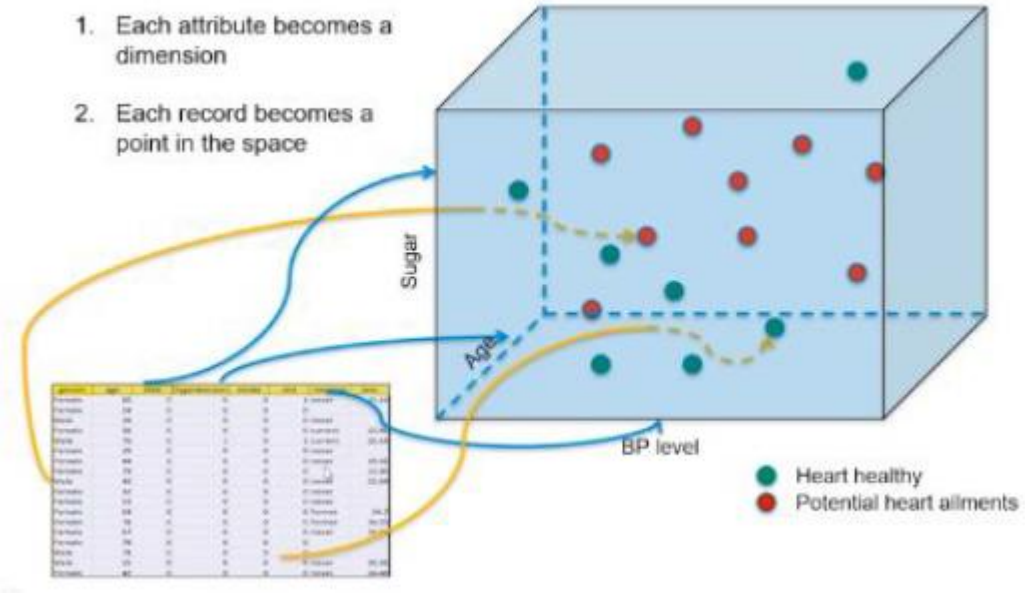
- The ability of a computer to do some task without being explicitly programmed.
- The ability to do the tasks come from the underlying model which is the result of the learning process.
- The model is generated by learning from huge volume of data, huge both in breadth and depth reflecting the real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

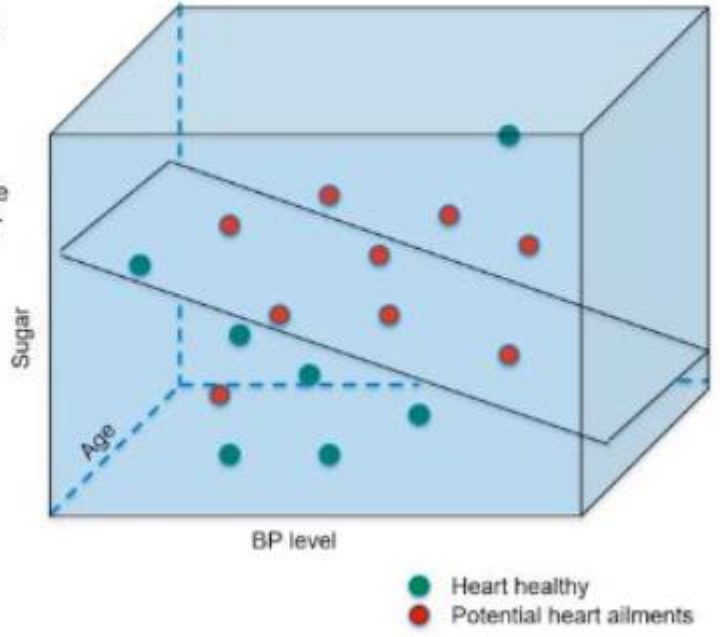
Machine Learning Contd.

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.



Machine learning happens in mathematical space / feature space:

3. A model represents the real world process that generated the different set of data points
4. The model could be a simple plane, complex plane, hyper plane
5. But multiple planes can do the job. Each representing an alternate hypothesis
6. The learning algorithm selects that hypothesis which minimizes errors in the test data



Supervised Machine Learning

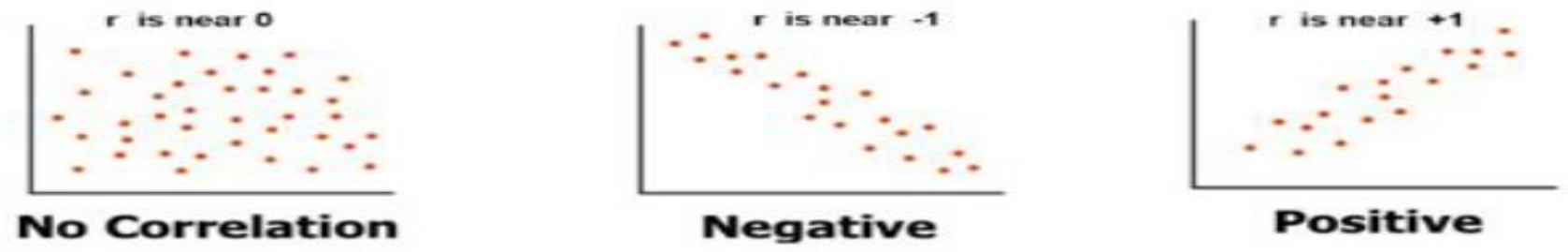
- Class of machine learning that work on externally supplied instances in form of predictor attributes and **associated target values**.
- The target values are the 'correct answers' for the predictor model which can either be a regression model or a classification model (classifying data into classes.)
- 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.
- **Correlation**
  - Correlation is also a measure of association between two variables.
  - Moreover, it is a dimensionless quantity and thus enables comparison beyond units**.**
  - 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

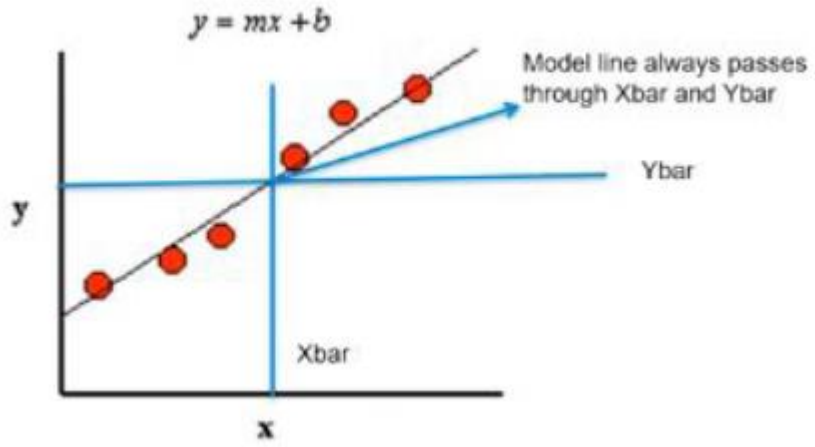
- The term “Regression” generally refers to predicting a target value, which is generally a real number, for a data point based on its attributes.

- 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

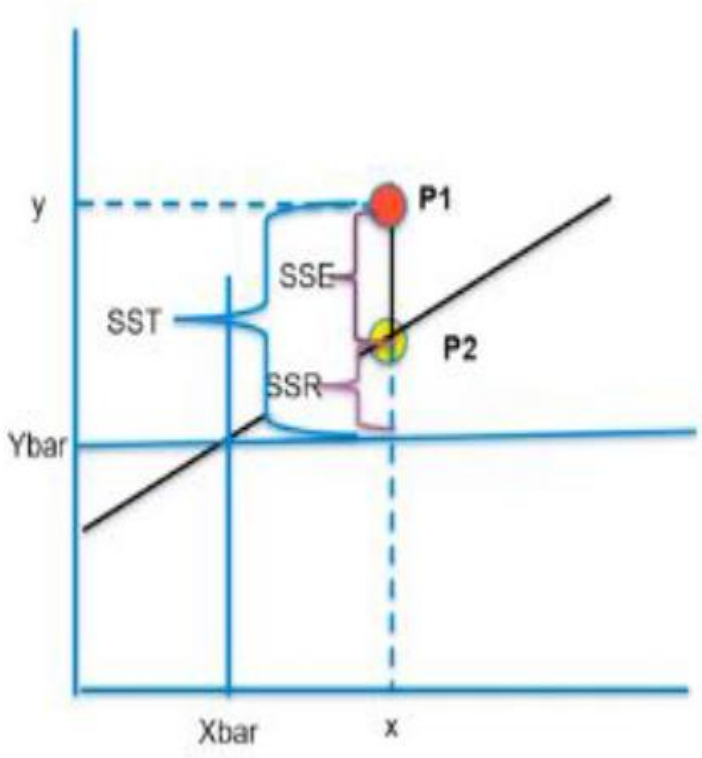
Best fit line

- 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 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.

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

Problem :

A certain bank wants to predict the credit loss based on the details provided by the customer while applying for loan. These details are Age, Years of Experience, Number of cars, Gender, Marital Status.

The objective is to come up with a regression model which can predict the credit loss based on the above parameters . Here are the details about the data set.

Case Study Contd.

Data Attributes:

- Ac_no: The account of customer used as identifier.
- Age: Age of the borrower
- Years of experience: Working experience
- Number of vehicles: Number of cars possessed
- Gender: M/F
- Losses in thousands: Target variable

Case Study Contd.

Steps to follow:

- Import libraries
- Get the data
- Find top 5 headings from the data
- Plot histograms
- Find the correlation between variables.
- Drop variables which are of no use to the model.
- Find the model using coefficients

Questions if any…