Introduction to machine learning

Machine Learning

Introduction to machine learning

“[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.” Arthur Samuel 1959:

Introduction to machine learning

What is machine learning?

1. Is a process of enabling a computer based system to learn to do tasks based on well defined statistical and mathematical methods

2. The ability to do the tasks come from the underlying model which is the result of the learning process. Sometimes the ability comes from an mathematical algorithm

3. The model generated represents behaviour of the processes that were earlier performed before machine learning

4. The model is generated from huge volume of data, huge both in breadth and depth reflecting the real world in which the processes are performed

5. The more representative data is of the real world, the better the model would be. The challenge is how to make it a true representative

Introduction to machine learning

What do machine learning algorithms do?

1. Search through data to look for patterns

2. Patterns in form of trends, cycles, associations, classes etc.

3. Express these patterns as mathematical structures such as probability equations or polynomial equations

Introduction to machine learning

When is machine learning useful ?

1. Cannot express our knowledge about patterns as a program. For e.g. Character recognition or natural language processing
2. Do not have an algorithm to identify a pattern of interest. For e.g. In spam mail detection
3. Too complex and dynamic. For e.g. Weather forecasting
4. Too many permutations and combinations possible. For e.g. Genetic code mapping
5. No prior experience or knowledge. For e.g. Mars rover
6. Patterns hidden in humongous data. For e.g. Recommendation system

Introduction to machine learning

Where are machine learning based systems used (examples only)

1. Fraud detection
2. Sentiment analysis
3. Credit risk management
4. Prediction of equipment failures
5. New pricing models / strategies
6. Network intrusion detection
7. Pattern and image recognition
8. Email spam filtering

Introduction to machine learning

Machine Learning & Data Science

1. Machine learning is part of a larger discipline called Data Science

2. Data science is the process of applying science and domain expertise to data to extract useful information from data.

3. It includes application of all the statistical and mathematical tools and techniques to glean out the useful information from data using machine learning

Introduction to machine learning

Machine Learning Pre-requisites

1. Rich set of data representing the real world

2. Knowledge and skills in
    a. Maths and statistics
    b. Programming (Python, R, Java, Go)
    c. Tools / frameworks such as Keras / TensorFlow
    d. Domain knowledge

Real World as Mathematical Space

Introduction to machine learning

Machine learning happens in mathematical space / feature space:

1. A data set representing the real world, is a collection attributes that define an entity

2. Each entity is represented as one record / line in the data set

Attributes / Dimensions

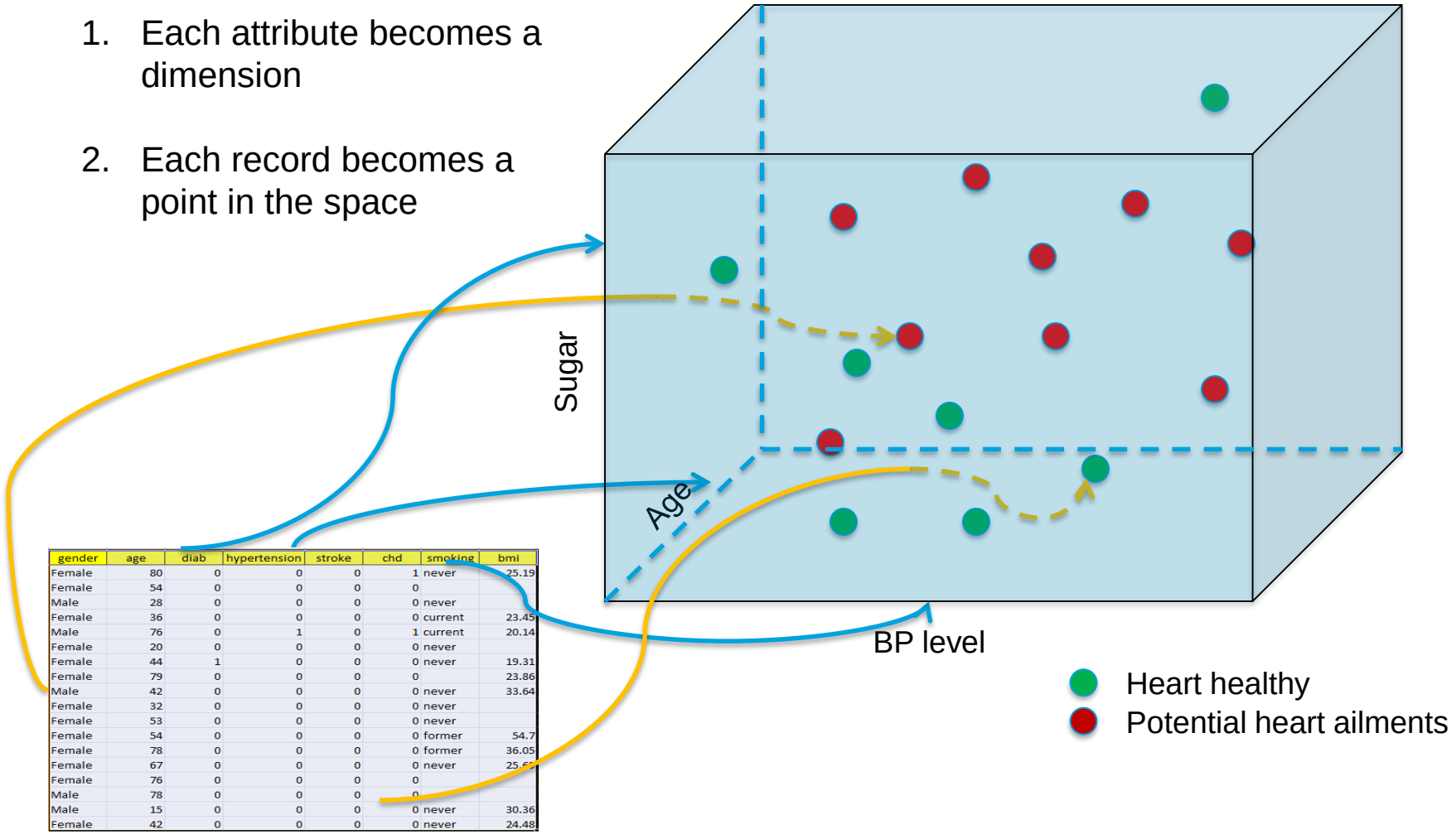
gender	age	diab	hypertension	stroke	chd	smoking	bmi
Female	80	0	0	0	1	never	25.19
Female	54	0	0	0	0		
Male	28	0	0	0	0	never	
Female	36	0	0	0	0	current	23.45
Male	76	0	1	0	1	current	20.14
Female	20	0	0	0	0	never	
Female	44	1	0	0	0	never	19.31
Female	79	0	0	0	0		23.86
Male	42	0	0	0	0	never	33.64
Female	32	0	0	0	0	never	
Female	53	0	0	0	0	never	
Female	54	0	0	0	0	former	54.7
Female	78	0	0	0	0	former	36.05
Female	67	0	0	0	0	never	25.69
Female	76	0	0	0	0		
Male	78	0	0	0	0		
Male	15	0	0	0	0	never	30.36
Female	42	0	0	0	0	never	24.48

Introduction to machine learning

Machine learning happens in mathematical space / feature space:

1. Each attribute becomes a dimension

2. Each record becomes a point in the space



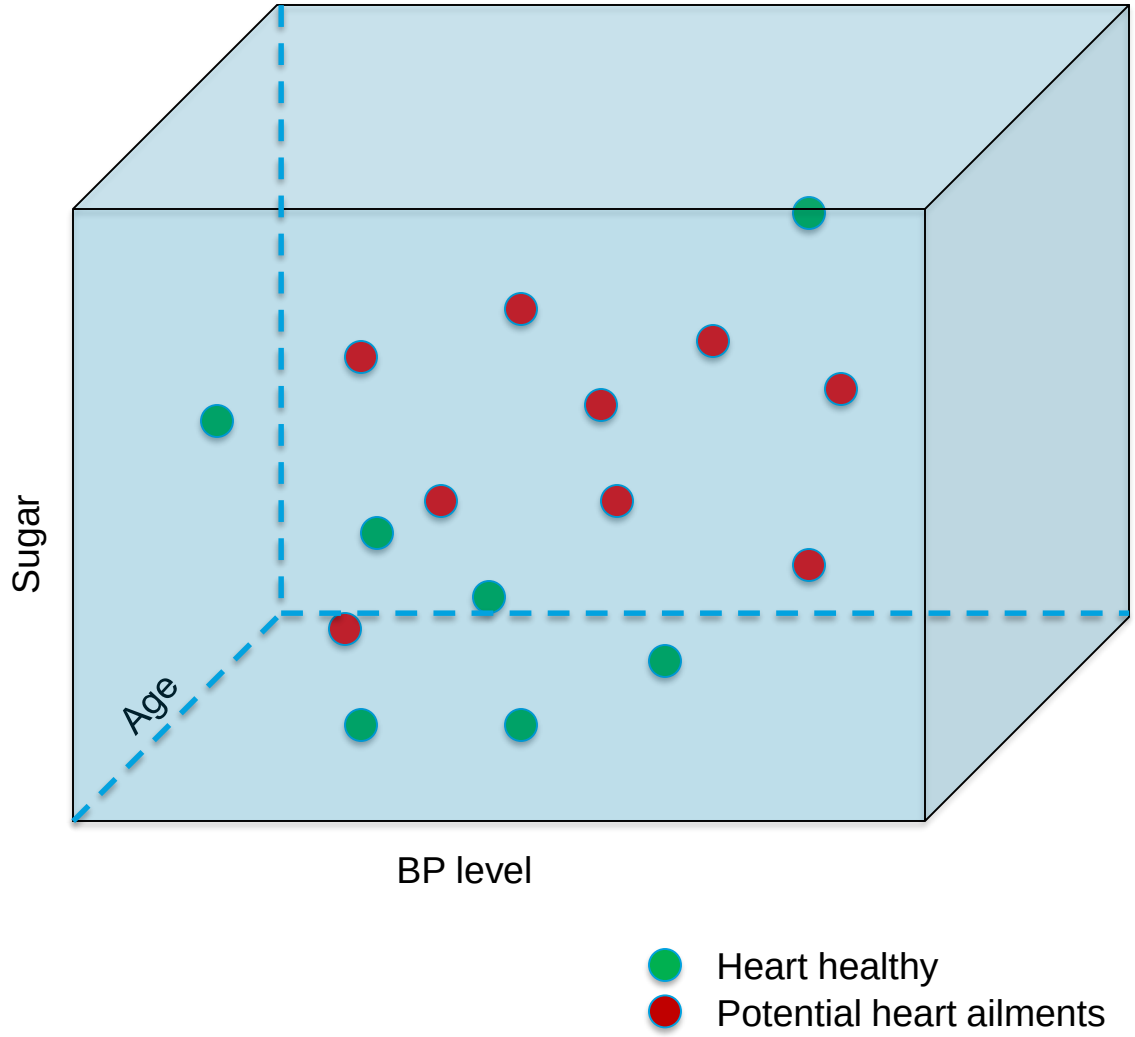
gender	age	diab	hypertension	stroke	chd	smoking	bmi
Female	80	0	0	0	1	never	25.19
Female	54	0	0	0	0		
Male	28	0	0	0	0	never	
Female	36	0	0	0	0	current	23.45
Male	76	0	1	0	1	current	20.14
Female	20	0	0	0	0	never	
Female	44	1	0	0	0	never	19.31
Female	79	0	0	0	0		23.86
Male	42	0	0	0	0	never	33.64
Female	32	0	0	0	0	never	
Female	53	0	0	0	0	never	
Female	54	0	0	0	0	former	54.7
Female	78	0	0	0	0	former	36.05
Female	67	0	0	0	0	never	25.6
Female	76	0	0	0	0		
Male	78	0	0	0	0		
Male	15	0	0	0	0	never	30.36
Female	42	0	0	0	0	never	24.48

Introduction to machine learning

Machine learning happens in mathematical space / feature space:

1. Position of a point in space is defined with respect to the origin

2. The position is decided by the values of the attributes for a point



Introduction to machine learning

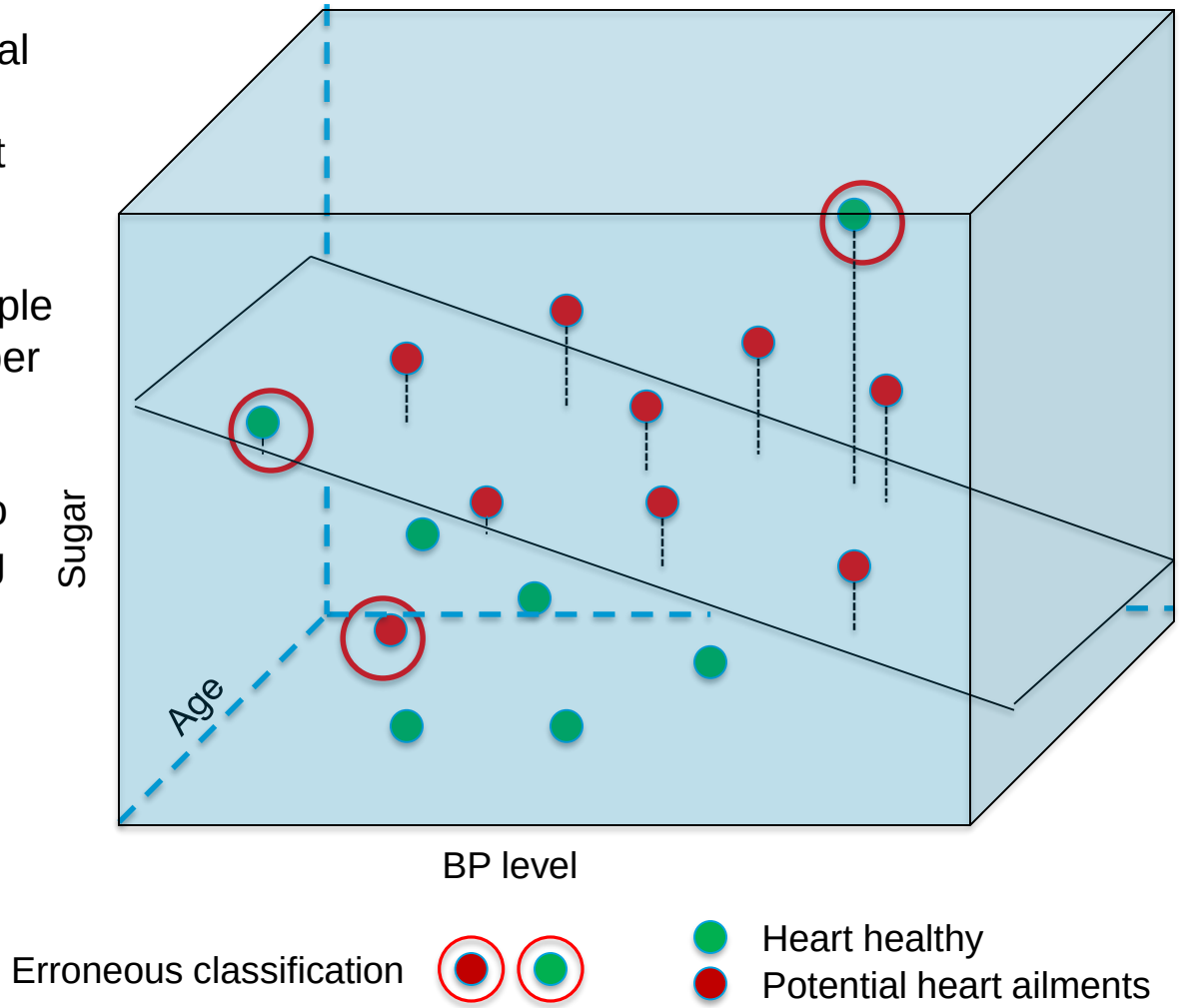
Machine learning happens in mathematical space / feature space:

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



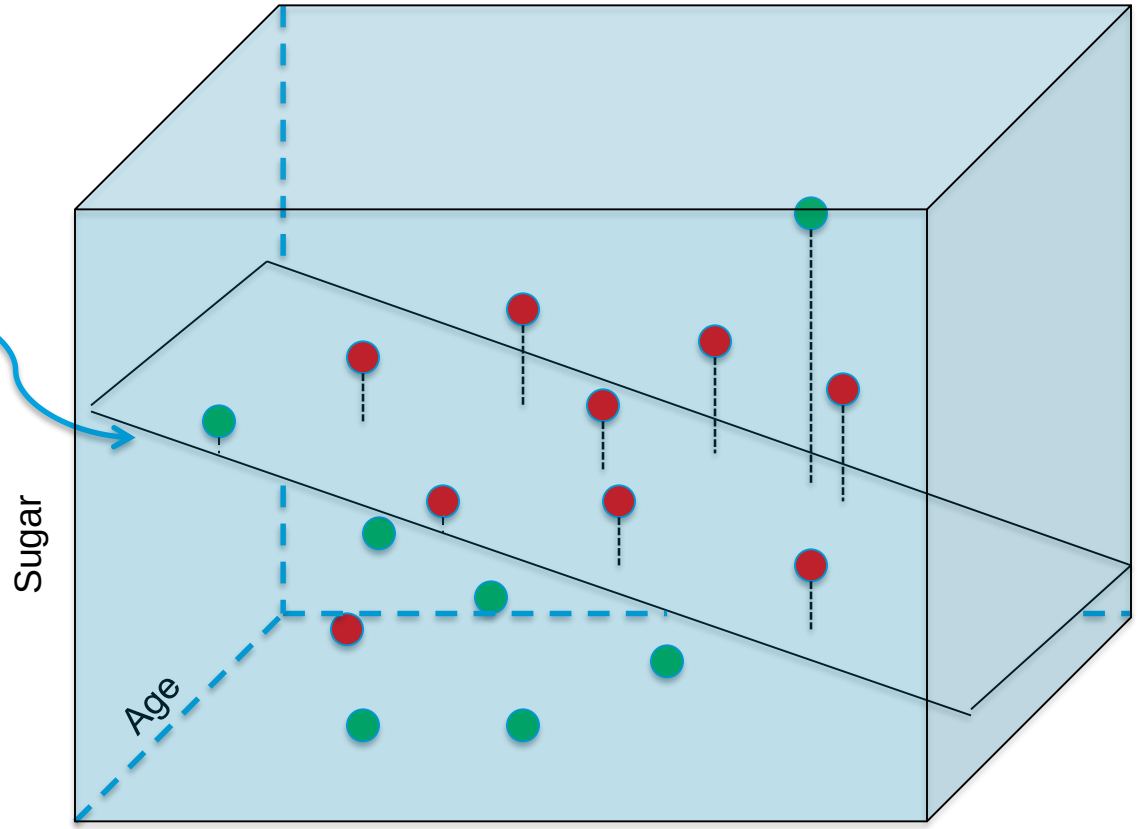
Introduction to machine learning

Machine learning happens in mathematical space / feature space:

7. In the figure, since the separator is a plane, the model will be the equation representing the plane

   ax + by + cz = d

8. x , y, z represent the three dimensions i.e. BP, Age, Sugar while d represents the color i.e. healthy or ailing heart



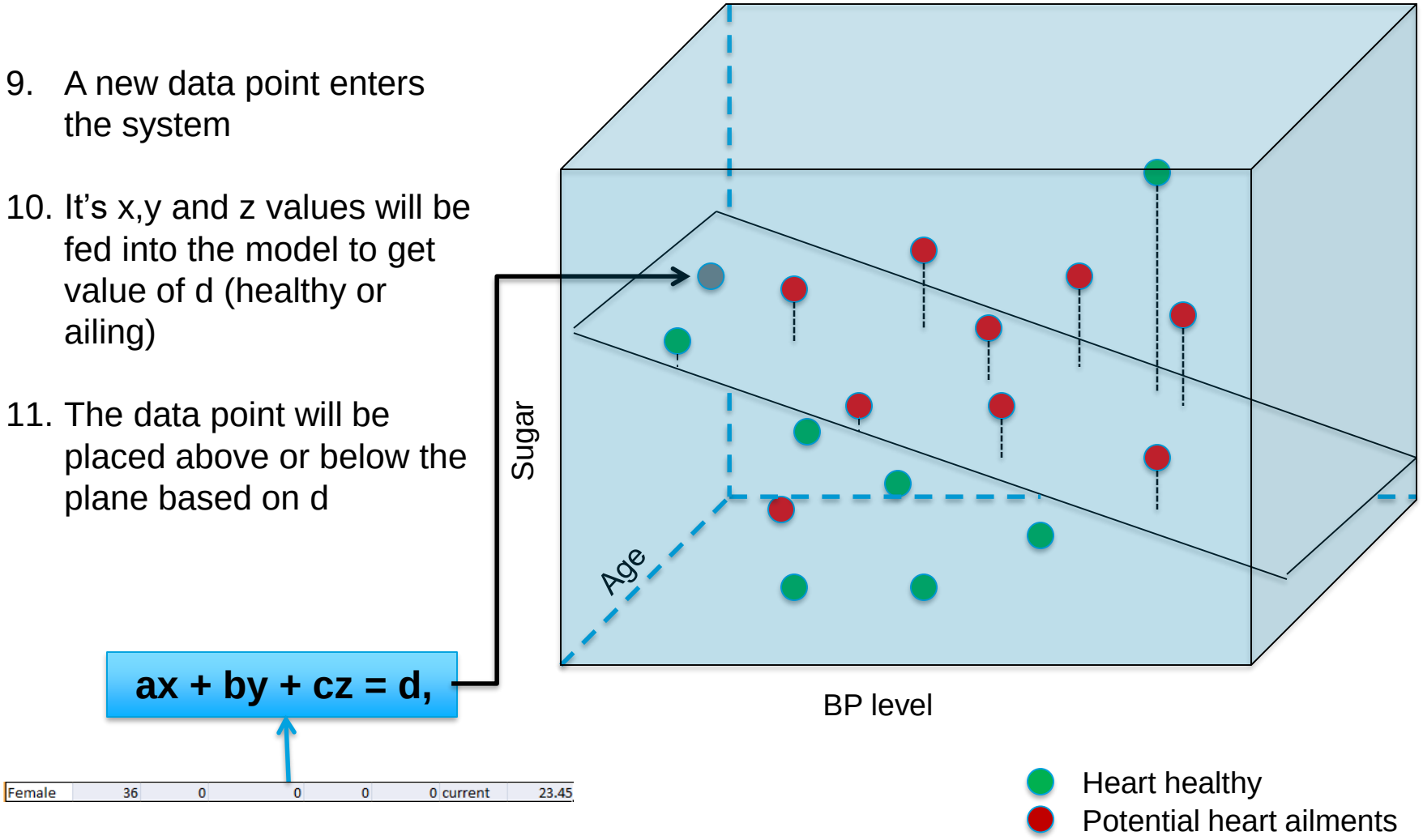
Introduction to machine learning

Machine learning happens in mathematical space / feature space:

9. A new data point enters the system

10. It's x,y and z values will be fed into the model to get value of d (healthy or ailing)

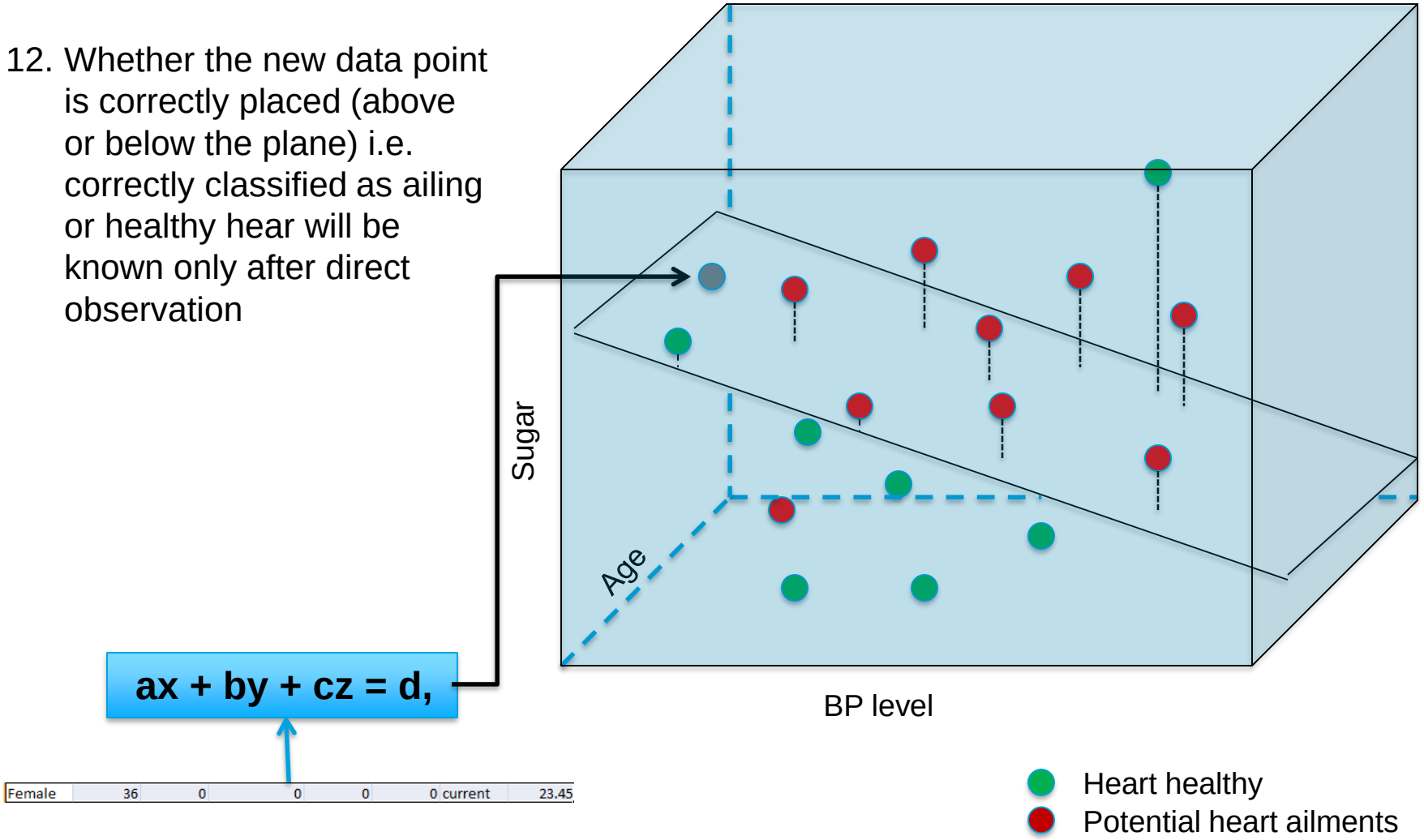
11. The data point will be placed above or below the plane based on d



Introduction to machine learning

Machine learning happens in mathematical space / feature space:

12. Whether the new data point is correctly placed (above or below the plane) i.e. correctly classified as ailing or healthy hear will be known only after direct observation



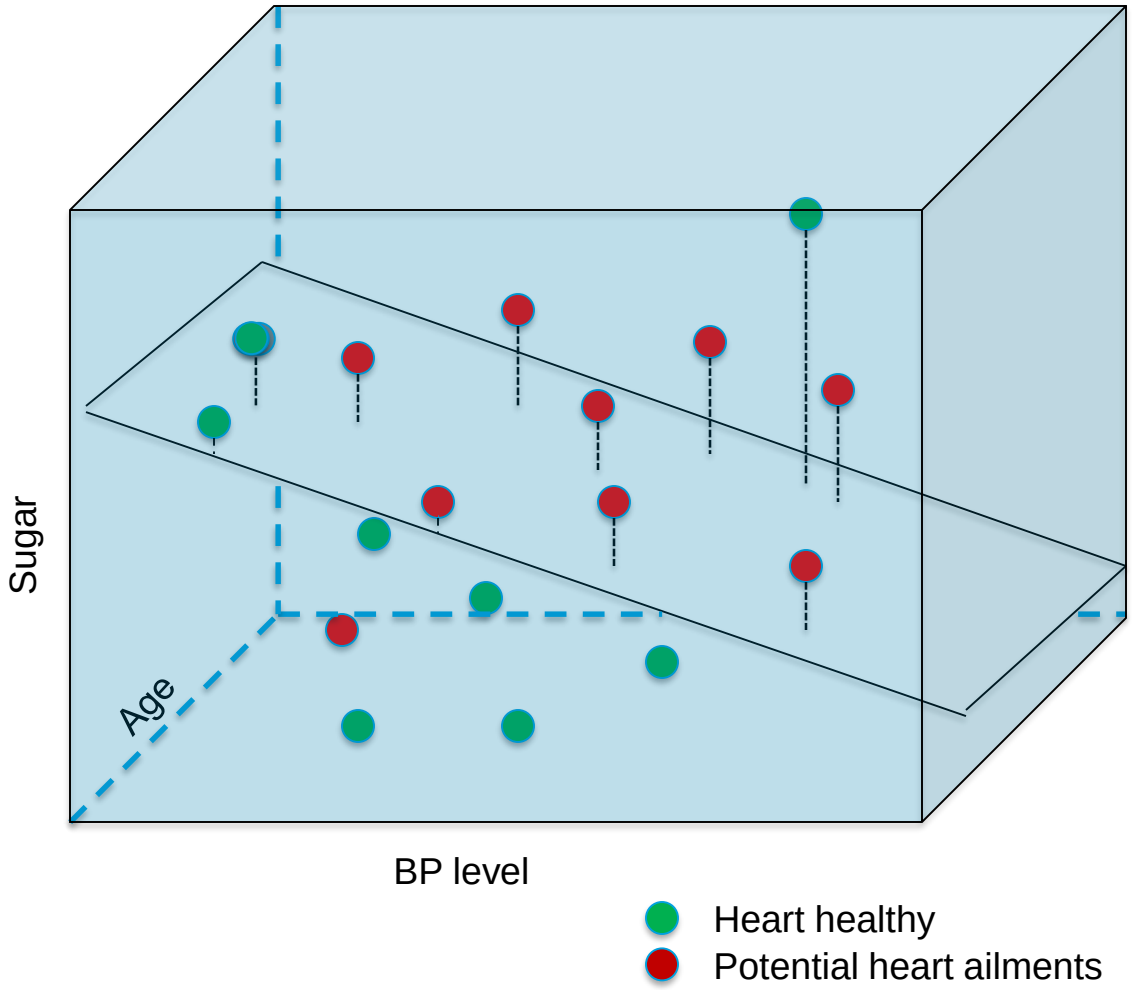
Introduction to machine learning

Machine learning happens in mathematical space / feature space:

13. Only direct test on the object of interest will tell whether the classification is correct or not

ax + by + cz = d,

14. If majority of new data points are correctly classified, the model is good else not



Introduction to Supervised Machine Learning

Introduction to machine learning

Characteristics of Supervised Machine Learning -

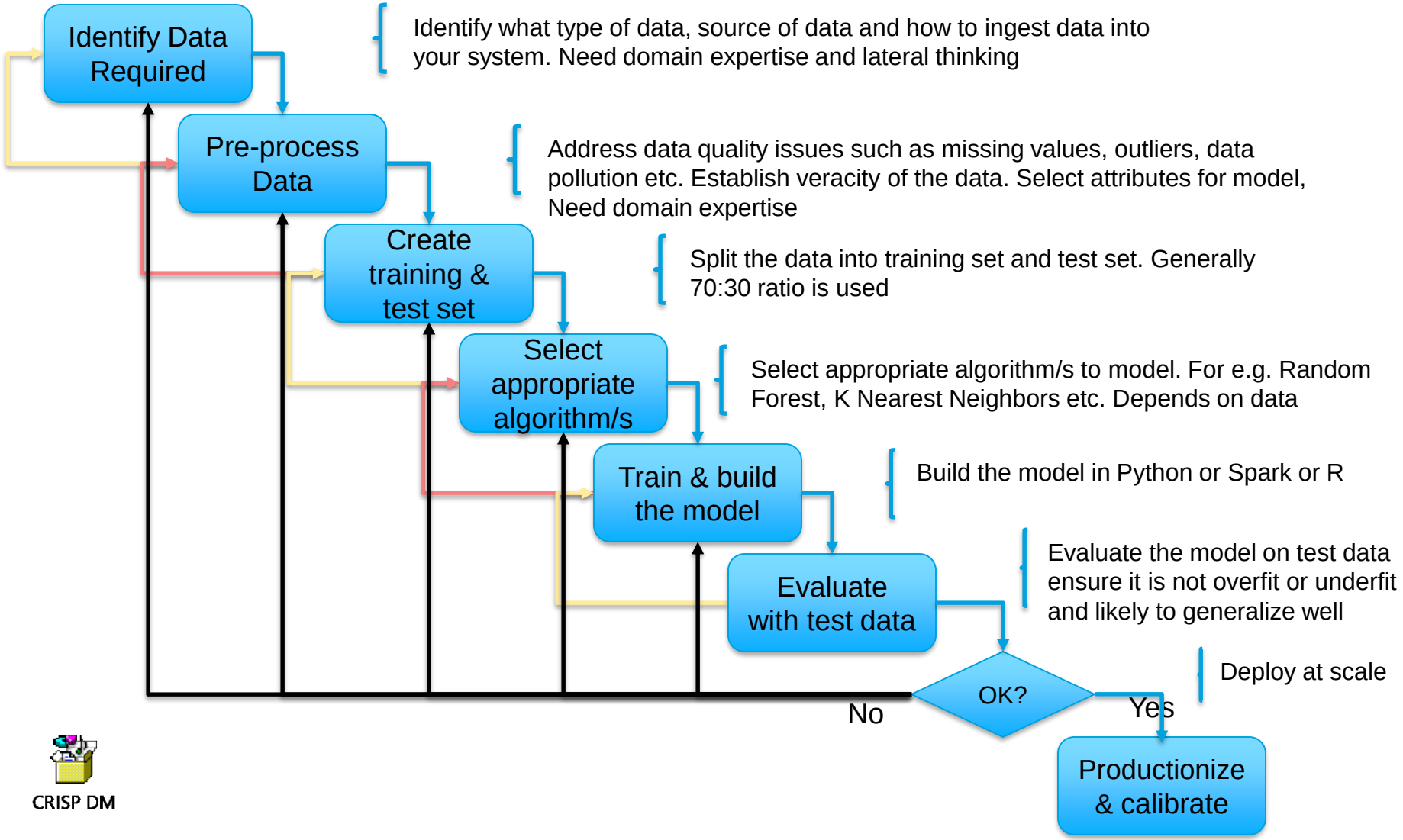
a. Class of machine learning algorithms that work on externally supplied instances (data) in form of predictor attributes and associated target values

b. They produce a model representing alternate hypothesis i.e. distribution of class labels in terms of predictor variables in the feature space

c. The model thus generated is used to make predictions about future instances where the predictor feature values are known but the target / class value is unknown
   a. E.g.-1 building model to predict the re-sale value of a car based on its current mileage, age, color etc.
   b. E.g.-2 Predicting the final year scores based on student performance in previous years.

Introduction to machine learning

Data Science Machine Learning Steps -



Introduction to machine learning

Linear Regression

Introduction to machine learning

Linear Regression Models -

a. The term "regression" generally refers to predicting a real number. However, it can also be used for classification (predicting a category or class.)

b. The term "linear" in the name “linear regression” refers to the fact that the method models data with linear combination of the explanatory variables.

c. A linear combination is an expression where one or more variables are scaled by a constant factor and added together.

d. In the case of linear regression with a single explanatory variable, the linear combination used in linear regression can be expressed as:

$$\text{response} = \text{intercept} + \text{constant} * \text{explanatory}$$

e. In its most basic form fits a straight line to the response variable. The model is designed to fit a line that minimizes the squared differences (also called errors or residuals.).

Introduction to machine learning

Linear Regression Models -

a. Before we generate a model, we need to understand the degree of relationship between the attributes Y and X

b. Mathematically correlation between two variables indicates how closely their relationship follows a straight line. By default we use Pearson's correlation which ranges between -1 and +1.

c. Correlation of extreme possible values of -1 and +1 indicate a perfectly linear relationship between X and Y whereas a correlation of 0 indicates absence of linear relationship

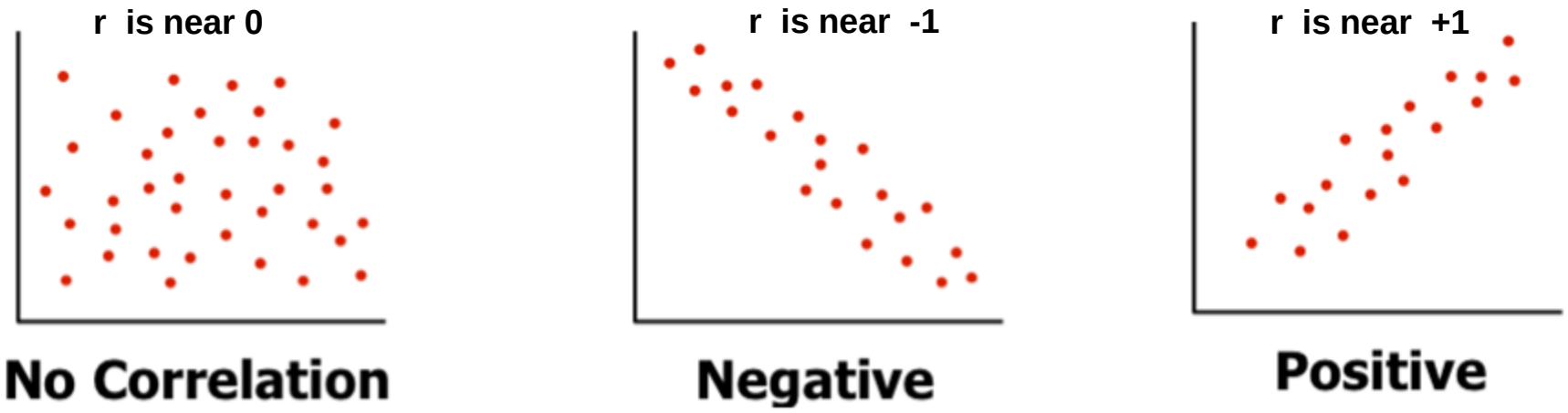
   I. When r value is small, one needs to test whether it is statistically significant or not to believe that there is correlation or not

Introduction to machine learning

Linear Regression Models -

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

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



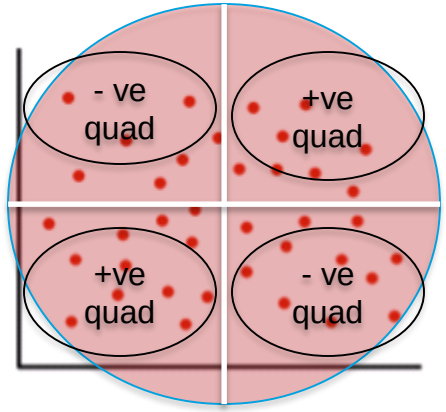
e. Generating linear model for cases where r is near 0, makes no sense. The model will not be reliable. For a given value of X, there can be many values of Y! Nonlinear models may be better in such cases

Introduction to machine learning

Linear Regression Models (Recap) -

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

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



$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = 0$$

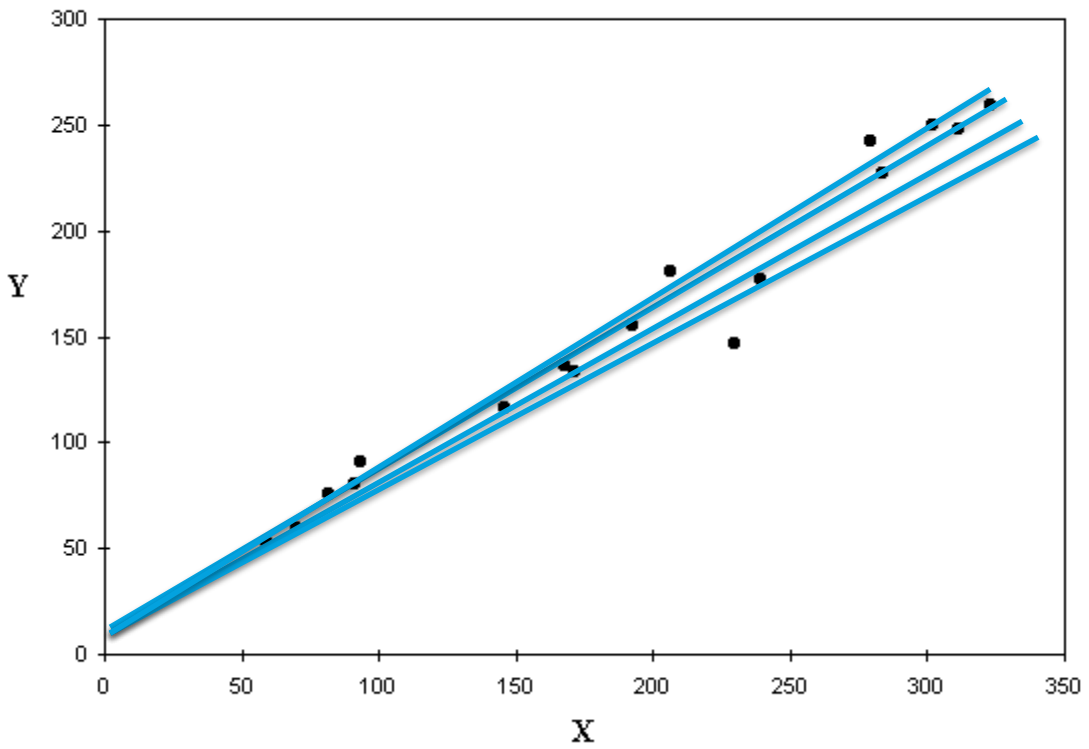
http://www.socscistatistics.com/tests/pearson/Default2.aspx

Introduction to machine learning

Linear Regression Models -

g. Given Y = f(x) and the scatter plot shows apparent correlation between X and Y
Let's fit a line into the scatter which shall be our model

h. But there are infinite number of lines that can be fit in the scatter. Which one should we consider as the model?



i. This and many other algorithms use gradient descent or variants of gradient descent method for finding the best model

j. Gradient descent methods use partial derivatives on the parameters (slope and intercept) to minimize sum of squared errors

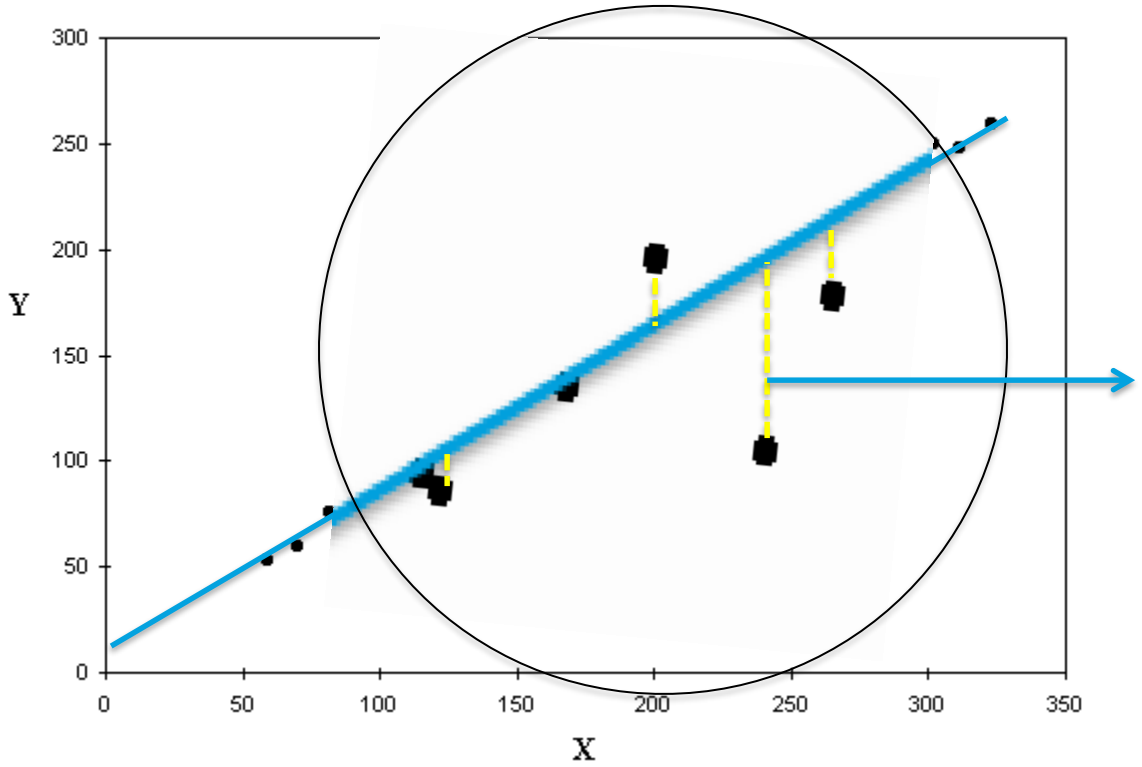
Introduction to machine learning

Linear Regression Models (Recap) -

k. Whichever line we consider as the model, it will not pass through all the points.

l. The distance between a point and the line (drop a line vertically (shown in yellow)) is the error in prediction

m. That line which gives least sum of squared errors is considered as the best line



Error = (T – (mx + C)

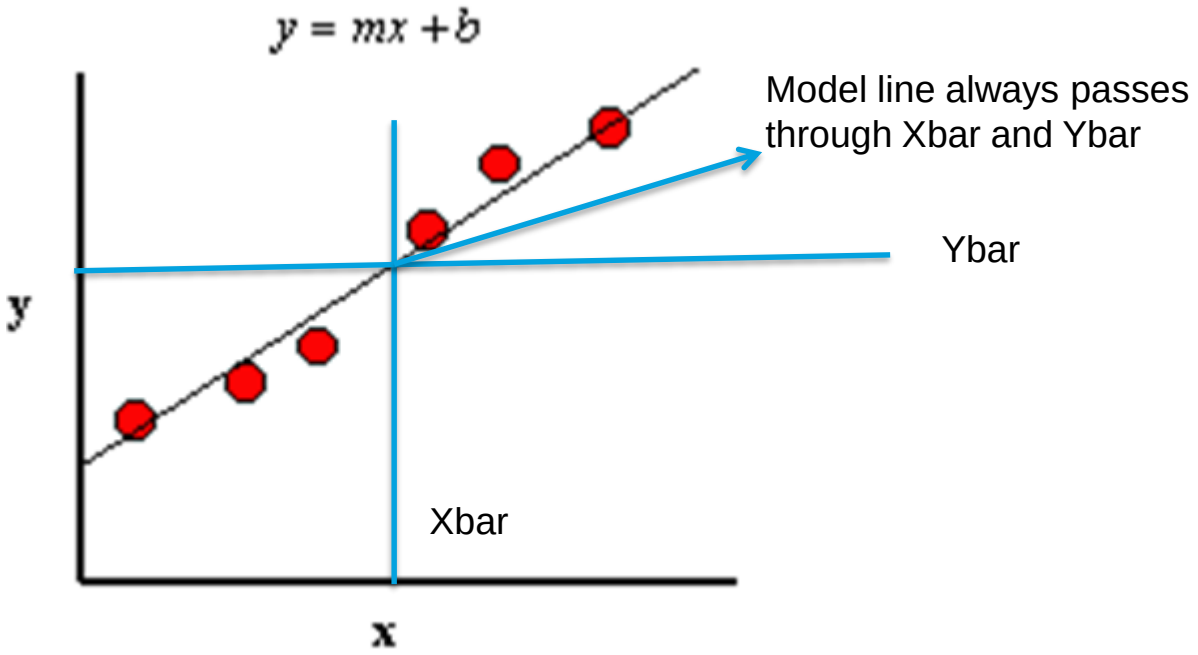
Sum of all errors can cancel out and give 0

We square all the errors and sum it up. That line which gives us least sum of squared errors is the best fit

Introduction to machine learning

Linear Regression Models -

n. Coefficient of determinant – determines the fitness of a linear model. The closer the points get to the line, the R^2 (coeff of determinant) tends to 1, the better the model is



Introduction to machine learning

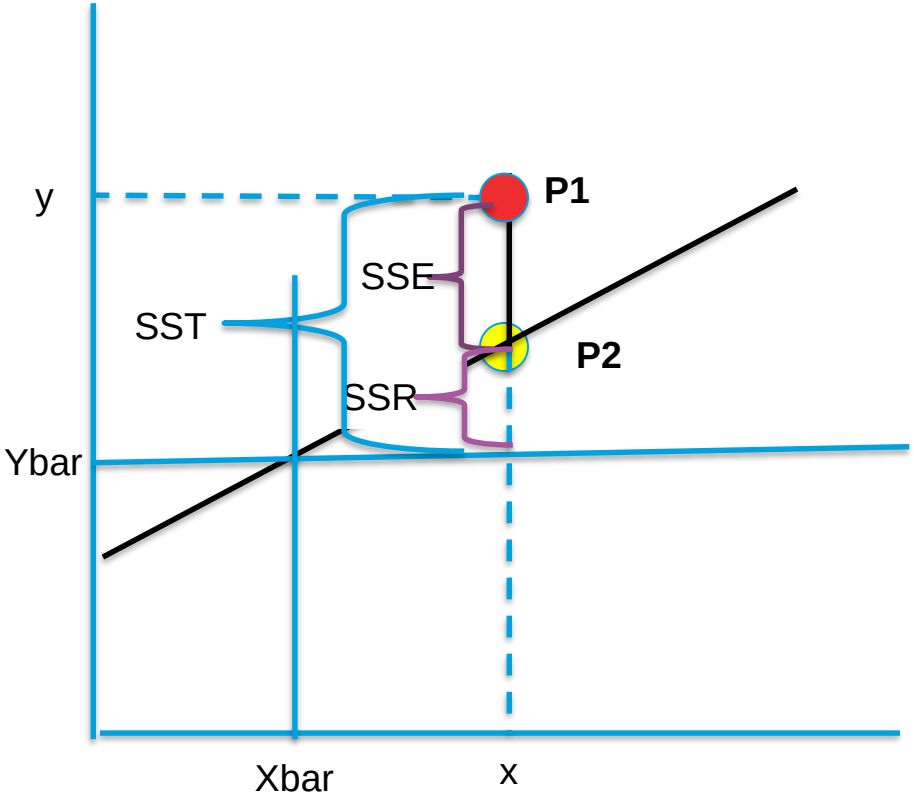
Linear Regression Models -

o. Coefficient of determinant (Contd…)

I. There are a variety of errors for all those points that don't fall exactly on the line.

II. It is important to understand these errors to judge the goodness of fit of the model i.e. How representative the model is likely to be in general

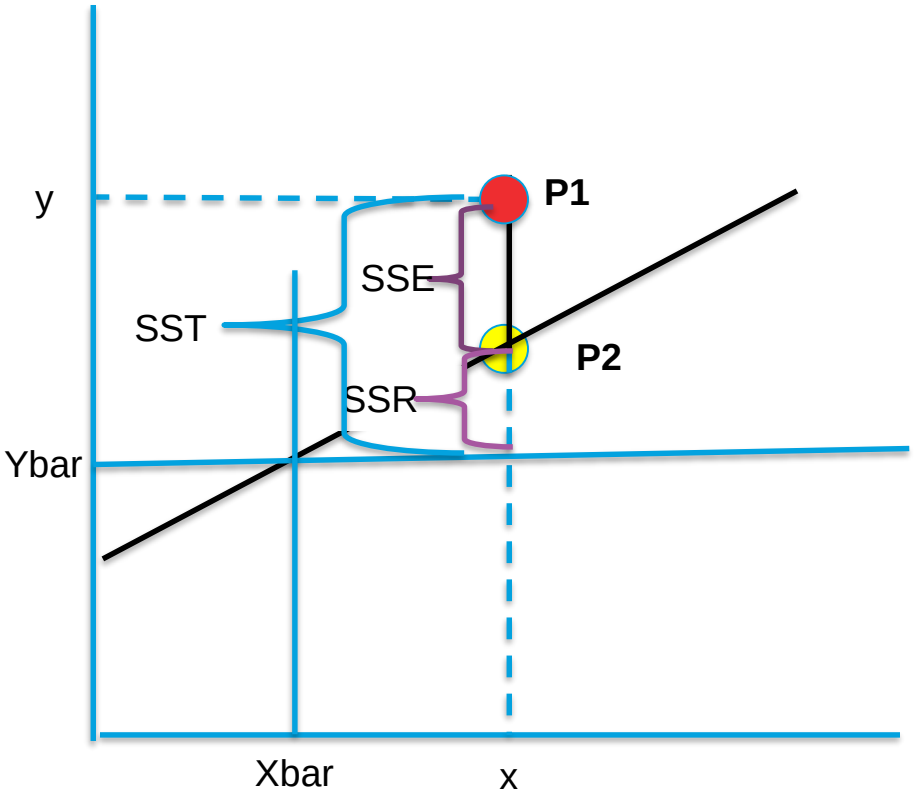
III. Let us look at point P1 which is one of the given data points and associated errors due to the model



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

Introduction to machine learning

Linear Regression Models -



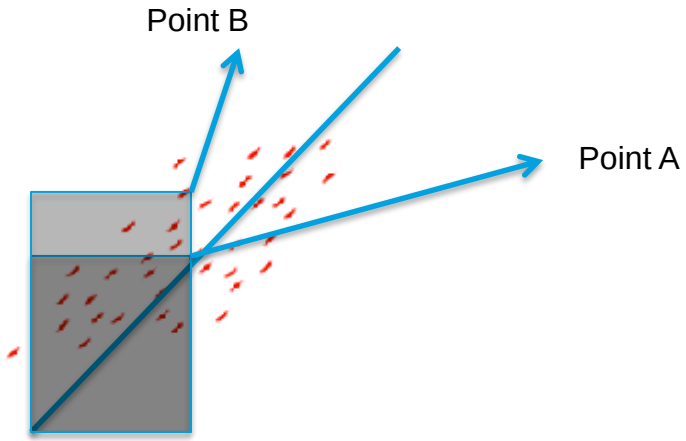
p. Coefficient of determinant (Contd…)

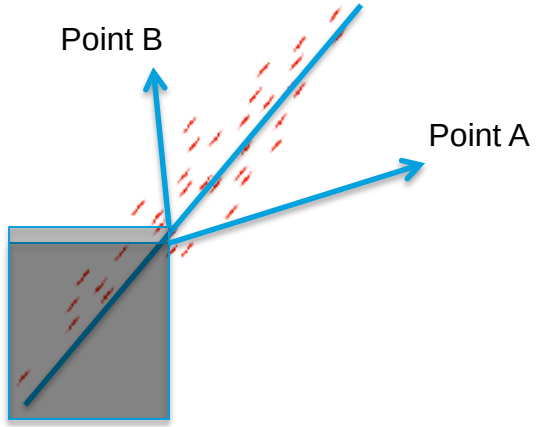
1. That model is the most fit where every data point lies on the line. i.e. SSE = 0 for all data points
2. Hence SSR should be equal to SST i.e. SSR/SST should be 1.
3. Poor fit will mean large SSE. SSR/SST will be close to 0
4. SSR / SST is called as r^2 (r square) or coefficient of determination
5. r^2 is always between 0 and 1 and is a measure of utility of the regression model

Introduction to machine learning

Linear Regression Models -

q. Coefficient of determinant (Contd…) -





In case of point “A”, the line explains the variance of the point

Whereas point “B” the is a small area (light grey) which the line does not represent.

%age of total variance that is represented by the line is coeff of determinant

Introduction to machine learning

Linear Regression Model -

Advantages –

1. Simple to implement and easier to interpret the outputs coefficients

Disadvantages -

1. Assumes a linear relationships between dependent and independent variables. That is, it assumes there is a straight-line relationship between them
2. Outliers can have huge effects on the regression
3. Linear regression assume independence between attributes
4. Linear regression looks at a relationship between the mean of the dependent variable and the independent variables.
5. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables
6. Boundaries are linear

Introduction to machine learning

Linear Regression Model -

Lab- 1- Estimating mileage based on features of a second hand car

Description – Sample data is available at
https://archive.ics.uci.edu/ml/datasets/Auto+MPG

The dataset has 9 attributes listed below that define the quality

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

Sol : mpg-linear regression.ipynb