Supervised Machine Learning

greatlearning
Learning for Life

Linear Regression

Supervised Machine Learning

greatlearning
Learning for Life

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:

response = intercept + constant * 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.).

Supervised Machine Learning

greatlearning
Learning for Life

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

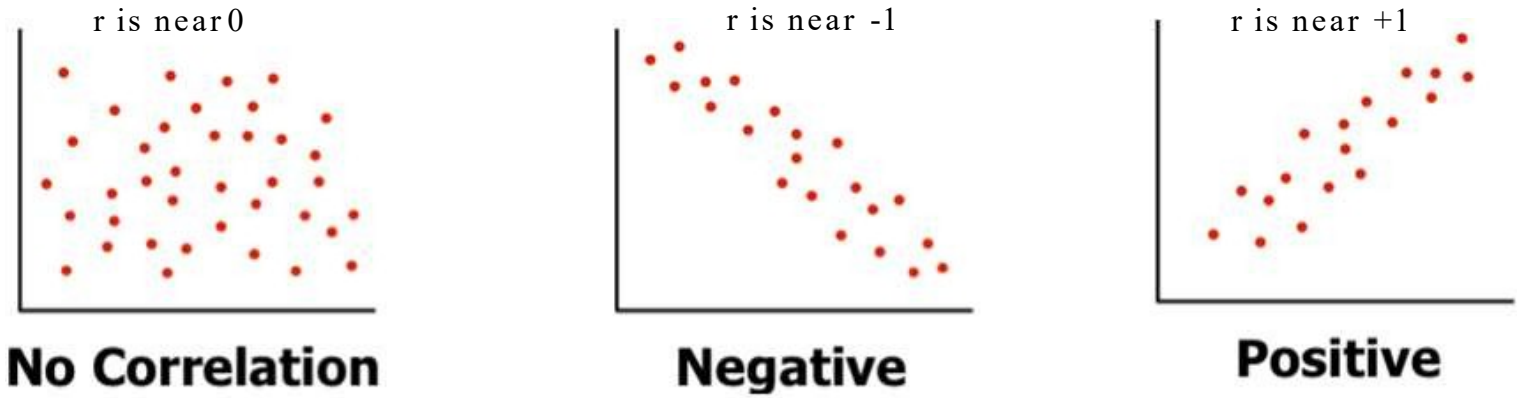
Supervised Machine Learning

greatlearning
Learning for Life

Linear Regression Models (Recap) -

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

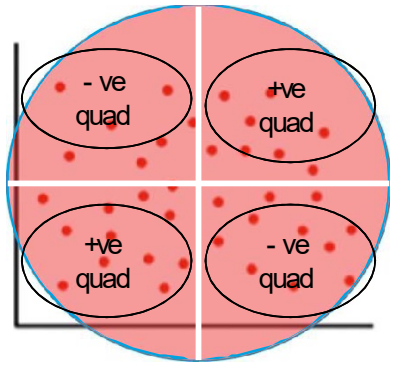
Supervised Machine Learning

greatlearning
Learning for Life

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

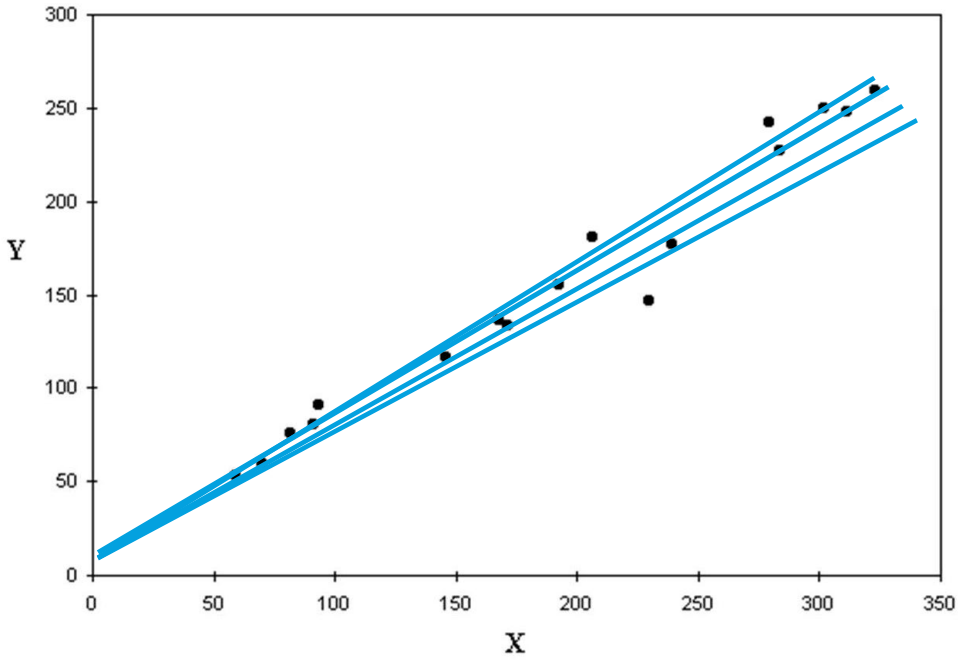
Supervised Machine Learning

greatlearning
Learning for Life

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

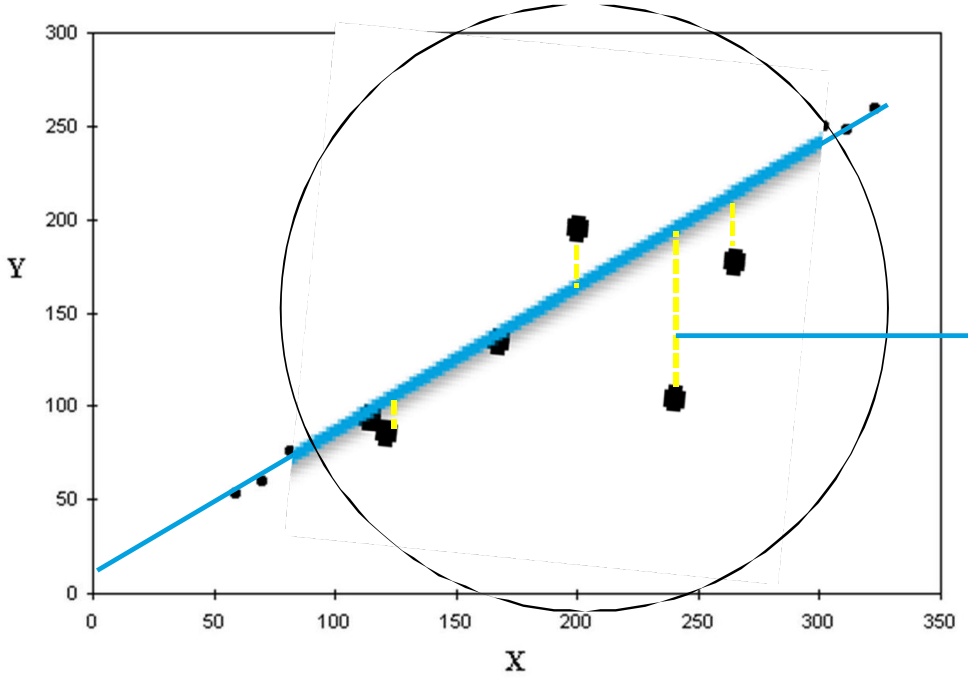
j. Gradient descent methods use partial derivatives on the parameters (slope and intercept) to minimize sum of squared errors

Supervised Machine Learning

greatlearning
Learning for Life

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

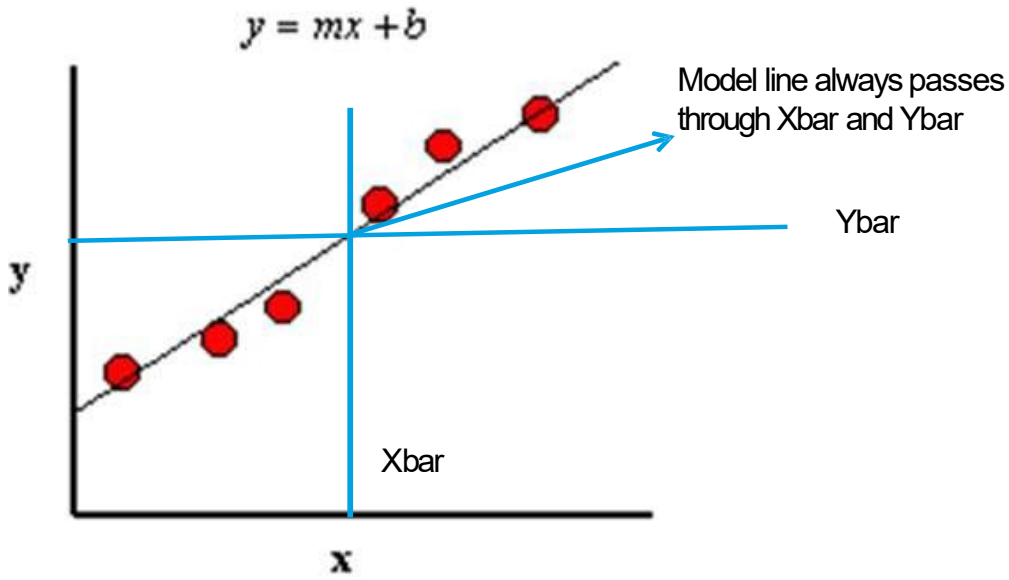
We square all the errors and sum it up. That line which gives us least sum of squared errors is the best fit

Supervised Machine Learning

greatlearning
Learning for Life

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



Supervised Machine Learning

greatlearning
Learning for Life

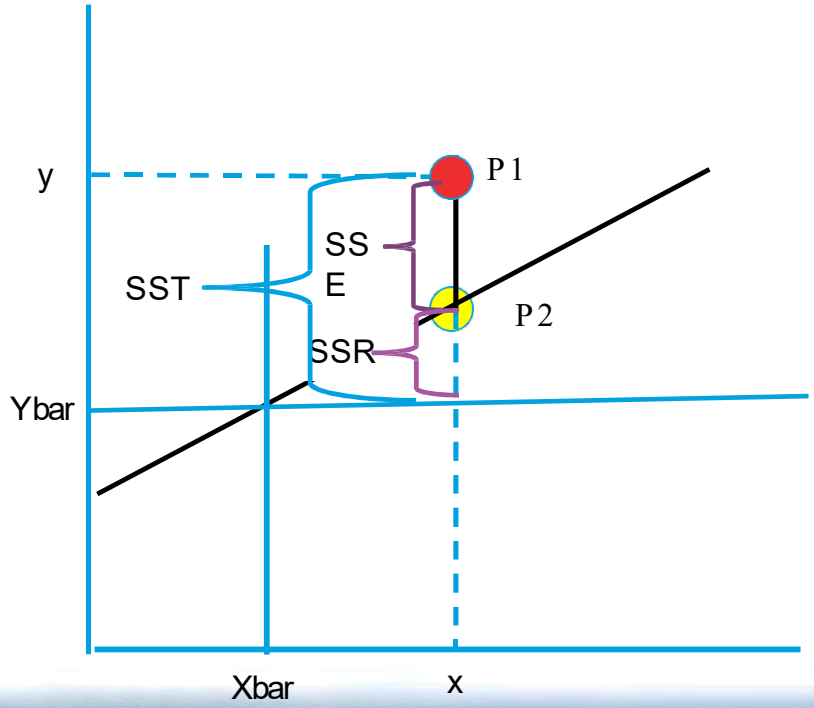
Linear Regression Models -

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

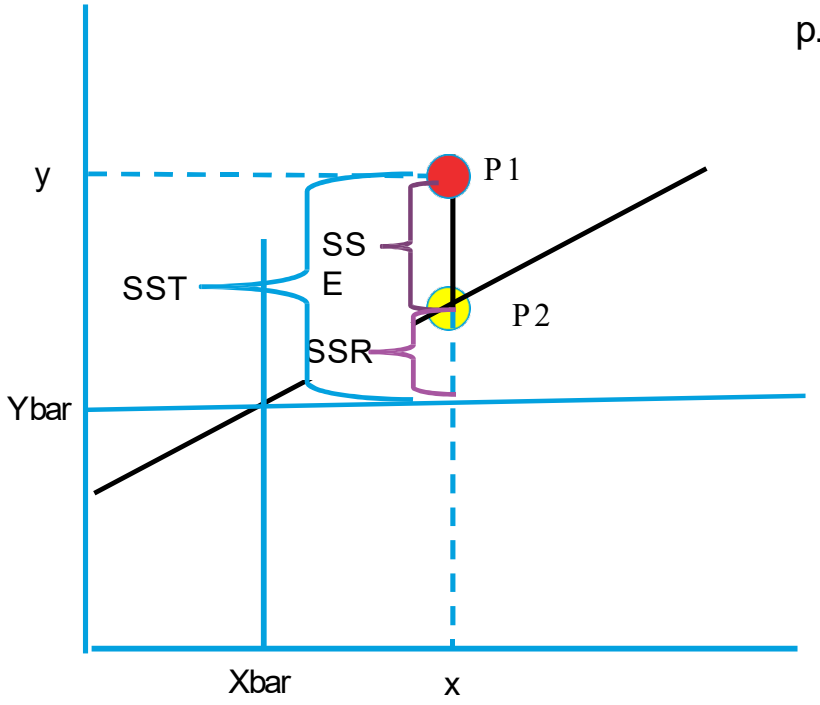


1. P1 – Original y data point for given x
2. P2 - Estimated y value for given x
3. Ybar – Average of all Y values in data set
4. SST – Sum of Square error Total (SST) Variance of P1 from Ybar (Y – Ybar)^2
5. SSR - Regression error (p2 – ybar)^2 (portion SST captured by regression model)
6. SSE - Residual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

Supervised Machine Learning

greatlearning
Learning for Life

Linear Regression Models -



p. Coefficient of determinant (Contd…)

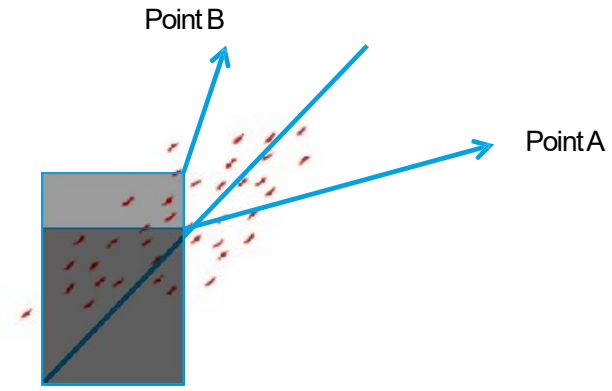
1. That model is the most fit where every data point lies on the line. i.e. SSE =0 for all data points
2. Hence SSR should be equal to SST i.e. SSR/SST should be 1.
3. Poor fit will mean large SSE. SSR/SST will be close to 0
4. SSR / SST is called as r^2(r square)or coefficient of determination
5. r^2is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

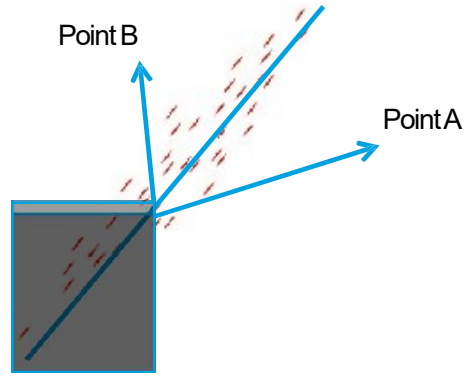
Supervised Machine Learning

greatlearning
Learning for Life

Linear Regression Models (Recap) -

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

Supervised Machine Learning

greatlearning
Learning for Life

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

Supervised Machine Learning

greatlearning
Learning for Life

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

Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression

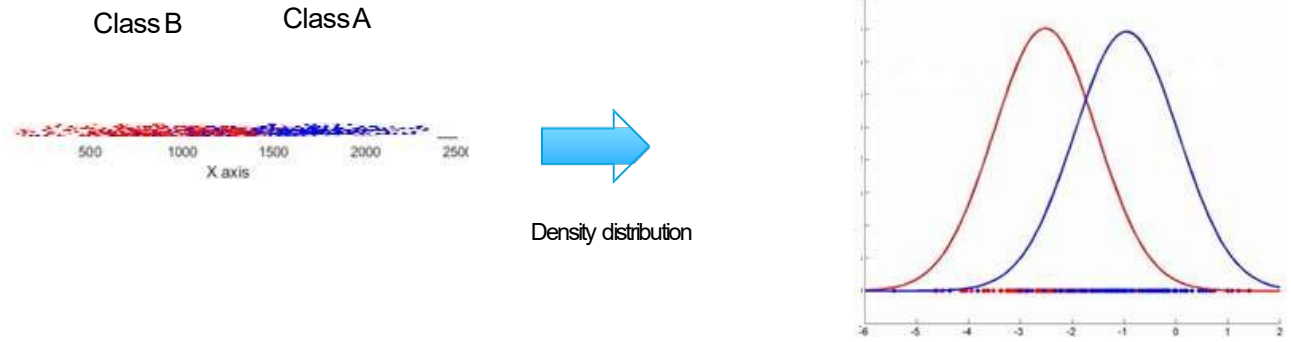
Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression Model -

a. A classification method built on the same concept as linear regression. The response variable is categorical. In it's simplest form, the response variable is binary i.e. belongs to one class or the other

b. Given the value of predictor (variable x), the model estimates the probability that the new data point belongs to a given class say "A". Probability values can range between 0 and 1.



Supervised Machine Learning

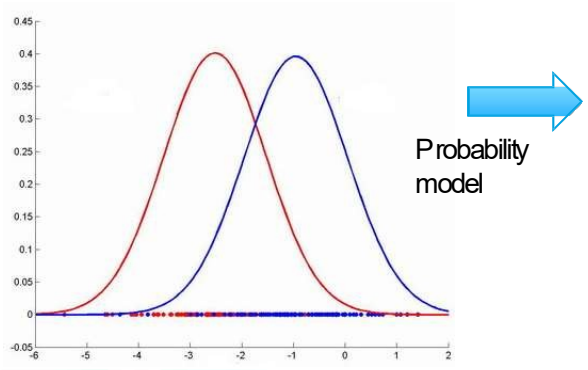
greatlearning
Learning for Life

Logistic Regression Model

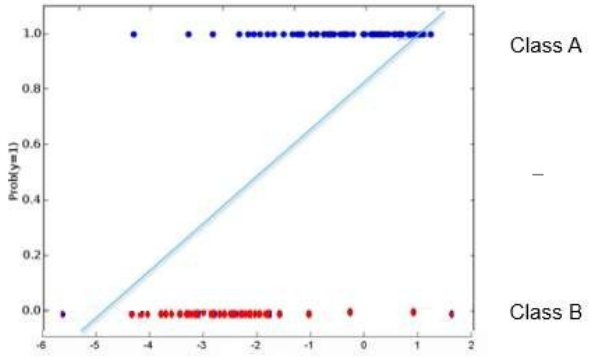
c. A new data point (shown with "?") needs to be classified i.e. does it belong to class A or B.

d. Given the distribution, closer the point is to the origin, it is unlikely to belong to class A. Farther away it is from the origin, likely it belongs to class A

e. One can try to fit a simple linear model (y =mx+c) where y greater than a threshold means point most probably belongs to class A. The challenge is, for extreme values of x, probability is <0 or >1 which is absurd



Probability model

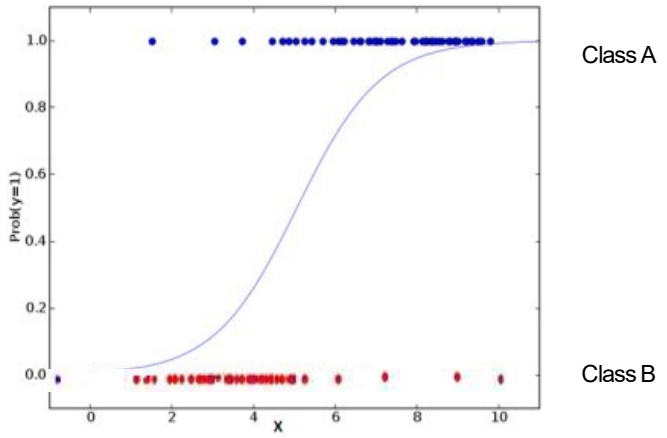


Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression Model -

f. The linear model is passed to a logistic function p =1/ 1 +e^-t the result of which is values between 0 and 1. Thus p represents probability a data point belongs to class "A" given x



Note: The linear model t(which is of the form mx+c), represents logit which is natural log(p / 1-p) where p is probability that a data point belongs to a class or not

Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression Model -

g. Uses logloss function to find the best fit line from the infinite possibilities where

$$logLoss = \frac{-1}{N}\sum_{i=1}^{N}(y_i(logp_i) + (1-y_i)log(1-p_i))$$

h. The objective is to make logLoss as large negative number as possible

i. There can be four difference cases for the value of yi and pi

Case 1 : $y_i = 1$, p_i = High , $1 - y_i = 0$, $1 - p_i$ = Low	Correct classification
Case 2 : $y_i = 1$, p_i = Low , $1 - y_i = 0$, $1 - p_i$ = High	Incorrect classification
Case 3 : $y_i = 0$, p_i = Low , $1 - y_i = 1$, $1 - p_i$ = High	Correct classification
Case 4 : $y_i = 0$, p_i = High , $1 - y_i = 1$, $1 - p_i$ = Low	Incorrect classification

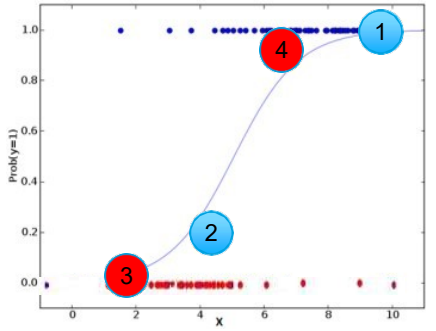
j. Incorrect classification contributes very minimal to the sum while a correct classification contributes large magnitudes

Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression Model -

k. In case1 y =1 and p =high implies that we have got things right!. It significantly inflates the sum because, Yi * log (Pi) would be high while the other term in the would be zero since 1 - Yi =1 - 1 =0.



l. So more occurrences of Case 1 would inflate the sum and consequently inflate the mean

m. In case2, y =1 and p is low which is incorrect classification. p =low, Yi * log (Pi) would not inflate the sum as much, the second term would be zero since 1- yi would be zero. So Case 2 would ultimately not affect the sum a lot.

n. Similarly the occurrences of Case 3 would inflate the sum significantly because first term would be 0 but second term will be high i.e. (1- yi)* log(1 – pi)

o. Case 4 first term will be 0 while in second term due to high pi, (1 – pi) will also be small hence contribution will be small

Supervised Machine Learning

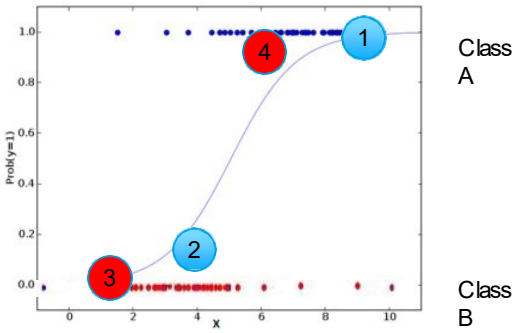
greatlearning
Learning for Life

Logistic Regression Model -

p. More of Case 1s and Case 3s increase the magnitude of the sum inside the logloss formula and because of the negative sign, make it overall error smaller and smaller

q. More Case2s and Case4s will not have as bit an impact on the overall value

r. The objective is to find the logistic curve that makes the overall logloss as negative as possible



Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression Model -

Advantages -

1. Makes no assumptions about distributions of classes in feature space
2. Easily extended to multiple classes (multinomial regression)
3. Natural probabilistic view of class predictions
4. Quick to train
5. Very fast at classifying unknown records
6. Good accuracy for many simple data sets
7. Resistant to overfitting
8. Can interpret model coefficients as indicators of feature importance

Dis advantages -

1. Constructs linear boundaries

Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression Model -

Lab- 2- Predict diabetes among Pima Indians

Description – Sample data is available at
https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.names

The dataset has 9 attributes listed below

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Sol: Logistic Regression - Lima Diabetes.ipynb

Supervised Machine Learning

greatlearning
Learning for Life

Confusion Matrix -

1. A tool to assess the performance of a classification model such as logistic regression model

1 = 231	Predicted 0	Predicted 1	Row Total
Actual 0	134	13	147
Actual 1	38	46	84
Col Total	172	59	231

2. Of the 84 actual diabetes case, the model correctly classified only 46 as diabetic

3. Of the 147 non diabetic cases, the model correctly classified 134 as non-diabetic

4. 13 cases who are normal but identified as diabetic are called Type 1 error

5. 38 cases of diabetic patients identified as normal is Type II error

Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier

Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

a. Naive Bayes classifiers are linear classifiers based on Bayes' theorem. The model generated is probabilistic

b. It is called naive due to the assumption that the features in the dataset are mutually independent

c. In real world, the independence assumption is often violated, but naive Bayes classifiers still tend to perform very well

d. Idea is to factor all available evidence in form of predictors into the naïve Bayes rule to obtain more accurate probability for class prediction

e. It estimates conditional probability which is the probability that something will happen, given that something else has already occurred. For e.g. the given mail is likely a spam given appearance of words such as "prize"

f. Being relatively robust, easy to implement, fast, and accurate, naive Bayes classifiers are used in many different fields

Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

Probability - is the number of trials in which an event of interest occurred by total number of trials. (what is a trial and what is an event?)

If it rained 3 out of 10 days in the past where the days were exactly like today, the probability it will rain today is 30%

a. In this example, the day is a trial / experiment
b. The event is rain
c. Probability that it will rain is P(A) (A denoting rains) =3/10
d. Every trial has at least two outcomes (event will P(A) or will not occur P(A”)
e. The multiple possible events are mutually exclusive i.e. cannot occur simultaneously
f. Total probability in a trial is sum of probabilities of all events =P(A)+P(A”) =100% =1

Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

Joint Probability – is the probability of multiple events occurring together (we are not talking of causality here i.e. one event leads to another). For e.g.

1. probability of drawing a king from a deck of cards is 4/52
2. Probability of drawing a red colour card from a deck of cards is 26/52
3. Probability of drawing a red colour king =2 / 52

Conditional Probability – it is the probability that an event has occurred (not yet observed) given another event has occurred. For e.g.

1. given the card drawn is red (an event has occurred)
2. what is the probability it is a king (event not yet observed)?
3. Since the card is red, there are 26 likely values for red
4. Of these 26 possible values we are interested in king which is 2 (king of diamonds and heart)
5. Thus the conditional probability that the card is a king given red card is 2 /26
6. Compare this with joint probability of red king (2/52).
7. Given an event has occurred, it increases the probability of the other event

Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

Posterior probability – Bayes' rule can be expressed as

$$\text{posterior probability} = \frac{\text{conditional probability} \cdot \text{prior probability}}{\text{evidence}}$$

The posterior probability, in the context of a classification problem, can be interpreted as: "What is the probability that a particular object belongs to class given its observed feature values?"

Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

Posterior probability – general expression

$$\text{posterior probability} = \frac{\text{conditional probability} \cdot \text{prior probability}}{\text{evidence}}$$

$$P(\omega_j \mid \mathbf{x}_i) = \frac{P(\mathbf{x}_i \mid \omega_j) \cdot P(\omega_j)}{P(\mathbf{x}_i)}$$

- $\mathbf{x}_i$ be the feature vector of sample i, $i \in \{1, 2, ..., n\}$,
- ω_j be the notation of class j, $j \in \{1, 2, ..., m\}$,
- and $P(\mathbf{x}_i \mid \omega_j)$ be the probability of observing sample $\mathbf{x}_i$ given that is belongs to class ω_j.

Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

The objective function is to maximize the posterior probability given the training data

$$\text{predicted class label} \leftarrow \underset{j=1\ldots,m}{\arg\max}\, P(\omega_j \mid \mathbf{x}_i)$$

person has diabetes if P(diabetes | xi) ≥ P(not-diabetes | xi), else classify person as healthy.

Introduction to machine learning

greatlearning
Learning for Life

Naïve Bayes Classifier (Assumptions) -

One assumption that Bayes classifiers make is that the samples are independent and identically distributed. Samples are drawn from a similar probability distribution.

Independence means that the probability of one observation does not affect the probability of another observation (e.g., time series and network graphs are not independent)

An additional assumption of naive Bayes classifiers is the conditional independence of features. Under this naive assumption, the class-conditional probabilities or (likelihoods) of the samples can be directly estimated from the training data instead of evaluating all possibilities of x

Thus, given a d-dimensional feature vector x, the class conditional probability can be calculated as follows:

$$P(\mathbf{x} \mid \omega_j) = P(x_1 \mid \omega_j) \cdot P(x_2 \mid \omega_j) \cdot \ldots \cdot P(x_d \mid \omega_j) = \prod_{k=1}^{d} P(x_k \mid \omega_j)$$

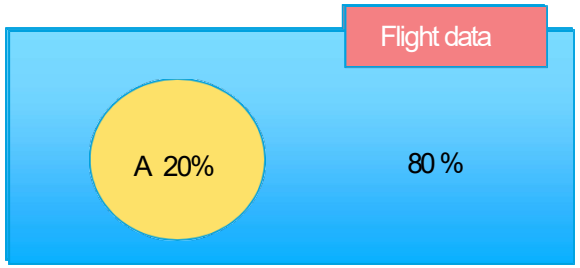
Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

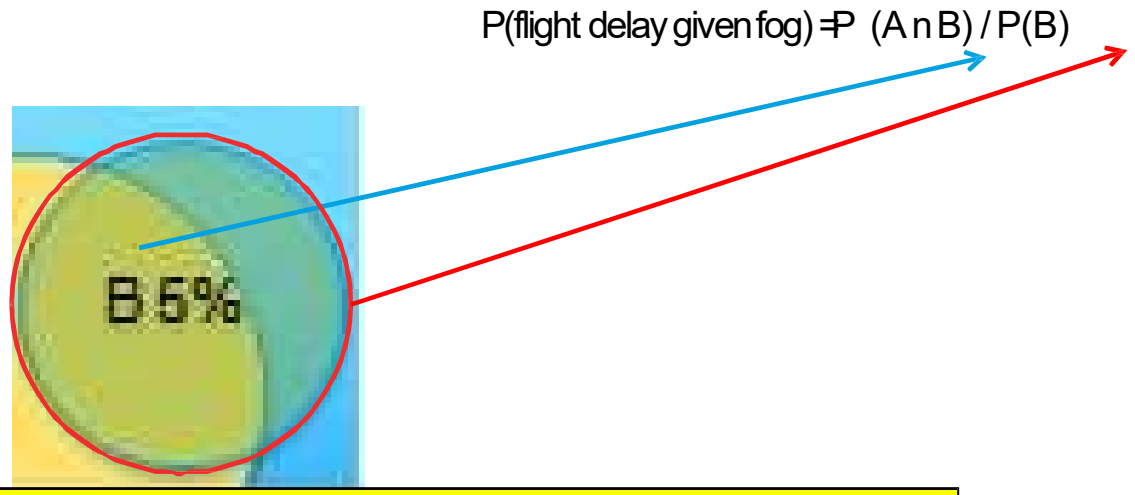
Joint Probabilities -

a. Imagine you represent all the flight experience you had till date as the blue area in a mathematical space. The dimensions of the boxes and circles are immaterial



Introduction to machine learning

Naïve Bayes Classifier -



More the overlap, more the occurrences of flight delay and fog P(A| B) =

P(A n B) / P(B) eq 1

Introduction to machine learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

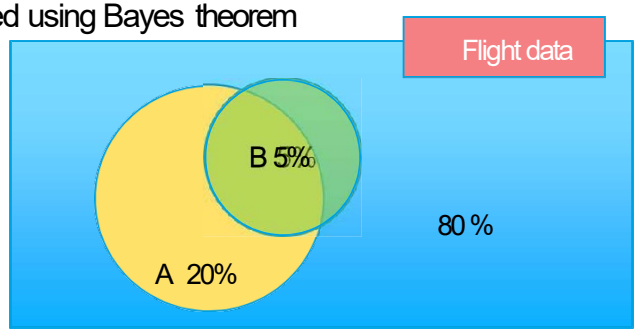
Joint Probabilities (Contd…) -

a. The relationship between dependent events is depicted using Bayes theorem

Likelihood Prior prob
Posterior

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Evidence



b. Probability of event A given that event B has occurred (fog has formed) depends on
   I. Apriori probability of fog occurring whenever there was flight delay – P (B/A)
   II. Apriori probability of flight delay P(A) which is 20% in the example
   III. Apriori probability of flight facing fog P(B) which is 5% in the example

c. When it is a matter of deciding the class of an output such as whether flight will get delayed or not, we calculate P(A/B) and P(!A/B), compare which is higher. Since in both the denominator is P(B), it is ignored as it has no influence on which class will it be

d. However, to calculate the updated probability of a class, denominator P(B) is required

Introduction to machine learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

a. The following two tables reflect the apriori probabilities of the events A and B. Probabilities based on past data of 100 points

T1	FOG		
Frequency	Yes	No	Total
Flight delayed	4	16	20
Not Delayed	1	79	80
Total	5	95	100

T2	FOG		
Likelihood	Yes	No	Total
Flight delayed	4 / 20	16 / 20	20
Not Delayed	1 / 80	79 /80	80
Total	5 / 100	95 / 100	100

b. In the likelihood table (T2) reveals that P(fog =Yes / flight delayed) =4/20 =.20 indicating that the probability is 20 percent that a flight is delayed due to fog

c. $P(A \cap B) \Rightarrow$ P(flight delay | fog) =P(fog / flight delay) * P(flight delay)

d. P(flight delay | fog) =((4/20) * (20 / 100)) =.04 (maximal probability) (no need to divide by P(B), probability of fog, as it is a constant. This is Naïve Bayes probability.

e. Naïve probability is when the event of flight delay and fog were unrelated (false independence) P( ) = $\bar{A} \cap B$ 100) * (5/100)) =.01 This indicates importance of Bayes theorem

Introduction to machine learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

Suppose there are multiple factors that could lead to flight delay (as shown in the likelihood table below)

	FOG		Technical Snag		Pilot Fatigue		Passenger		
Likelihood	Yes	No	Yes	No	Yes	No	Yes	No	Total
Flight delayed	4/20	16/20	10/20	10/20	0/20	20/20	12/20	8/20	20
Not Delayed	1/80	79/80	14/80	66/80	8/80	71/80	23/80	57/80	80
Total	5/100	95/100	24/100	76/100	8/100	91/100	35/100	65/100	100

a. Probability that the flight will be delayed, given that Fog =Yes, Technical Snag =No, Pilot Fatigue =No, Passenger related =Yes is given as –

P(flight delay / Fog n !Technical Snag n !Pilot Fatigue n Passenger delay) =(P(Fog n !Technical Snag n ! Pilot Fatigue n Passenger delay) / flight delay) * P(flight delay)) / P(Fog n !Technical Snag n !Pilot Fatigue n Passenger delay)

P(Fog n !Technical Snag n !Pilot Fatigue n Passenger delay) / flight delay) = P(Fog / flight delay) * P(!Technical snag |flight delay * P(!Pilot Fatigue | flight delay) * P(Passenger Delay| flight delay)

Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

Advantages -

1. Simple , Fast in processing and effective
2. Does well with noisy data and missing data
3. Requires few examples for training (assuming the data set is a true representative of the population)
4. Easy to obtain estimated probability for a prediction

Dis-advantages -

1. Relies on and often incorrect assumption of independent features
2. Not ideal for data sets with large number of numerical attributes
3. Estimated probabilities are less reliable in practice than predicted classes
4. If rare predictor value is not captured in the ** training set but appears in the test set the probability calculation will be incorrect

** For e.g. input record has fog="yes", Technical snag ="yes", Pilot Fatigue ="Yes" and passenger delay = "yes" .If this combination is not in the training set for delayed flights in the past then the probability calculation in step "a" on previous slide will become 0!

Supervised Machine Learning

greatlearning
Learning for Life

Naïve Bayes Classifier -

Lab- 4 Model to predict diabetes among Pima Indians

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

The dataset has 9 attributes listed below

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Sol: Naive+Bayesian+Pima+Diabetes+.ipynb