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
b. The response variable is categorical. It can be two category (binary class) or multi category (multi-class) variable
c. It is used to predict class given the predictor variable values of an observation
d. It can also be used to find the propensity (probability) with which an observation belongs to the various classes

Supervised Machine Learning

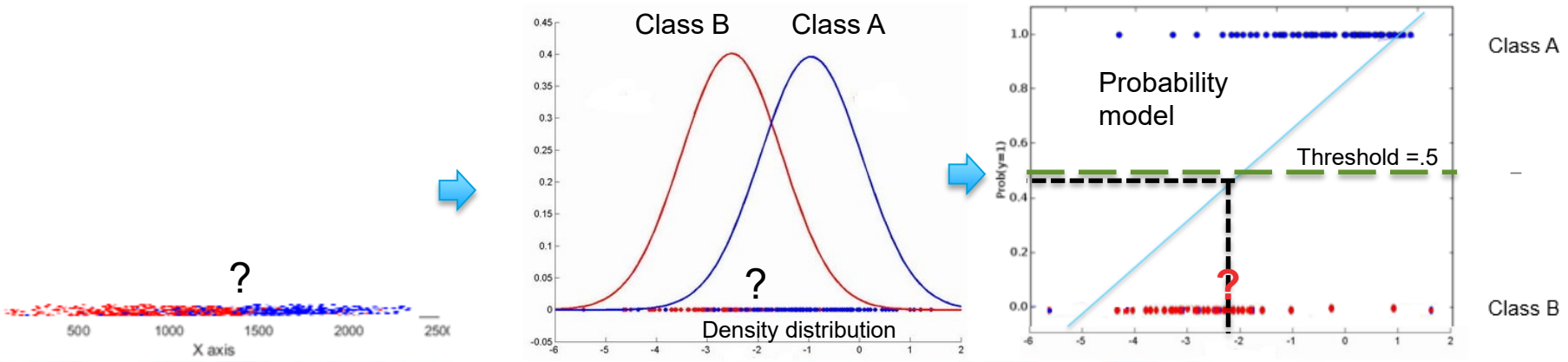
greatlearning
Learning for Life

Logistic Regression Model

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

b. 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. Let class A be 1 and class B be 0 on the vertical axis

c. 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. for extreme values of x, probability is <0 or >1 which is absurd

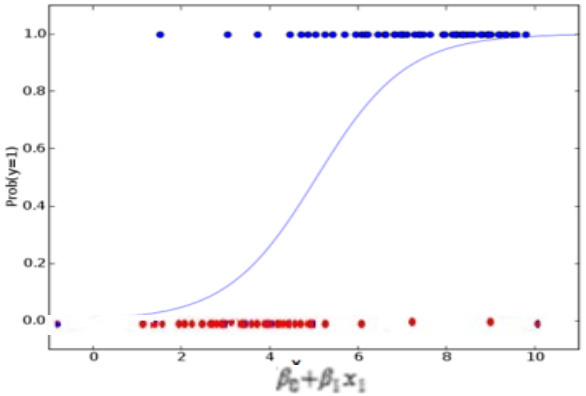


Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression Model -

d. 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



Class A

Class B

e. Instead of using y of linear model as dependent, it's function shown as "p" is used as dependent variable $p = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_c x_c)}}$. This is logistic response function

f. It is a two step model. In first step, the propensity to belong to class 1 i.e P(1|X), followed by next step of using cut-off to decide the class

Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression Model – The learning process

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

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

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

c. 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

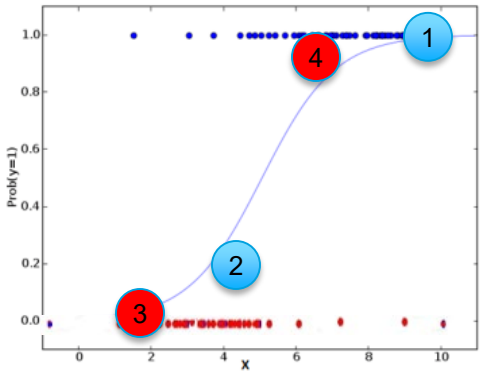
d. Incorrect classification contributes very minimal to the sum while a correct classification contributes large magnitudes

Supervised Machine Learning

greatlearning
Learning for Life

Logistic Regression Model -

e. 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.



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

g. 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.

h. 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)

i. 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

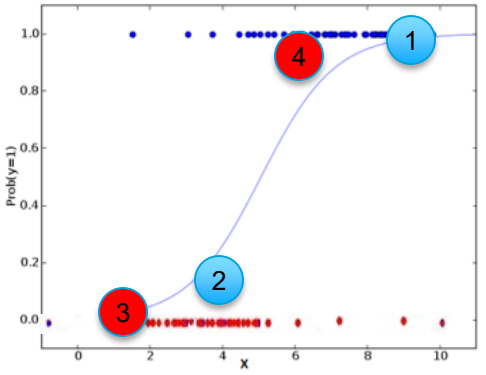
greatlearning
Learning for Life

Logistic Regression Model -

j. 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

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

l. 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