upGrad

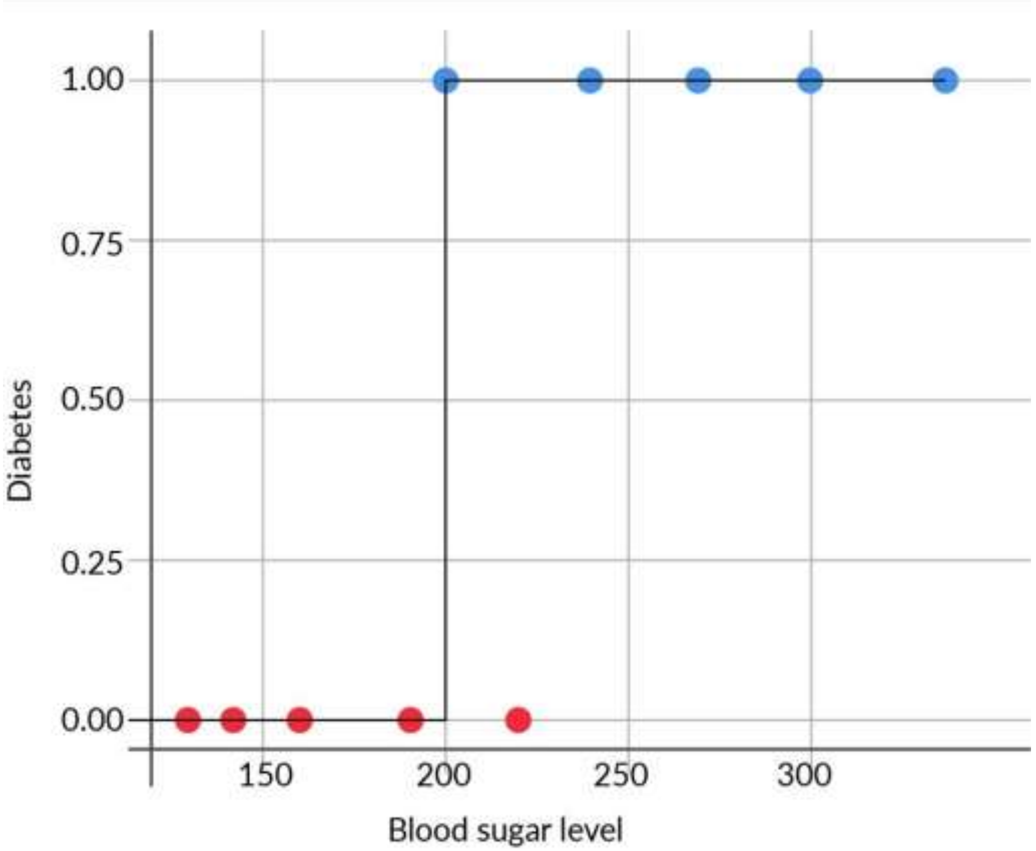
Logistic Regression

LOGISTIC REGRESSION

- Logistic regression models are used for classification problems.
- Binary Classification: in which the target variable has only 2 possible values.
  - Finance: Customer will default on loan or not
  - Email: Spam or not
  - Diabetes: Yes or No
- Multi-class classification: in which the target variable has more than 2 possible values.
  - Categorize email into primary, social, promotions.

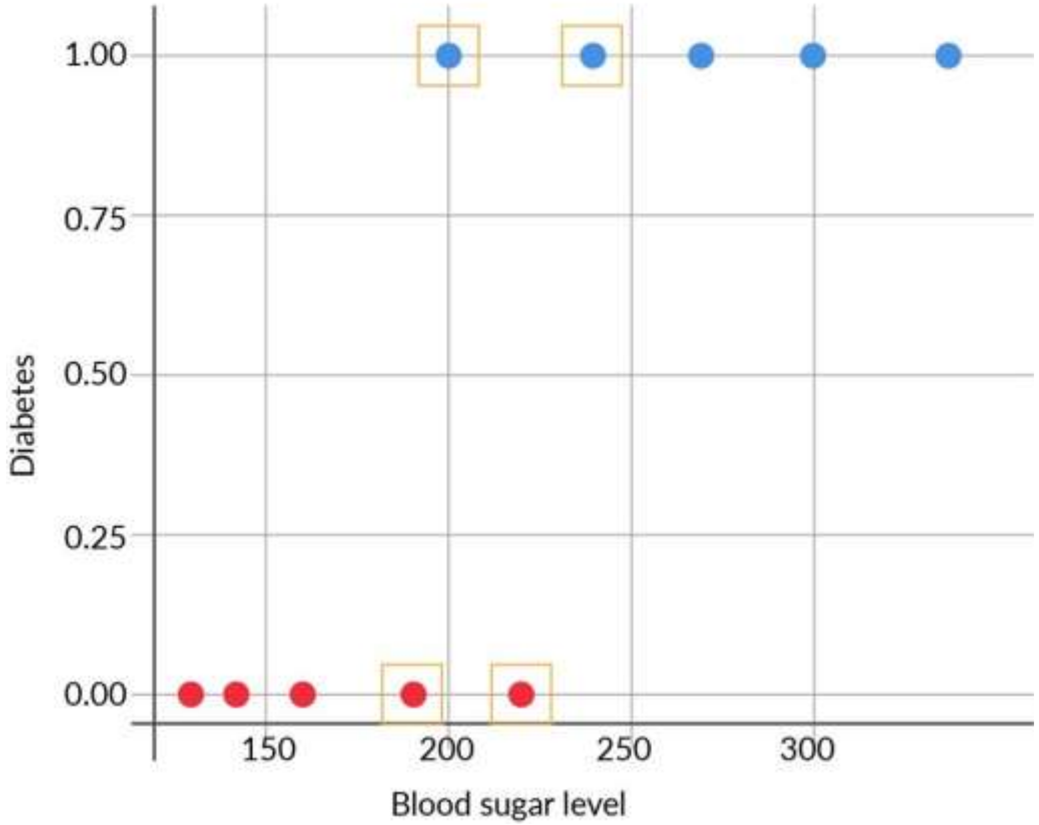
LOGISTIC REGRESSION

Diabetes problem: predict whether a person has diabetes or not based on that person’s blood sugar level.



LOGISTIC REGRESSION

Diabetes problem: predict whether a person has diabetes or not based on that person's blood sugar level.

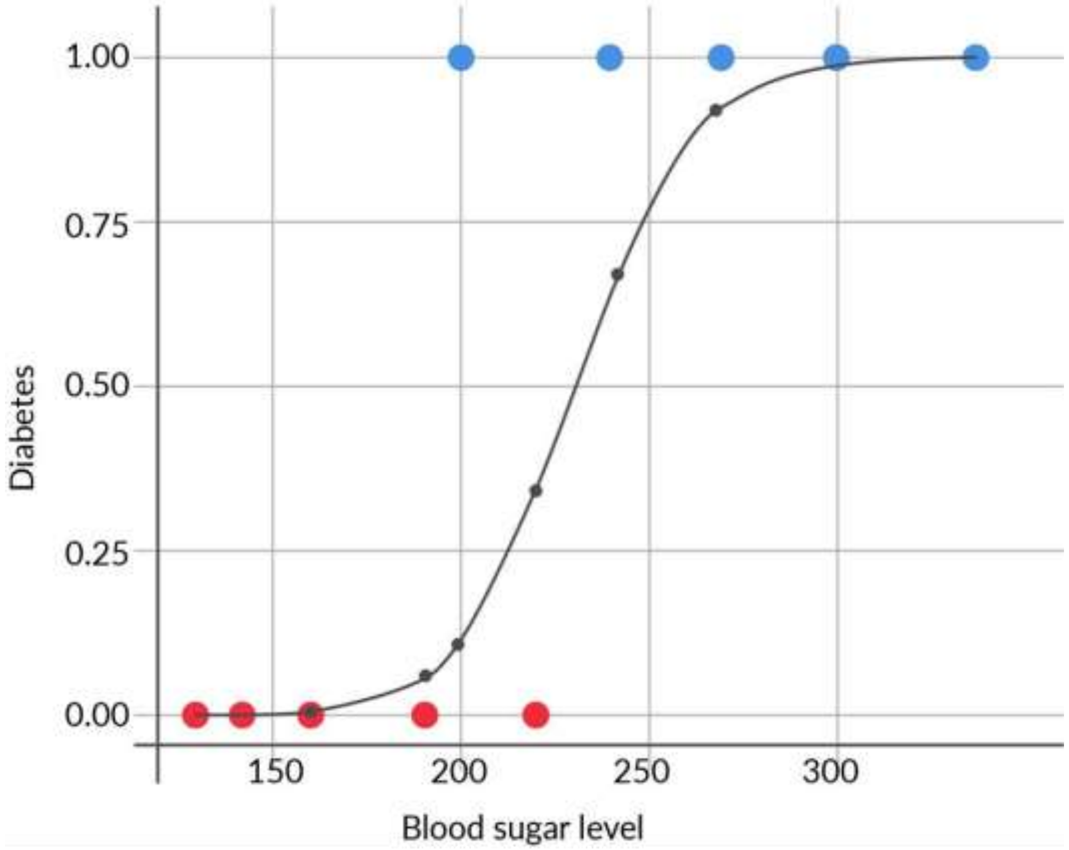


LOGISTIC REGRESSION

- Simple decision boundary approach does not work very well.
- It would be too risky to decide the class blatantly on the basis of the cutoff because, especially in the middle, the patients could belong to any class — diabetic or non-diabetic.
- So instead of sharp decision boundary will use a smooth curve:
  - Sigmoid Curve.
  - It gives probability of diabetes at any x [blood sugar level]

LOGISTIC REGRESSION

Sigmoid Curve

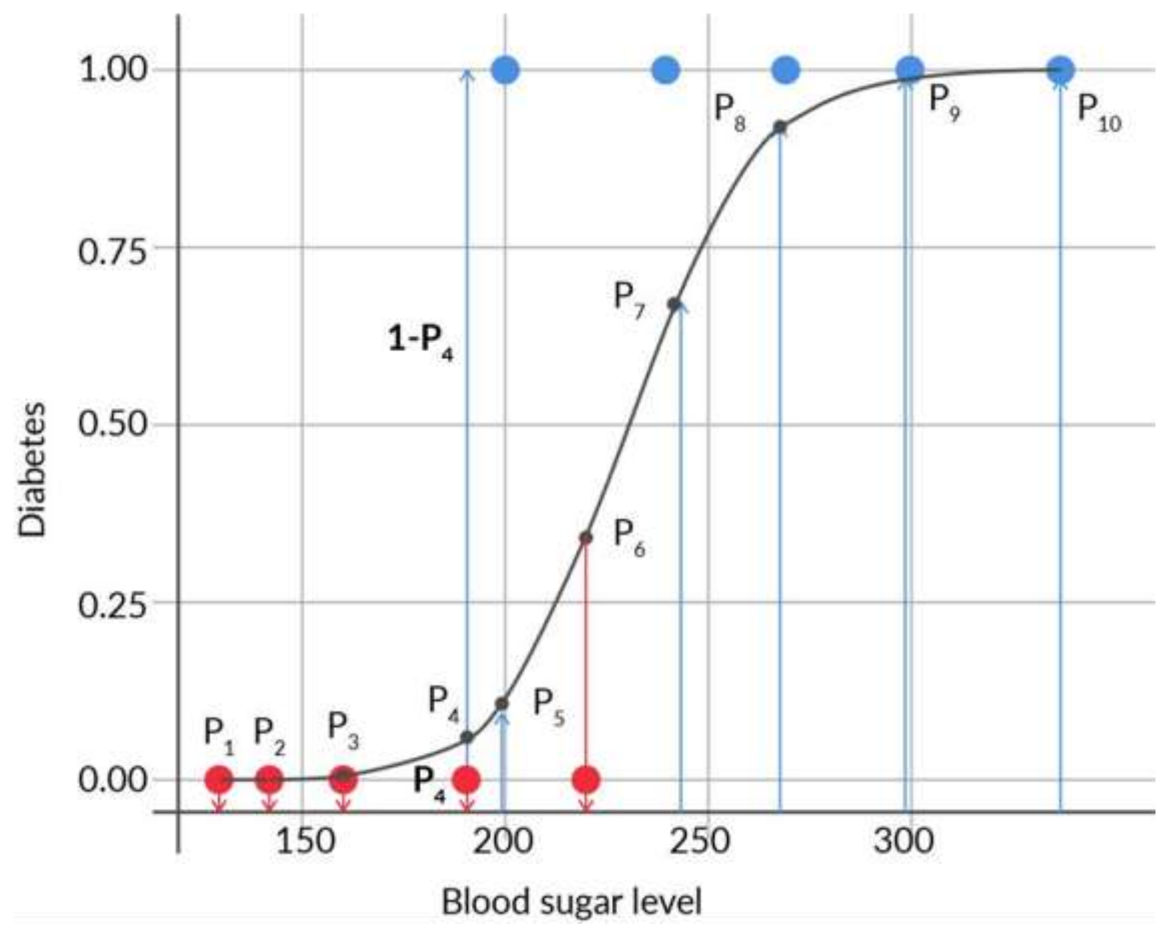


LOGISTIC REGRESSION

Equation For Sigmoid Curve:

$$y\ (\text{Probability of Diabetes}) = \frac{1}{1+e^{-(\beta_0 + \beta_1 X)}}$$

LOGISTIC REGRESSION



LOGISTIC REGRESSION

- Finding the best fit sigmoid curve
- Find the combination of β_0 and β_1 which maximises the likelihood.
- For the diabetes example, the likelihood is given by the expression:

$$\text{Likelihood}=(1-P_1)(1-P_2)(1-P_3)(1-P_4)(P_5)(1-P_6)(P_7)(P_8)(P_9)(P_{10})$$

- The best fitting sigmoid curve would be the one which maximises the value of this product.

LOGISTIC REGRESSION

- Different values of β_o and β_1 gives different shape of the sigmoid curve.
- At some combination of β_o and β_1 the 'likelihood' will be maximised.
- To find the optimal values of β_o and β_1 :
  - The optimisation method maximum likelihood estimation (MLE) is used.

Case Study: CTR Prediction

Case Study: Online Advertising

- In online advertising, click-through rate (CTR) is a very important metric for evaluating ad performance.
- As a result, click prediction systems are essential and widely used for sponsored search and real-time bidding.
- CTR is basically rate of how many users clicked on Ad with respect to how many times the Ad was displayed.
- CTR = Clicks/Impressions

Dataset

- Show dataset in Excel
- Further details about Data can be explored in Kaggle:

  https://www.kaggle.com/c/avazu-ctr-prediction/data

Data Fields

1. id: ad identifier
2. click: 0/1 for non-click/click
3. hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.
4. C1 -- anonymized categorical variable
5. banner_pos
6. site_id
7. site_domain
8. site_category
9. app_id
10. app_domain
11. app_category
12. device_id
13. device_ip
14. device_model
15. device_type
16. device_conn_type
17. C14-C21 -- anonymized categorical variables