Logistic Regression Intro

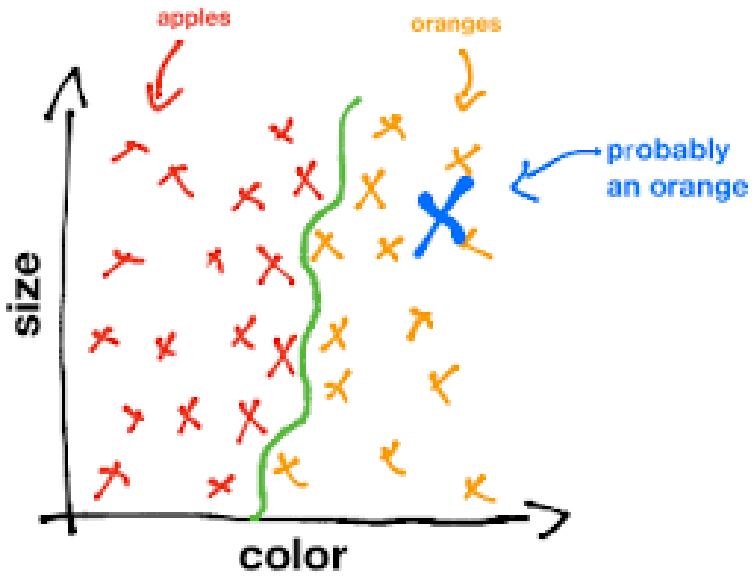
By

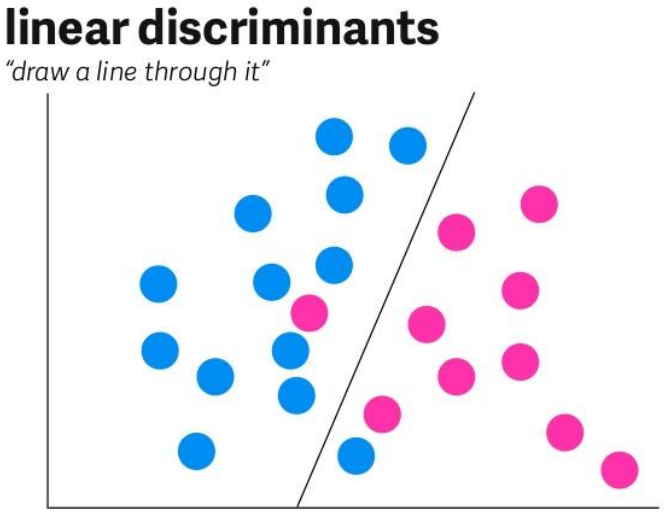
Amritansh

M.S. in CS (Concentration in AI)

AI Researcher (Core AI, Deep Learning, Swarm Robotics)

Mentor of Change @ Atal Innovation Mission (AIM) by Niti Aayog





Classification

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known

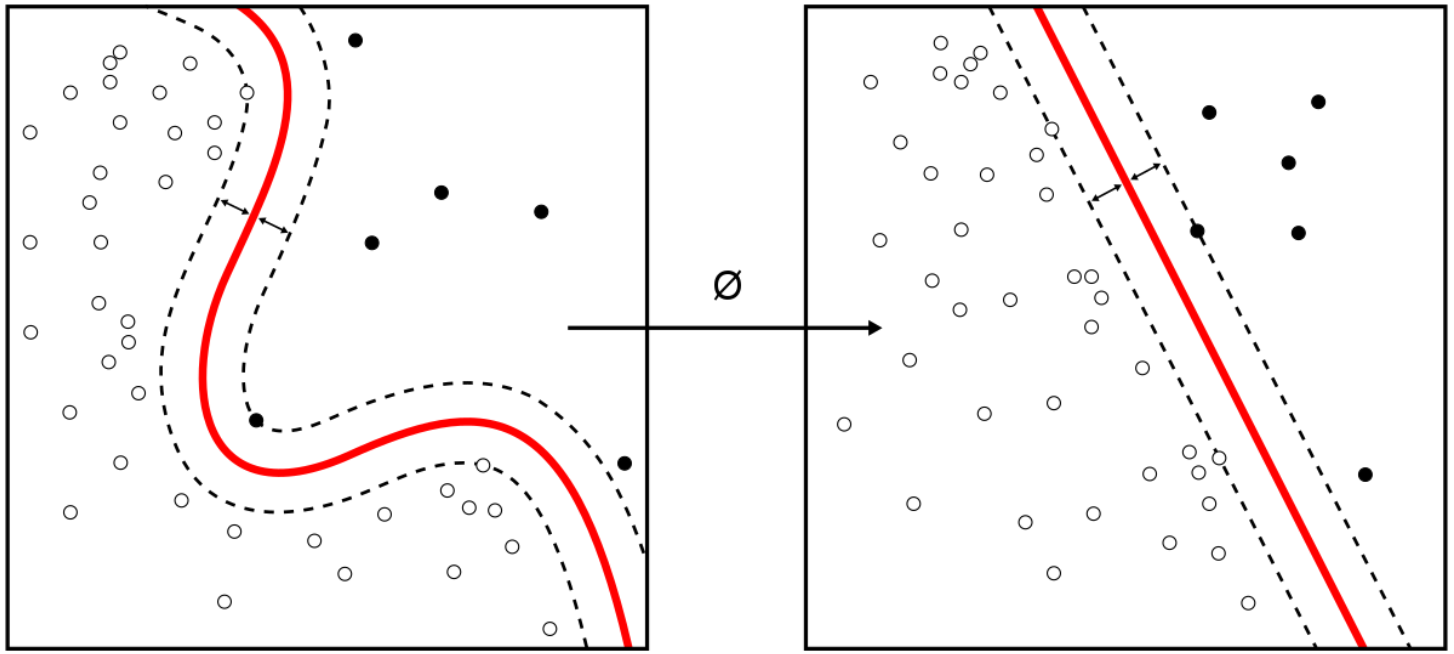
Classification

- Y is discrete (0,1) or (A, B)

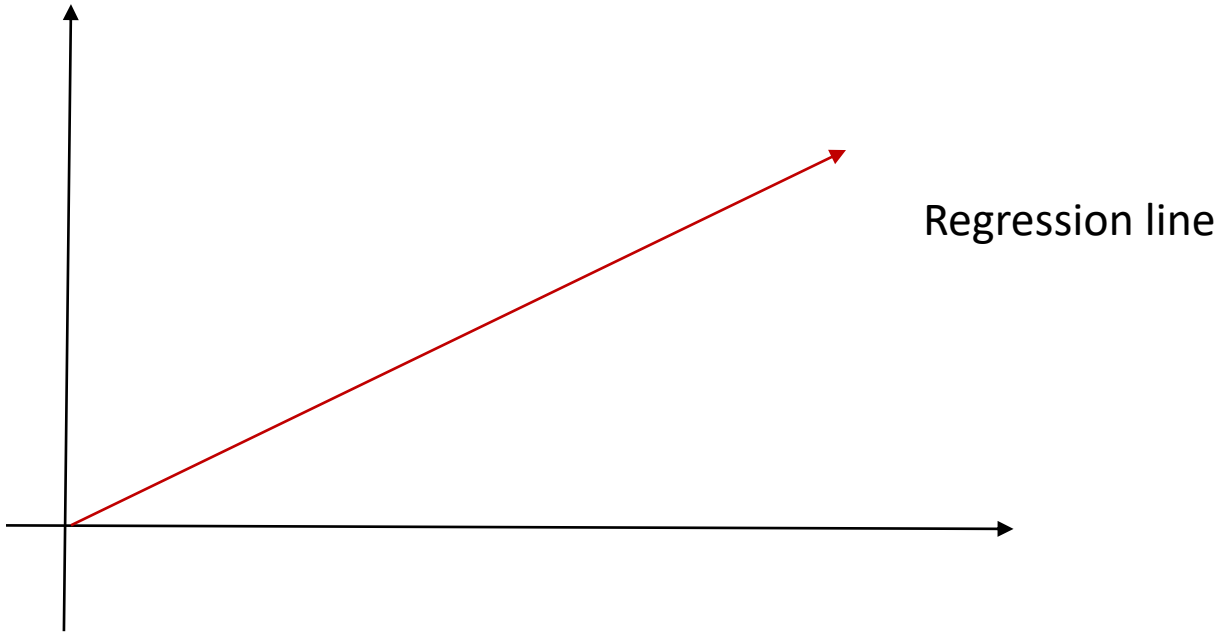
  Email : Spam or Not Spam

  Cancer : Benign or Malignant

  Food Types: Vegan, Vegetarian, Non – Vegetarian



Why can’t we use Linear Regression ?



Logistic Regression

- Logistic Regression was developed by statistician David Cox in 1958

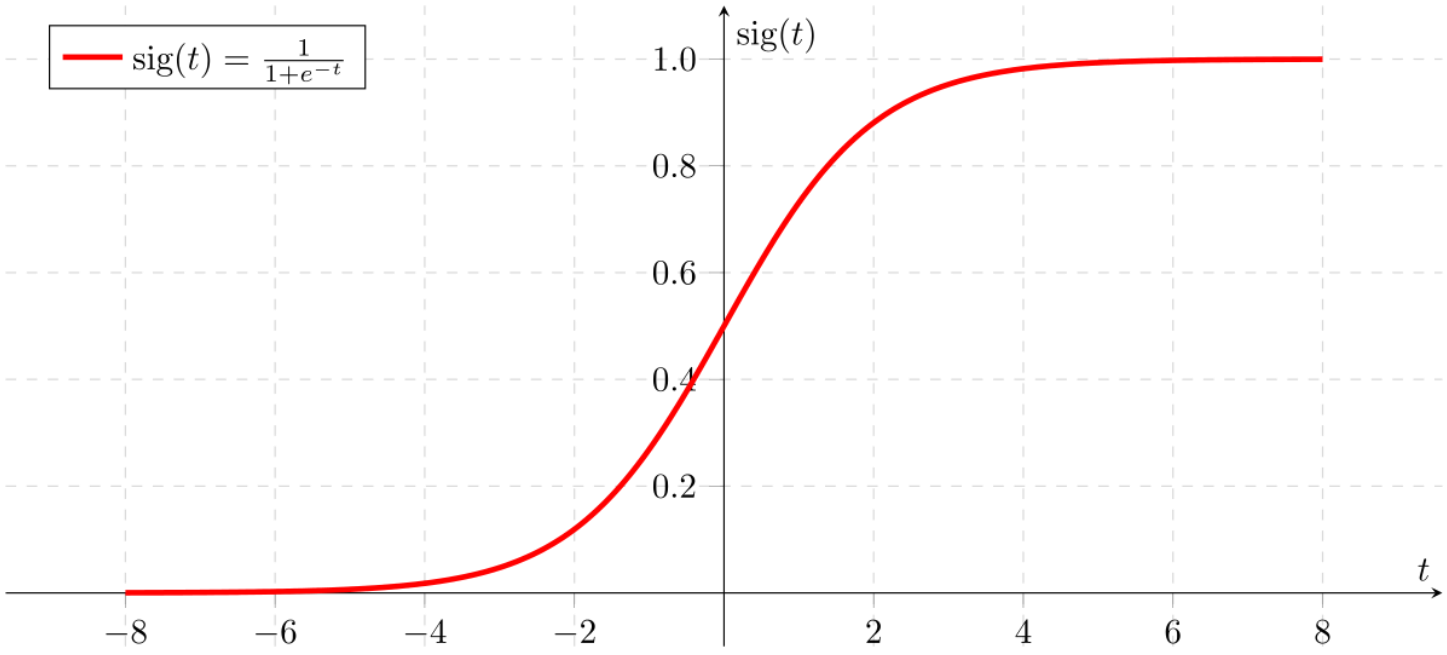
- The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor.

- Logistic Regression was used in the biological sciences in early twentieth century.

Sigmoid Function

A sigmoid function is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point.

Sigmoid is synonymous to Logistic function



Interpretation

- Sigmoid function outputs a probability of P(1)

- Thus if we get P = 0.75, it means there is 75% probability of happenstance of 1
- If we get P = 0.3, then there is 30% probability of happenstance of class 1, and simultaneously 70% probability of Happenstance of Zero(0)

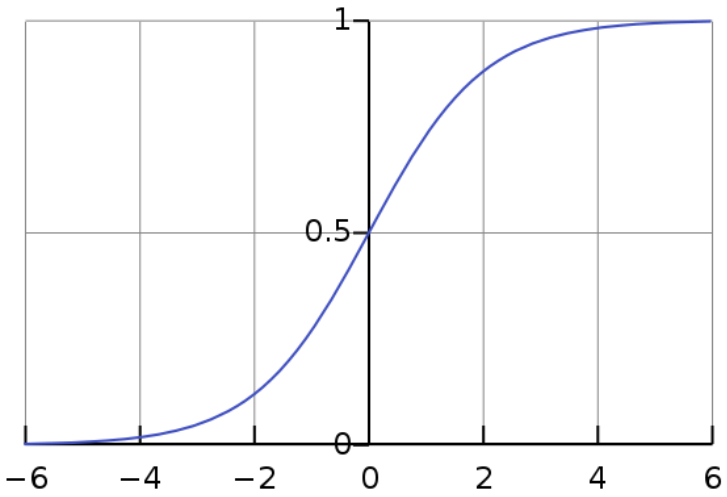
- Probability needs to be 1 when all partial probabilities are added : P(1) + P(0) = 1

  Thus, P(0) = 1- P(1)

Hypothesis

- If you remember hypothesis in linear regression:
  - $y = \Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \ldots\ldots\ldots + \Theta_n x_n$

Hypothesis (Logistic)



As we see in sigmoid if the input is positive then we get value more than 0.5

Thus, we need input hypothesis to be greater than Zero (0)

Cost function

Instead of Mean Squared Error, we use a cost function called Cross-Entropy, also known as Log Loss. Cross-entropy loss can be divided into two separate cost functions: one for y=1 and one for y=0.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\text{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad \text{if } y = 1$$

$$\text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad \text{if } y = 0$$

The benefits of taking the logarithm reveal themselves when you look at the cost function graphs for y=1 and y=0. These smooth monotonic functions (always increasing or always decreasing) make it easy to calculate the gradient and minimize cost.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

$$h = g(X\theta)$$

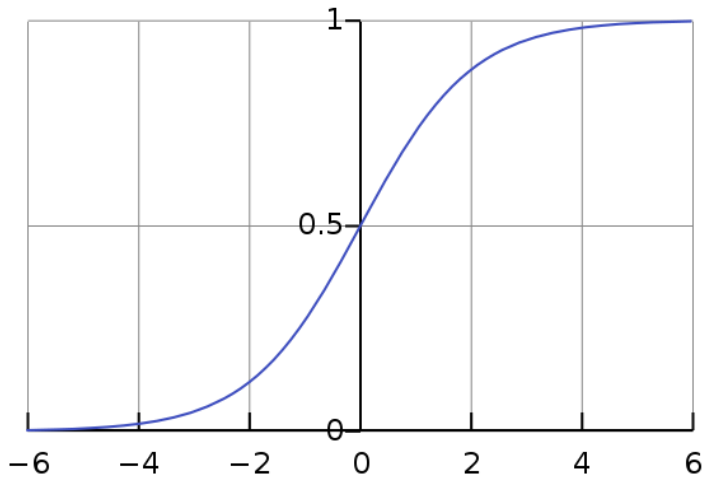
$$J(\theta) = \frac{1}{m}\cdot\left(-y^T\log(h) - (1-y)^T\log(1-h)\right)$$

Vectorized form of Cost function

Decision Boundary

- In case of Binary Classification
  - To predict which class a data belongs, a threshold can be set.
    if predicted_value ≥ 0.5, then classify email as spam else as not spam.

- We want sigmoid(x) > 0.5, thus we need hypothesis to be positive



Logistic regression for Multiclassifcation

One vs All

One vs All

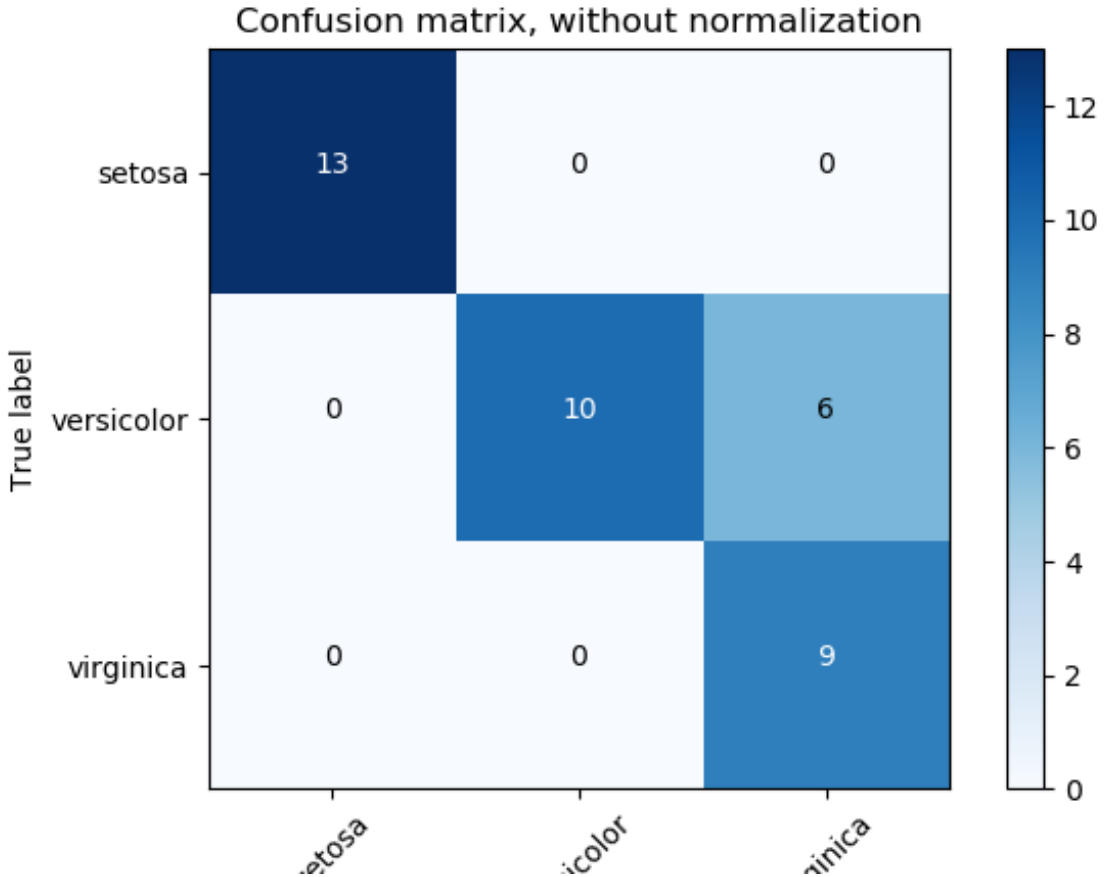
One vs All

- Run individual logistic regression on all 3 hypothesis considering 1 class group as positive class
- Calculate probabilities for 3 such hypothesis

- Final output =

Confusion Matrix

- A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.



Terminologies

- **Accuracy**: Overall, how often is the classifier correct?

  (TP+TN)/total = (100+50)/165 = 0.91

- **Error Rate**: Overall, how often is it wrong?

  (FP+FN)/total = (10+5)/165 = 0.09
  Also equivalent to (1 – Accuracy)

- **Precision**: When it predicts yes, how often is it correct?
  TP/(FP + TP) = 100/(100 +10) = 0.91

- **Recall**: How many true positives were found?
  TP /(TP + FN) = 100 / (100 +5) = 0.95

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

TN = True Negatives
TP = True Positives
FN = False Negatives
FP = False Positives

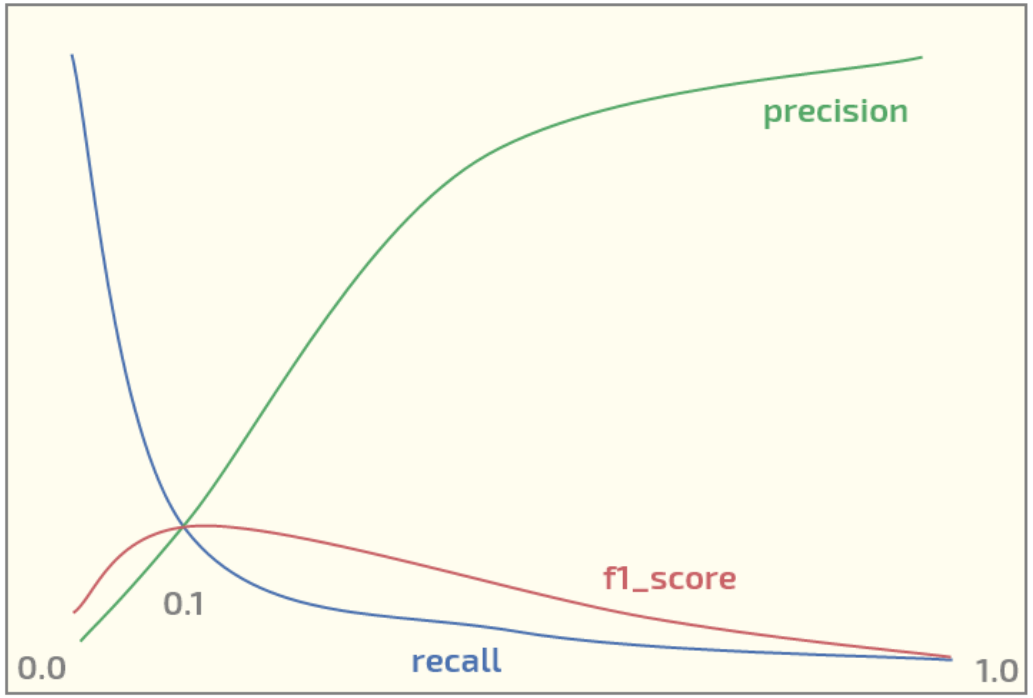
F1 Score

F1 score is the harmonic mean of Precision and Recall values of a trained classifier model

Mathematically,

F1 = 2 *(Precision x Recall)/(Precision + Recall)

The F1 score(or F-Score) conveys the balance between the precision and the recall



Advantages

- The output of a logistic regression is more informative than other classification algorithms. Like any regression approach, it expresses the relationship between an outcome variable (label) and each of its predictors (features).

- Easy to design and run across various datasets

- Correlations are explainable thus providing transparency in creating decision boundaries

Use Cases

- Image Classification
- Spam filtering
- Medical Diagnosis
- Drug trial analysis
- Customer segmentation
- Fraud detection
- Object segmentation
- Digit recognizer
- OCR

Thank You!!!