Nov - 26:

Classification:

Why not linear regression on classification data!

* As the output is not continuous in the range of $-\infty$ to $+\infty$ we can't go with too linear regression.
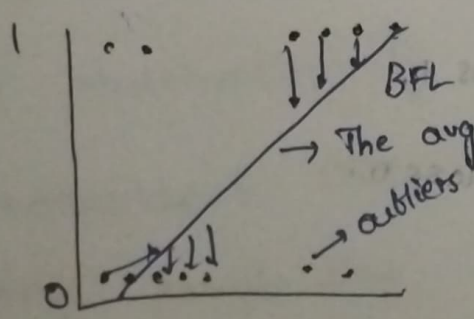
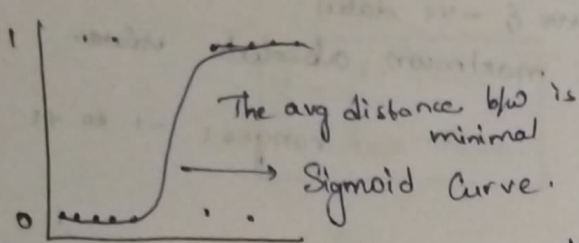* The linear regression's best fit line will not be suitable for categorical data.

Ex:

BP	BT	Cold	Headache	Fever

Fever
$Y - 1$
N - 0   categorical
N
Y

continuous values
$-\infty$ to ∞

$$\sigma(z) = \dfrac{1}{\frac{1}{x} - 9 + 1}$$

→ BFL

→ The avg distance b/w Actual & predicted is maximum.

→ outliers

The avg distance b/w is minimal

→ Sigmoid Curve.

Continuous variables

0 to 1.

>0.5 = 1

>= 0.5 = 0.

* A Sigmoid curve gives continuous output so it is referred as regression but based on probability, condition it is used for classification

Logistic Regression:

Logistic Regression is a supervised classification algorithm.

* It is used when the dependent variable (target) is categorical.

* Logistic Regression is most suitable for (0 and 1) **Binary**

* There are 2 types of classification.

1. Binary classification → Y/N, 0/1, T/F.

2. Multi-class classification → Weather (S/W/R), Rating (1/2/3/4/5)

Example: - Predicting whether a customer will buy. (Y/N)
- Predicting if an email is spam (spam/not spam)
- Predicting patient fever level (LBP, NBP, HBP) [BP]

Why not linear Regression?

* Linear Regression outputs continuous values that can go beyond 0 & 1.

* For classification we need probabilities b/w 0 and 1. So Logistic regression uses a 'sigmoid' function instead of a linear line equation. to convert prediction into probability.

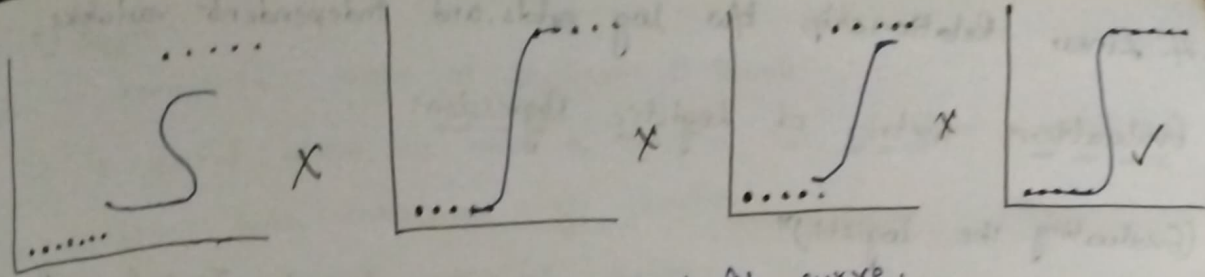## Sigmoid Function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

e: euler's value.
$e \approx 2.71828$.

* outputs always lies b/w 0 & 1.

* If probability >0.5 → predicts class 1.

* If probability =< 0.5 → predicts class 0.

The top of the page shows four small hand-drawn graphs with sigmoid/step curves, marked (left to right): ✗, ✗, ✗, ✓.

* Logistic Regression finds the best fit curve.

Logistic Regression Equation:

$$Z = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_n x_n + \beta.$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

probability of class 1:

$$P(y = 1/x) = \frac{1}{1+e^{-z}}$$

$$P(y = 1/x) = \frac{1}{1+e^{-(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots \alpha_n x_n + \beta)}}$$

$$\hat{y} = \frac{1}{1+e^{-(\alpha_1 x_1 + \alpha_2 x_2 - \cdots + \alpha_n x_n + \beta)}}$$

Types of logistic Regression:

1. Binary Logistic Regression (Two classes)

Ex: 0 = no disease

1 = disease.

2. Multinomial Logistic Regression (More than 2 unordered classes)

Ex: Predicting which brand customer chooses (A/B/c)

3. Ordinal Logistic Regression (More than 2 ordered classes)

Ex) High / Medium / Low Customer Satisfaction.

Assumption of Logistic Regression:

1. No multicollinearity b/w independent variables.

2. Independent Observation.

3. Large sample size is preferred.

4. Linear Relationship b/w log odds and independent variables.

Evaluation metrics of Logistic Regression:

(Evaluating the logistic)*

Classification problems aim to predict discrete categories. To evaluate the performance of classification model, we use the following metrics:

1. Accuracy:

Accuracy is a fundamental metric used for evaluating the performance of a classification model. It tells us the proportion of correct predictions made by the model out of all predictions.

$$Accuracy = \frac{No. of \ correct \ predictions}{Total \ No. of \ predictions}$$

Accuracy is good but it gives a false positive sense of achieving high accuracy. The problem arises due to the possibility of misclassification of minor class samples being very high.

2. Precision:

It measures how many of the positive predictions made by the model are actually correct. It's useful when the cost of false positives is high such as in medical diagnoses where predicting a disease when it's not present can have serious consequences.

$$Precision = \frac{TP}{TP+FP}$$

where,
TP = True Positive   FP = False positive

3. Recall:

Recall or Sensitivity measures how many of the actual positive cases were correctly identified by the model. It is important when a missing positive case (FN) is more costly than False positive

$$Recall = \frac{TP}{TP+FN}$$

4. F1 Score :

It is harmonic mean of precision & Recall.

* It is useful when we need a balance b/w precision & recall as it combines both into a single number.

* Higher F1 Score the model is better

* Range of F1 score 0 to 1.

* Lower recall & Higher precision → great accuracy but it then it misses a large no.of instances.

$$F1 \text{ Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. Logarithmic Loss (Log Loss) :

* It measures the uncertainty of model's predictions.

* It is calculated by penalizing the model for assing low probabilities to the correct classes.

* It is used in multi-class classification. & helpful to access model's confidence in it's predictions.

Ex: N samples belonging to M class.

$$Log \text{ Loss} = -\frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{M} y_{ij} \cdot log(P_{ij}) \right)$$

where y_{ij} : Actual class (0/1) for sample i & class j

Pij : Predicted probability "   "   "   "

Goal: Minimize log loss function. Lower logg loss → higher prediction accuracy.

6. Area Under Curve & ROC curve » Binary Classification Tasks

AUC → Represents the probability that the model will rank a randomly chosen positive example higher than a randomly chosen negative example.

* AUC ranges from 0 to 1

* Higher values → better model performance.

TPR (True positive Rate)/Recall:

$$TPR = \frac{TP}{TP + FN}$$

TNR (True Negative Rate)/Specificity:

$$TNR = \frac{TN}{TEN + FP}$$

FPR (False Positive Rate):

$$FPR = \frac{FP}{FP + TN}$$

False Negative Rate (FNR):

$$FNR = \frac{FN}{FN + TP}$$

ROC Curve:

It is a graphical representation of TPR & FPR at different classifie thresholds.

AUC = 1   perfect model.

AUC = 0.5   model is avg.

AUC < 0.5   worst model.

7. Confusion Matrix:

It creates N×N matrix, where N is no. of classes that are to be predicted.

Ex:

n=165	Predicted No	Predicted Yes
Actual No	50	10
Actual Yes	5	100