Data Science & AI

Machine Learning

Supervised Learning

Lecture No.- 03

By- Krish Naik Sir

Recap of Previous Lecture

- Topic
- Topic
- Topic
- Topic
- Topic

Regression

Topics to be Covered

Topic Polynomial

Topic Logistic

Topic

Topic

Topic

Machine Learning

Agenda

i) Multicollinearity ✓ [Bias Variance Tradeoff] ✓

ii) Polynomial Regression ✓

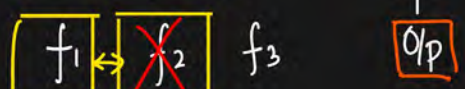
iii) Logistic Regression ✓

iv) Performance metrics [Accuracy, True Positive, False Positive, Precision, Recall] ✓.

0.90 ⇒ 0.92

① Multicollinearity.

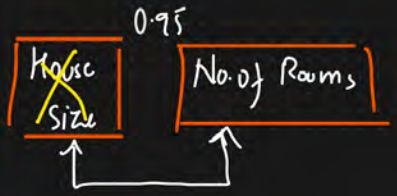
Dependent features



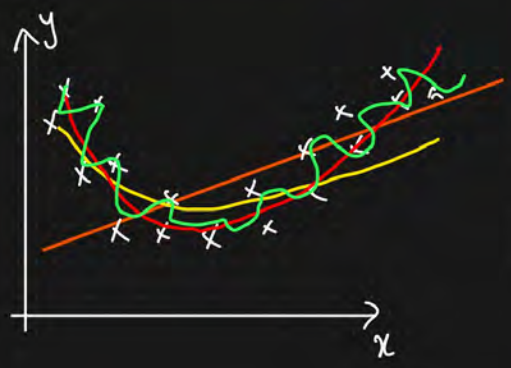
{f1, f2, f3} → Independent features

0.96 {Duplicate feature}

House Price Prediction



② Polynomial Regression



Underfitting

TRAINING ACC ↓↓ = 60%.

Test ACC ↓↓ = 65% 55%

Overfitting

Low Bias

Traning Acc ↑↑ ⇒ 95%.

Test Acc ↓↓ ⇒ 60%.

High Variance

$h_\theta(x) = \theta_0 + \theta_1 x_1$

Underfitting ⇒ High Bias, High Variance

Normalized model

High Bias, Low Variance ⇒ Underfitting.

Overfit.

Generalized or Perfect Model

TRAINING ACC ↑↑ } Low Bias
Test ACC ↑↑ } Low Variance

Polynomial Regression

Polynomial degree = 0 $h_\theta(x) = \theta_0 \times x_1^0 = \theta_0 * 1 = \theta_0$ ⟶ Intercept

Polynomial degree = 1 $h_\theta(x) = \theta_0 x_1^0 + \theta_1 x_1^1 = \theta_0 + \theta_1 x_1$ → {Simple Linear Regression}

Polynomial degree = 2 $h_\theta(x) = \theta_0 * x_1^0 + \theta_1 x_1^1 + \theta_2 x_1^2$ ⇐

Polynomial degree = 3 $h_\theta(x) = \theta_0 x_1^0 + \theta_1 x_1^1 + \theta_2 x_1^2 + \theta_3 x_1^3$

⋮

x_1 x_2 x_3

Polynomial degree = 2

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_1^2 + \theta_5 x_2^2 + \theta_6 x_3^2$$

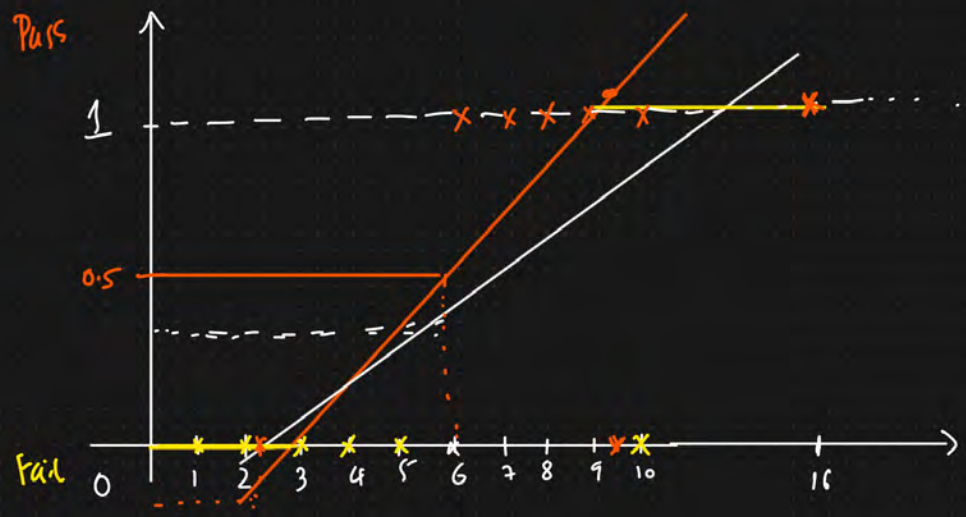
⑧ Logistics Regression ⟶ Classification ⟶ Binary classification

Dataset

UPSC

No. of study hours	Pass / Fail
1	Fail
2	Fail
3	Fail
4	

5	Fail
6	Pass
7	Pass
8	Pass
9	Pass
10	Fail



Logistic Regression

Why we cannot use linear Regression for classification

① Best fit line changes because of outlier → prediction goes wrong

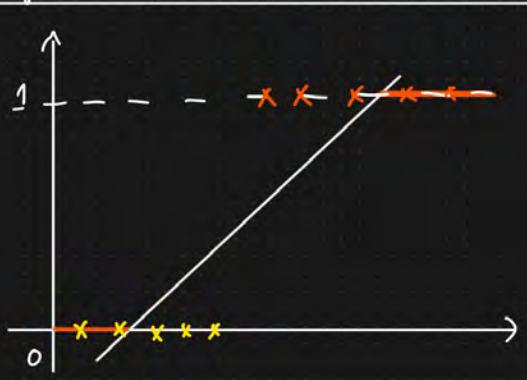
② The outcome comes >1 and < 0

To solve this problem we use Logistic Regression

⇓

0 to 1 ⇒ Squashing Technique

How Logistic Regression Solves classification Problem



$h_\theta(x) = \theta_0 + \theta_1 x_1$ ⇒ Best fit line

⇓

[Sigmoid Activation function]

⇓

0 to 1

Logistic Regression

$$h_\theta(x) = \sigma(\theta_0 + \theta_1 x_1)$$

z

$$\sigma = \frac{1}{1+e^{-z}} \Rightarrow 0 \text{ to } 1.$$

$$h_\theta(x) = \frac{1}{1+e^{-z}} \Rightarrow h_\theta(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}}$$

Linear Regression Cost fn

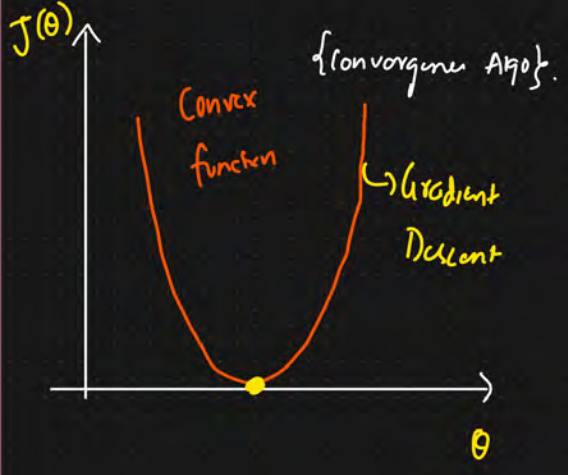
$$J(\theta_0,\theta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\theta(x)_i)^2$$

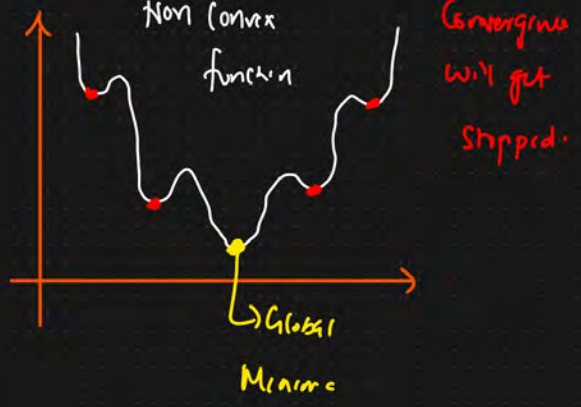
$$h_\theta(x) = \theta_0 + \theta_1 x_1$$

Logistic Regression Cost fn

$$J(\theta_0,\theta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\theta(x)_i)^2$$

$$h_\theta(x) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}}$$





Log Loss → Cost function → Logistic Regression

$h_\theta(x) \Rightarrow$ predicted o/p

$$J(\theta_0,\theta_1) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

$$J(\theta_0, \theta_1) = -y \log(h_\theta(x)) - (1-y) \log(1 - h_\theta(x))$$ ⇒ Convex function

↓

1 global Minima.

⇓

Minimize $J(\theta_0, \theta_1)$ changing θ_0 and θ_1

⋇ Performance Metrics, Accuracy, Precision, Recall and F-Beta

Topics to be covered

1. Confusion Matrix ✓
2. Accuracy ✓
3. Precision
4. Recall
5. F-Beta Score

DATASET

x_1	x_2	Actual y	Predicted y	
–	–	0	1	→ Wrong Prediction
–	–	1	1	→ Correct Prediction
–	–	0	0	
–	–	1	1	
–	–	1	1	
–	–	0	1	
–	–	1	0	

① Confusion Matrix

Predicted \ Actual	1	0
1	3	2
0	1	1

(Actual Value →, Predicted Value ↓)

Predicted \ Actual	1	0
1	TP	FP
0	FN	TN

$$\text{Accuracy Score} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{Accuracy Score} = \frac{3+1}{3+2+1+1} = \frac{4}{7} =$$

DATASET →

1000 Datapoints → 900 - 1, 100 - 0 } Imbalanced dataset

Dumb Model — 1 ⇒ Accuracy = 90%

③ Precision = $\frac{TP}{TP+FP}$ Out of all actual values how many are correctly predicted.

	1	0
1	TP	FP
0	FN	TN

False Positive → Important ↓↓↓.

④ Recall = $\frac{TP}{TP+FN}$ ⇒ FN ↓↓↓.

Usecase 1 : Spam classification → Spam / Not Spam.

	1	0
1	TP	FP
0	FN	TN

FP ↓↓

Mail → Spam {Truth}
Mail → Spam
} ⇒ Good Scenario.

FN ↓↓

0 ← Mail → Not a Spam
Model → Spam
} ⇒ False positive.

[Precision]

Mail → Spam } ⇒ FN

Model → Not a Spam ∫

Usecase 2 : Predict whether a person has diabetes or not ⇒ FN ↓↓↓.

⇓

Recall

X Bank

↓↓.

Assignment : Tommorrow the stock market is going to crash.

⇓

F - Beta Score.

Harmonic Mean.

$$F1\ Score = 2 * \left[\frac{Precision * Recall}{Precision + Recall}\right].$$

Question1:

In a simple linear regression model, the R-squared (R2) value is 0.70, and the Adjusted R-squared (R2) is 0.68. What does this difference between R2 and Adjusted R2 suggest?

a) The model is underfitting the data.
b) The model is perfectly accurate in making predictions.
c) The independent variable is not relevant to the dependent variable.
d) The model includes unnecessary polynomial terms(features).

Question2:

What is the primary purpose of the Adjusted R-squared (R2) in regression analysis?

a) To quantify the proportion of the variance in the dependent variable explained by the independent variables.

b) To provide a measure of the model's accuracy in predicting the dependent variable.

c) To account for the number of predictors in the model, penalising excessive complexity.

d) To calculate the residual sum of squares (RSS) of the regression model.

THANK - YOU