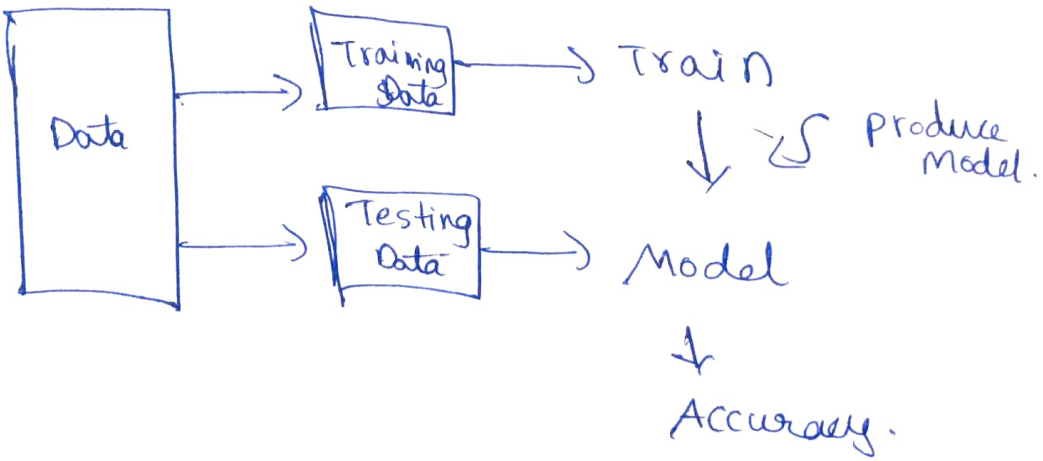
Supervised learning :-

1. In this, A model is getting trained on a labelled dataset.
2. It is a process of providing sample input data as well as correct output data.
3. This learning is to find a mapping function to map the input to the output.
4. In Supervised learning, the main data set is divided into 2 data sets.
   a. Training Data set.
   b. Testing Data Set.



Learning a class from Examples:

1. Set of Cars

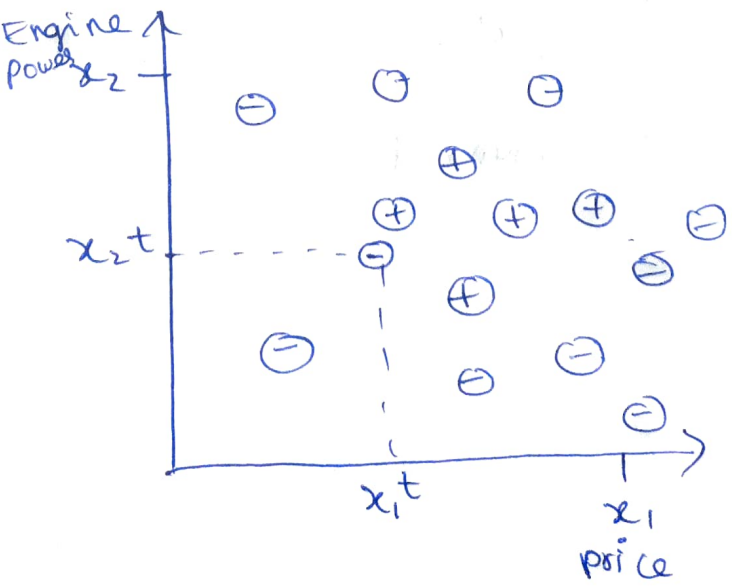
   "Class - C : Family Cars"

2. A group of people look at the cars and label them as family Car(⊕) not by using main 2 attributes
   price
   Enginepower.

3. The Cars that they believe are family cars are (+) positive examples and other cars are (-ve) negative examples.

4. We can ignore other attributes such as seating Capacity and colour and Consider those of irrevelant.

Training Set-Family Car



* The data point corresponds to one sample car.
* Co-ordinates : price and engine power.
* (+) ; positive examples of class (a family car).
* (-) ; negative examples (not a family car).

Variables 'x' and 'r'

1. Price is the 1st attribute x_1 (eg. ; in Rupees)
2. engine power as the second attribute x_2.
3. It can be denote as $\text{Car}\left\{x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right\}$.

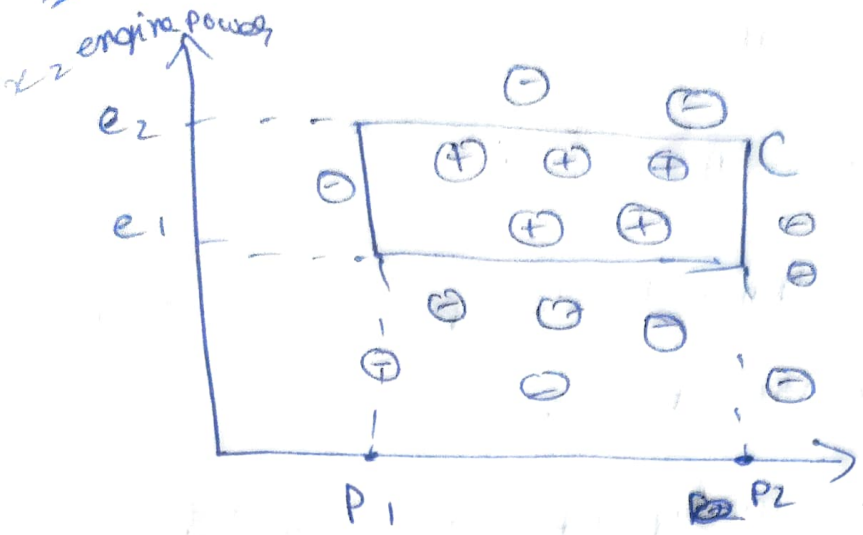
$$r = \begin{cases} 1 & \text{if } x \text{ is a positive example.} \\ 0 & \text{if } x \text{ is a negative example.} \end{cases}$$

4. Each car is represented by such an ordered pair (x, r) and the training set contains N such example.

$$X = \{x^t, r^t\}_{t=1}^{N}$$

5. where t is training set.

Hypothesis class:-



1. If a car to be a family car, its price and engine power should be in certain range.
2. $(p_1 \leq \text{price} \leq p_2)$ and $(e_1 \leq \text{engine power} \leq e_2)$.
3. The class of family car is a rectangle in the price-engine power space.
4. hypothesis, $h \in H$, specified by a particular quadruple of $(p_1^h, p_2^h, e_1^h, e_2^h)$ to approximate C.

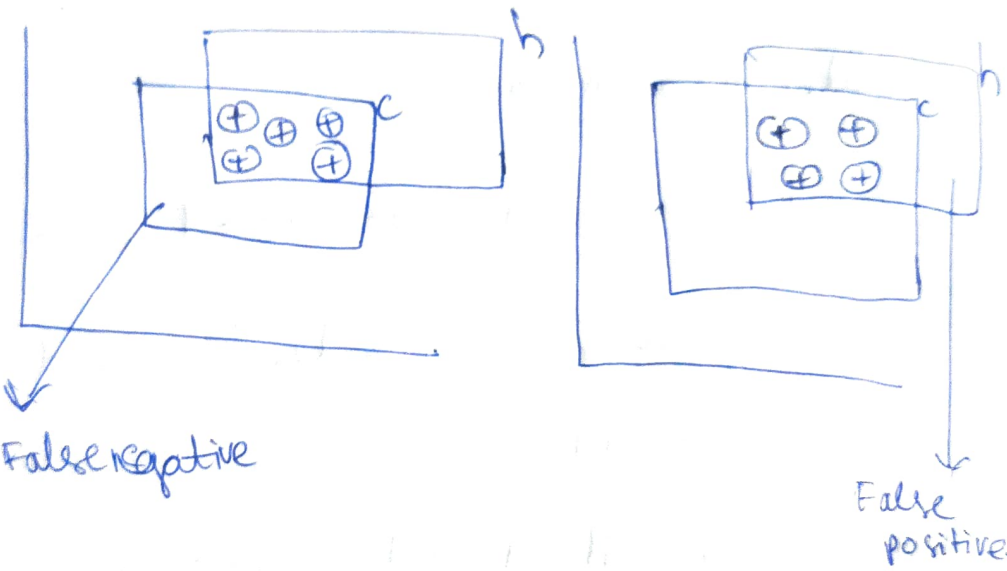
$$h(x) = \begin{cases} 1 & \text{if } h \text{ classifies } x \text{ as a positive example.} \\ 0 & \text{if } h \text{ classifies } x \text{ as a negative example.} \end{cases}$$

1. In real life, we do not know $c(x)$, so we cannot evaluate how well $h(x)$ matches $c(x)$.

2. c - Target function

3. Instances within rectangle represents family cars and outside are not family cars.

4. Hypothesis h - closely approximate C, and there may be error region.



5. The point where c is 1 but h is 0 is False negative.

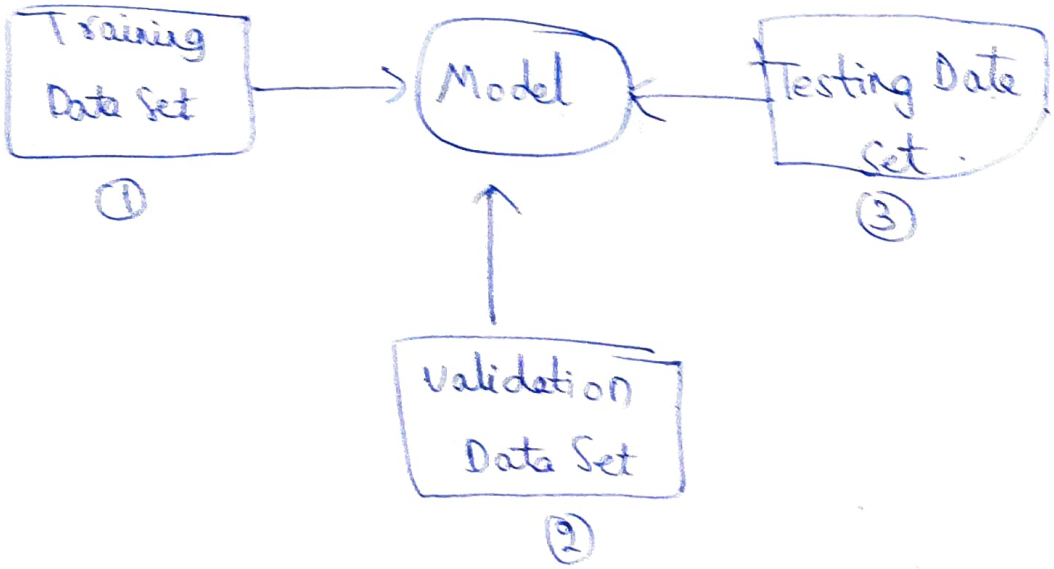
6. The point where c is 0 and h is 1 is called false positive.

7. True positives and True negatives are correctly classified.

Model Selection and generalization :-

There are 3 steps to determine a model with lowest error.

1. (Train) the system with Training Data Set.
2. (Validate) the output of model using validation data set.
   - perform model selection.
3. (Test) the model using testing dataset - Evaluate the model with test data set



In Constructing a Model

1. The set of assumptions, used to learn the algorithm is called the "inductive Bias".

2. We introduce "Inductive Bias is when we assume a hypothesis class "H".

3. In learning the class of Family cars, there are infinitely many ways of seperating the (+ve) examples from the (-ve) examples.

4. Assuming the shape of a rectangle is one inductive Bias, then the rectangle with the largest margin, is another Inductive Bias

5. The class of functions, that can be learned and extended by using a hypothesis class with larger capacity, Containing more complex hypothesis.

6. In Regression, as we increase the order of the polynomial, the capacity and complexity increases

Underfitting:-

1. Underfitting means the training errors and testing errors are more and the system will be simple.
2. Error rate will be very high.

Overfitting:-

1. If we perfectly match the data points in our training dataset, our model probably won't generalize very well, because the data isn't perfect (there always a bit of noise.

Model selection

1. A Model to generate the right output for input instances (the given training set).
2. A model trained on the training set, predicts the right output for new instances is called Generalization.
3. Match the Complexity of the hypothesis class H with the Complexity of the function underlying the data.

4 Triple Trade off :-

In all learning algorithms that are trained from example data, there is a trade-off b/w 3 factors

a The complexity of hypothesis that fit to data i.e, the Capacity of the hypothesis class.

b The amount of training data.

c The generalization error on new examples.

5 amount of training data increases, the generalization error decreases.

6 As the complexity of the model class H increases, the generalization error decreases first and then starts to increase.

Training Set and Validation Set

1 Dividing the data set to 2 parts.

2 one part is for (training), and remaining part is called (validation set) used to test generalization ability

3 If large training and validation sets, then the hypothesis is the most accurate on the validation set. is the (best inductive bias).

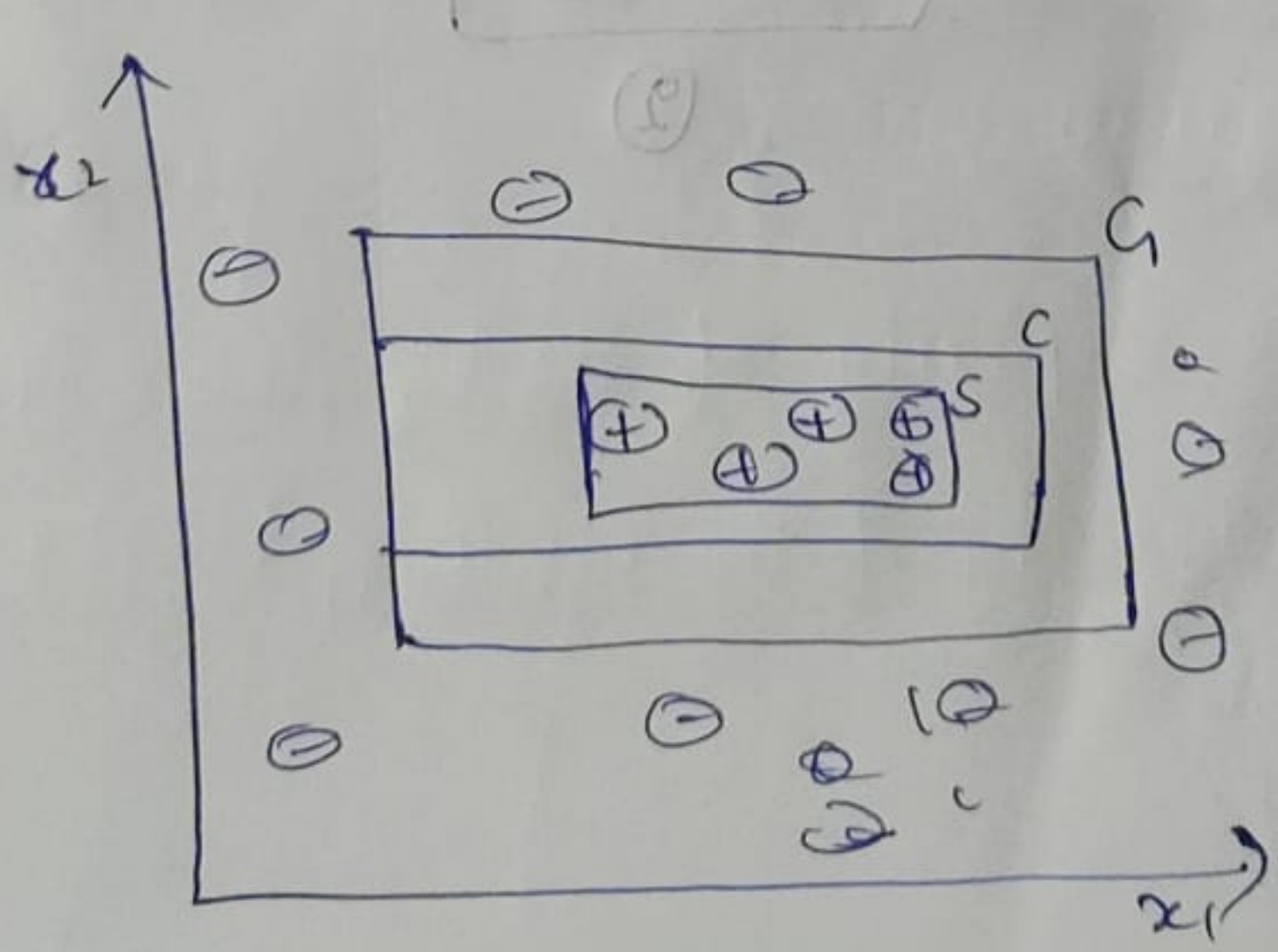
4 This process is called "(cross-validation)"

Generalization:- that is, how well our hypothesis will correctly classify future examples that are not part of the training set.

Most general hypothesis:- G, is the largest rectangle, that includes all the positive examples and none of the negative examples.

Most specific hypothesis:- (S), that is the hypothesis (tighest rectangle) that includes all the positive examples and none of the negative examples.

C is always larger than S..



Regression

- Regression analysis is said to be under supervised learning
- it is a statistical method to model the relationship b/w dependent and independent variables
- The dependent variables are otherwise called as target
  Independent variables are otherwise called as predictors
- This type of regression models help us to understand how the value of the dependent variable is changing corresponding to an independent variable
- the regression model predicts continuous real values
  eg: temperature, age, salary etc.

Types of Regression

1. Linear regression
2. logistic regression
3. polynomial regression
4. support vector regression
5. decision tree

1. Linear regression

- it is a statistical method in predictive analysis
- if there lies only one I/p then such type of Linear regression is called simple Linear regression.
- if there exists more no. of I/p variables then such Linear Regression is called Multiple Linear regression.

* This shows Relationship b/t the independent variable which lie on the X-axis and the dependent variable which lie on the Y-axis

* Simple linear Regression formula

$$y = B_0 + B_1 x + \varepsilon$$

where y = predicted value of dependent variables

B_0, B_1 = are coefficients

x = is an independent variable

ε = Error occured (variation in our estimation)

$$y = mx + C$$

* y is said to the o/p value

x is said to the I/p value

m is slope of the line

c is given Contant

Linear Regression

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

mean square error $\frac{1}{N} \sum_{i=1}^{N} (y - \bar{y})^2$

Eg: Consider the following data → Question

person	Rating Manual (x)	Rating automation (y)
1	4	3
2	2	4
3	3	2
4	5	5
5	1	3
6	3	1

Sol:

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$	$(y - \bar{y})^2$
4	3	1	0	1	0	0
2	4	-1	1	1	-1	1
3	2	0	-1	0	0	1
5	5	2	2	4	4	4
1	3	-2	0	4	0	0
3	1	0	-2	0	0	4

$\sum x = 18$ $\sum y = 18$ $\sum (x - \bar{x})^2 = 10$ $\sum (x - \bar{x})(y - \bar{y}) = 3$ $\sum (y - \bar{y})^2 = 10$

$$\bar{x} = \frac{\sum x}{T} = \frac{18}{6} = 3$$

$$\bar{y} = \frac{\sum y}{T} = \frac{18}{6} = 3$$

Q1. find the value of B_0 & B_1 with respect to ... model which best fits the given data

Sol

$$b_1 = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

$$b_1 = \frac{3}{10} = 0.3$$

Consider $\hat{y} = \bar{y}$

$x = \bar{x}$

$\hat{y} = b_0 + b_1 x$

$3 = b_0 + 0.3(3)$

$b_0 = 3 - 0.9$

$\boxed{b_0 = 2.1}$

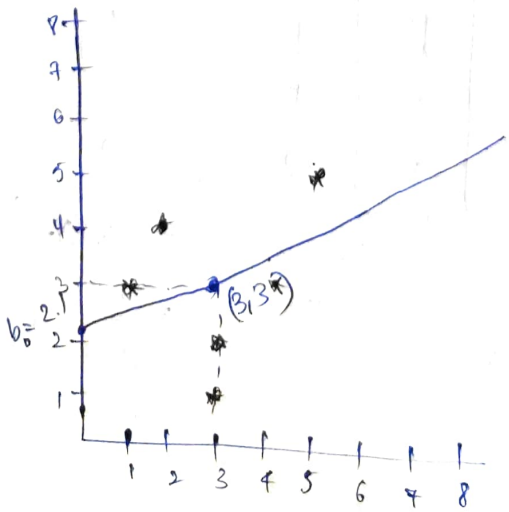
Q2. find the regression line, the best fit for given sample data.

Sol $\hat{y} = b_0 + b_1 x$

$\hat{y} = 2.1 + 0.3 x$

$b_0 = 2.1$

$(\bar{x}, \bar{y}) = (3, 3)$



Q3. Interpret & explain Equation of Regression line

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

$$= \frac{10}{6}$$

$$= 1.6$$

Best fit 0 - 1.0

→ The Error generated for the given sample data exceeds 1

∴ The Regression line is not a best fit for the given of Data

→ The Error can be minimized by Doing the no. of Samples consider

Q4. if new person rates manual car as 4 then predict the rating of same person for automatic cars.

Ans-

$$y^1 = b_0 + b_1 x \qquad x = 4$$

$$= 2.1 + 0.3(4)$$

$$= 2.1 + 1.2$$

$$= 3.3$$