Machine learning

Date 23/08/23

Learning :- Improving with experience(E) at some task (T)

~~Eg:~~

Learn to play checkers.

T → Play checkers

P → % of game won

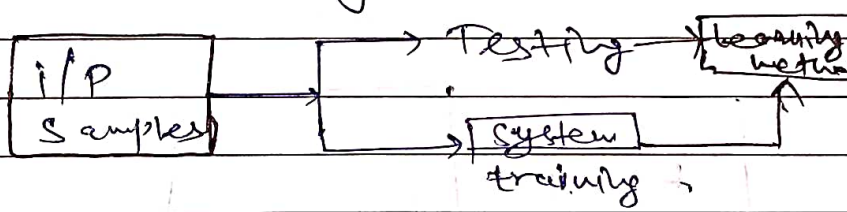
E → Opportunity to play game.
Past record / History.

M.L :- Optimizes a performance criteria using example data or past experience.

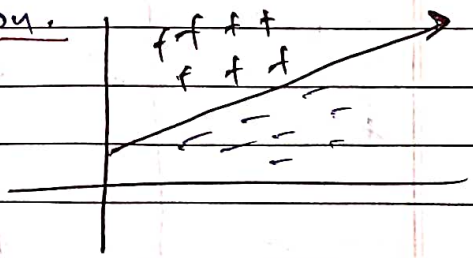
Role of Statistics (Inference from a sample)

Role of C.S (computer Science) efficient algorithm to solve optimization problem representing and evaluating the model from inference

1. framework



→ M.L has three components (i) representation, (ii) evaluation, (iii) optimization.



Classi

Classification Matrix

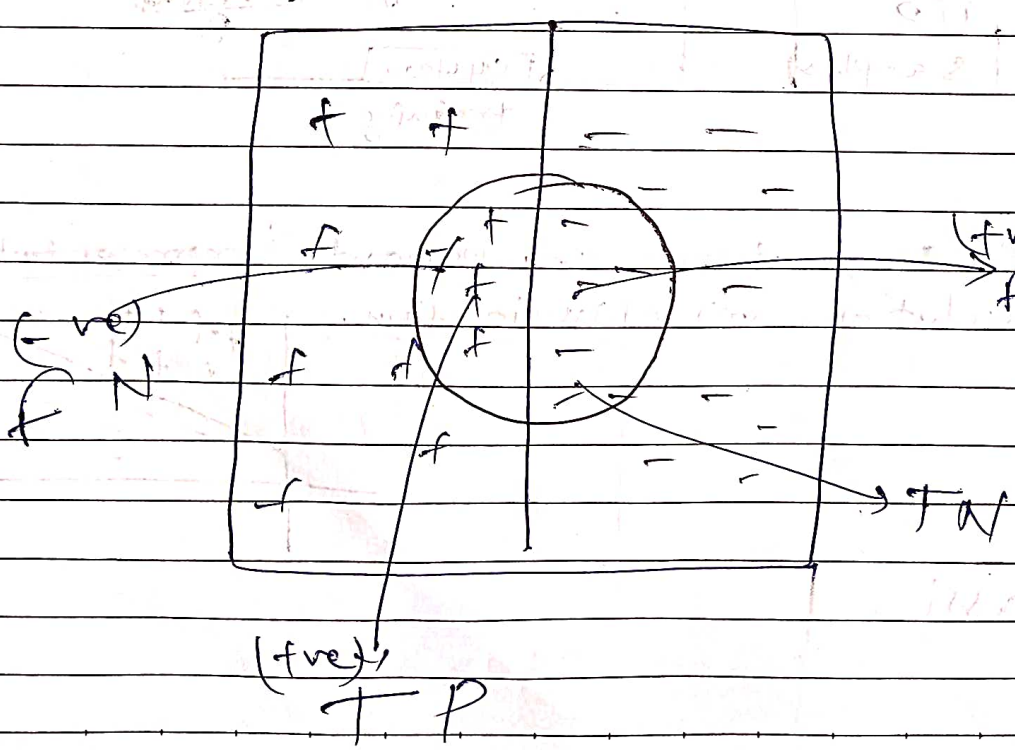
→ Performance measure for ML Classification Algorithm.

True +ve are when you predict an observation belong to class and it actually does belongs to that class.

True -ve → When you predict an observation does not belongs to a class and it actually does not belongs to that class

false +ve: Occurs when you predict observation belong to one class when it really it does not.

false -ve → Occurs when you predict observation doesnot belongs to when infact it does.



Classification measure

Date: Page:

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

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

Precision

How many selected items are relevant.

It is defined as fraction of relevant ex. (TP) among all the examples which are predicted to belong in a certain class.

No. of correctly classified +ve examples divided by total no. of examples that are classified as +ve.

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

Recall

How many relevant items are selected

T.P rate / Sensitivity is defined as fraction of examples that were predicted to belong to a class out of all of the examples are truely belongs to in the class.

Recall is the no. of correctly classified +ve examples divided by total no. of actual +ve examples

Date: _____ Page: _____

M the test ~~said~~ set.

Ex:- Calculating accuracy, Precision, recal
for the following model that classi
100 tumors as malignant (the +ve class
or benign the (-ve class)

True tve (TP)	False tve
Really: malignant	Really: Benign
• ML model predicted: malignant	• ML model pred
• No of TP result = 1	• No. of Pred

TP	FP
Really: malignant	Really - Benign
→ ML model predicted: malignant	→ ML model predicted: mal
→ No of Tp result = 1	→ No. of FP result

FN	TN
→ Really: malignant	→ Really: Benign
ML model predicted: Benign	ML model predicted: Ben
No of FN result = 8	No of TN = 90

Precision $\frac{TP}{TP+FP} = \frac{1}{1+1} = \frac{1}{2} = 0.5$

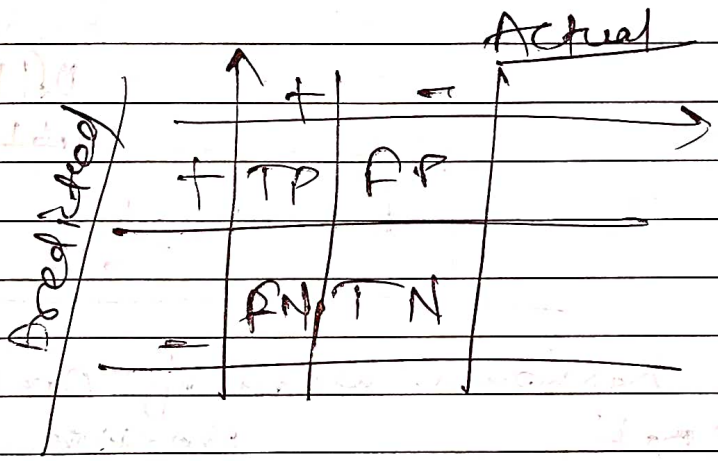
Date: _____ Page: _____

Recall :- $\frac{TP}{TP + FN}$

Confusion matrix

Actual \ Predicted	+	−
+	TP	FN
−	FP	TN

(Positive)



Malignant = 1

Accuracy :- It is defined as the % of correct prediction for test data. It can be calculated by dividing the no. of correct prediction by the no. of total prediction.

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

F_1 value/score/mesure

p - precision
r - recall

$$f_1 = \frac{2pr}{p+r}$$

$p = 0.5$

$r = 0.11$

$$f_1 = \frac{2 \times (0.5)(0.11)}{0.5 + 0.11}$$

$$= \frac{0.11}{0.61}$$

$$f_1 = \frac{11}{61}$$

→ f_1 is harmonic mean of precision & recall.

	Predicted +ve	Predicted -ve
Actual +ve	$T_P = 1$	$F_N = 99$
" -ve	$F_P = 0$	$T_N = 100$

$$p = \frac{T_P}{T_P + F_P} = \frac{1}{1+0} = 1.0$$

$$r = \frac{T_P}{T_P + F_N} = \frac{1}{1+99} = \frac{1}{100} = 0.01$$

$$f_1 = \frac{2 \times 1 \times \frac{1}{100}}{1 + 0.01} = \frac{0.02}{1.01} = \frac{2}{101}$$

(Recall - TPR)

Date: Page:

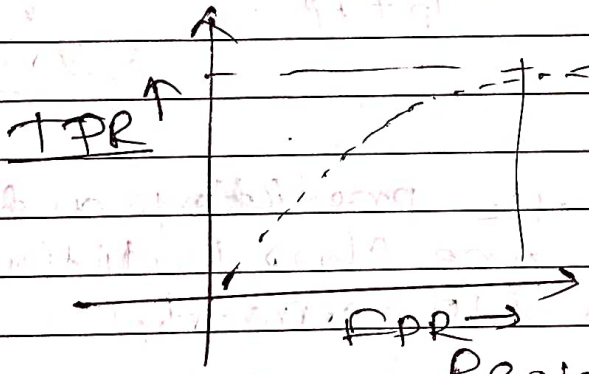
Specificity = $\frac{TN}{TN+FP}$

(False positive rate) = 1 - Specificity

$= 1 - \frac{TN}{TN+FP}$

$= \frac{TN + FP - TN}{TN+FN}$

$FPR = \frac{fp}{TN+fN}$



Receiver operating Characteristics curve
(ROC curve)

It is a graph showing the performance of classification model at all classification threshold. this curve plots two parameter TPR & FPR.

Area under ROC curve (AU. ROC curve

Date: _____ Page: _____

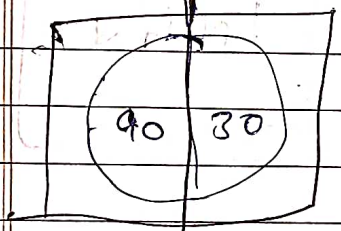
from sklearn.metric import precision_score

Regression Performance metrix

Mean Square esros :

Precision

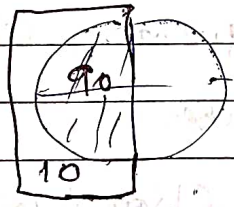
→ A model make prediction and predicts 120 examples as belonging to minority class 90 of which are correct, 30 of which are incorrect



$$P = \frac{TP = 90}{TP + FP} = \frac{90}{90+30} = \frac{90}{120} = 75\%$$

recall

→ A model make prediction and predicts and 90 of the class prediction are correctly and 10 incorrectly



$$\text{recall} = \frac{90}{90+10} = \frac{90}{100} = 90\%$$

A model make prediction and predicts 50 example for first minority class where 50 are correct and 20 are incorrect it predict, 80 for 2nd class with 99 correct and 51 incorrect find precision for it.

$$P_{1st} = \frac{50}{70} \qquad P_{2nd} = \frac{99}{150}$$

Precision = $\frac{80+99}{70+110}$

X_j
$\wedge$
$|^2$

Straintessen matrix Multiplication $\frac{n^3}{n}$

numpy, scipy, matplot lib, Pandas Sklearn

SKlearn
↓
All the library & matrix

n1tk — NLP

Monday, Tuesday

~~Tue~~, Wed, ~~Thur~~
Fri, sat

$P = \frac{Tp}{Tp+fp}$

Suppose
Q). 10000 patient get tested for flus of them 9000 are actually healthy and 1000 are actully seek. for seek people a test were done they got ~~+ve for~~ 620 +ve and 380 -ve. for healthy people same test are done 180 +ve and 8820 get -ve. You have to formulate Confusion matrix and find out precision & Recall

Prediction →

	+	−
Seek	Tp = 620	fp = 380
healthy	FN = 180	8820

Date 26/08/23

(multiple h(x))

Hypothesis Space & Inductive Bias

Inductive learning: It involves using evidence to determine the outcome.

(Learn from examples / past experience / etc.)

→ **Induction**: Deriving the function from given data

→ Specific data to (General model)

→ Given a collection of example $(x^{(i)}, f(x^{(i)}))$, $i = 1, \ldots$

of a $f(x)$, return $h(x)$ that approximates $f(x)$

Where $f(x)$ is called true function which correctly maps the input space 'x' to the output space y & approximating function $h(x)$ is called hypothesis function

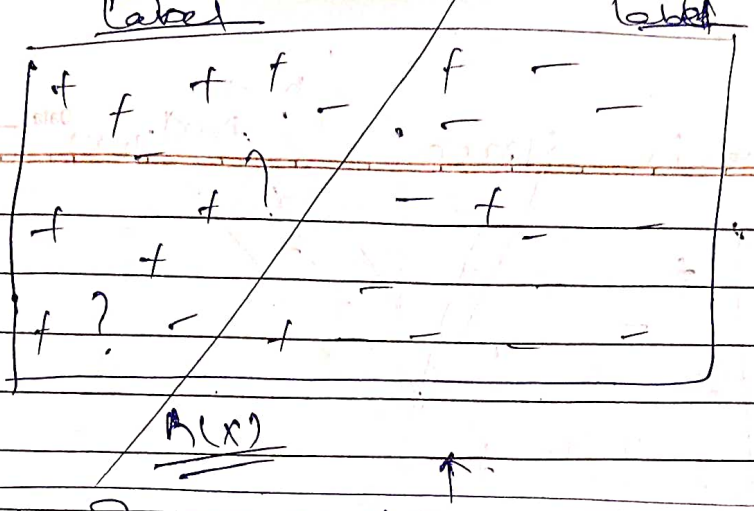
→ Given examples of a function $(x, f(x))$

→ Predict function $f(x')$ for new example (x')

Hypothesis Space-(H) is the set of all legal hypothesis(s) that describe using feature you have chosen $\boxed{h \in H}$

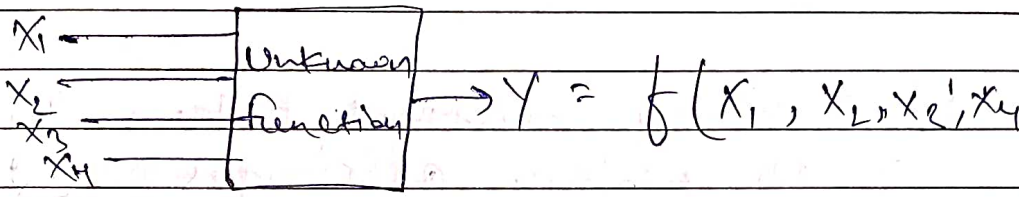
→ Supervised learning machine as a device that explores hypothesis space.

→ Each setting of a parameter in machine is a different hypothesis about the function that maps i/p vectors to o/p vectors.



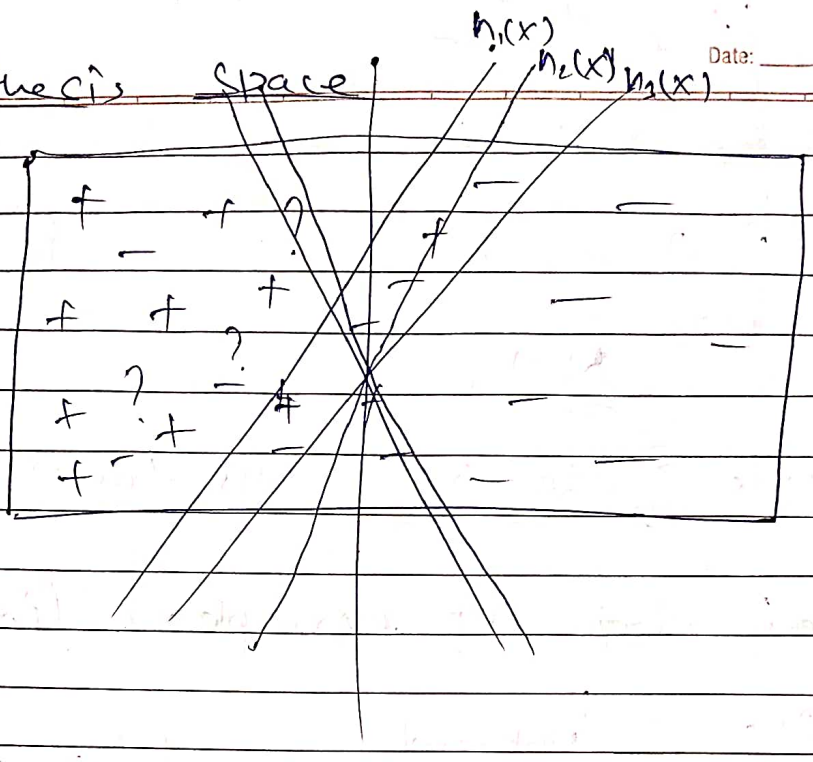
Example of the hypothesis function

Given data or examples, find the function



example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Hypothesis Space



→ How many distinct Boolean functions of n boolean attributes
→ 2^{2^n}

→ How many distinct

→ Inductive Bias :-
→ Need to make assumption & experience alone doesn't allow us to make conclusion & about unseen data instances

→ Inductive Bias : Explicit or implicit assumption(s) about what kind of model is wanted.
→ The constraints on the hypothesis space is called Inductive bias.

01/sept/2023

Bias

AND of OR
(product of sum)

(i) Preferences (ii) restriction

(i) Restriction :- Limit the hypothesis space (ex:- look at rules - only small 'h' belongs H ($h \in H$) are allowed (Language bias) (hypothesis Hypothesis function space)

(ii) → Preference :- Impose ordering on hypothesis space / includes all possible hypothesis.

Search bias

Restriction Bias - Hypothesis space restriction bias (or restrain the language of hypothesis space ex:- k-DNF OR of AND (Shorter hypothesis) sum of product

k-CNF AND of OR
product of sum

The algorithm search through only the subset of possible hypothesis (& incomplete hypothesis space) yet searches these space completely. This type of bais called restriction & language bias. Because no. of possible hypothesis is restricted.

Preference Bias :- It is any order or limit measure matter that serve as best to relation or preference in hypothesis space this algorithm search incompletely through the set of possible hypothesis preferencely select that lead to small decision tree. This type of bias is called preference or search Bias.

Date: Page:

Ex: (Occam's razor) - We prefer a simple formula for a function.

Principal of minimal description length (An extensibility of Occam's razor)

The Best hypothesis is one that minimizes the total length of hypothesis and description of exceptions of this hypothesis.

Occam's Razor

i) Razor

Given two model with same generalization error the simple one is preferred because simplicity desirable itself.

(ii) Razor

Given two model with same training set error the simple one should be preferred because it is likely to have lower generalization error.

Date: _____ Page: _____

Cross Validation

Evaluation method

Hold out set :- The available data set d is divided into two disjoint set.

(i) Training set - d_{tr} {for learning a model}

(ii) Test data set - d_{test} {for testing model}

→ Training set should not be used in testing and the test set should not be use in learning

(Test set - unseen set)

→

Leave One Out Cross Validation

→ This method is used when data set is very small, this is a special case of cross validation. Each fold of the cross validation has only a single test example and all rest data set is training.

d_1	d_2	d_3	d_4	d_5

- Training - $d_1 d_2 d_3 d_4$ | Train: $d_1 d_3 d_4 d_5$ Testing d_5 | Testing - d_2
- Training - $d_1 d_2 d_3 d_5$ | Training - $d_2 d_3 d_4 d_5$
  Testing - d_4 | Testing - d_1
- Training - $d_1 d_2 d_4 d_5$
  Testing - d_3

→ If original dataset has 'm' examples this is 'm' fold cross validation.

→ 'n' fold cross validation :-
(Leaveout)

'C' - fold

Validation set :- The available data is divided into 3 subset

(i) Training set

(ii) Validation set →

(iii) Test set.

↳ It is used frequently for setting parameter in learning algorithm. In such cases the values give the best accuracy on the validation are use as parameters values

Cross validation for estimating as well.

02/Sep/2023

Date: Page:

Overfitting & Underfitting

Signal & Noise :- The underlying pattern that your machine learning aim to learn from data.

Noise :- The irrelivent and random data in data set. If data set is large the ML model with learn the signal is better when data set is small.

Overfitting :- The ML model is very complex In such a case model learn noise in training data perform very well on it However when you use to model to test other data set. the model does not perform well gives high error.

→ This model has huge variance.

Underfitting :- The ML model is very simple The model has very few feature and irregulized way more than needed. Huge bias less variance in its prediction which lead to large error.

Bias :- Error Caused because the model can not represent the concept.

Variance :- Error Caused because the learning algorithm overreacts to small change (noise) in the the training data.

2nd Module

Andrew Ng

Date: Page:

Linear Regression -> (i) Regression Analysis

Regression Analysis → It is statistical technique for the linear relationship.

→ Dependent Variable & Independent Variable

Regression

Regression → It is used in more and more of the Analytical research that affects daily life from how much we pay automobile insurance to what ads appears in social media.

Prediction

→ Linear regression is standard Mathematical technique for predicting numerical outcome. The linear model is also an important model of parametric model.

$$y = c + mx$$

Dependent variable — Independent variable

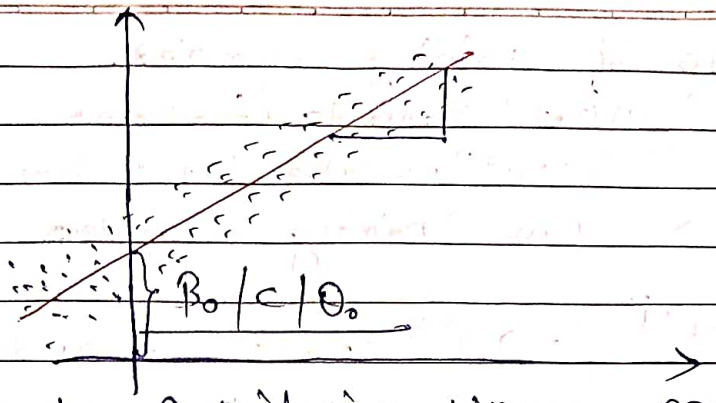
Eg. M.L → $y = \beta_0 + \beta_1 x$

{ β_0, β_1 – Parameter }

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

Estimate the value of θ_0 & θ_1

Date: 5 sept 2023 Page:



We begin by considering linear regression

$h(x, \theta) = \theta_0 + \theta_1 x$, where $\theta = [\theta_0, \theta_1]^T$ are parameters

Weneed to set.

We can measure the prediction loss in term of square error.

Cost function / ~~Cost~~ Loss $(y, \hat{y})$

$= (y - \hat{y})^2$

(y: actual value; $\hat{y}$: predicted value)

x	y
100	100
200	200
300	300

$\theta_0 = 0$	$\theta_0 = 0$
$\theta_1 = 1$,	$\theta_1 = 0.5$

$0 + 1 \times 100 = 100$ $\qquad$ $0 + 0.5 \times 200 = 100$

$(100 - 100)^2 = 0$ Cost function

Performance Measure

Cost function / Loss $= (y - \hat{y})^2$

Mean Square error

Square root mean error.

Date: Page:

n = no. of training examples.

x's = "input" variable / features

y's = "Output" variable / target examples

(x, y) :→ One training example

$(x^{(i)}, y^{(i)})$ → i^{th} training example

Example

$x^{(1)} = 1600$, $y^{(1)} = 350$

$x^{(2)} = 280$ $y^{(2)} = ?$

→ Size of home (Sq feet) → [h] → Estimated price

hypothesis

→ 'h' maps from x's to y's

→ Representation of h

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

values of parameters

$\theta_0 = 0$, $\theta_1 = 0.5$

$\theta_0 = 1.5$, $\theta_1 = 0$

$\theta_0 = 1$, $\theta_1 = 1.5$

(i) Choose θ_0, θ_1 such that $h_\theta(x)$ is close to 'y' for training example (x, y)

(ii) Minimize (θ_0, θ_1) $(h_\theta(x) - y)^2$ predicted $(\hat{y})$

(iii) Minimize $\frac{1}{2} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

$(\theta_0 + \theta_1 x^{(i)})$

Optimize → Minimize
→ Maximize
either

Gradient Descent method

Minimize (θ_0, θ_1) $\frac{1}{2n} \sum_{i=1}^{n} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$

Idea to find approx parameters (θ_0, θ_1) which minimize the cost function.

→ Start with some θ_0, θ_1
→ update the value of θ_0, θ_1 to minimize the cost function
→ Stop till get good result or no change in θ_0, θ_1.

Gradient Descent method.

→ Initilize with random θ_0, θ_1
→ Calculate the gradients of cost function w.r.t parameters $\frac{d}{d\theta_0}$ cost function (θ)

$$= \frac{d}{d\theta_0}\left\{\frac{1}{2}(h_\theta(x) - y)^2\right\}$$

→ $\frac{d}{d\theta_1}$ cost function (θ) =

$$\frac{d}{d\theta_1}\left\{\frac{1}{2}(h_\theta(x) - y)^2\right\}$$

→ update the parameters

$$(\theta_0)_{new} = (\theta_0)_{old} - \alpha \cdot \frac{d}{d\theta_0}\left\{\frac{1}{2}(h_\theta(x) - y)^2\right\}$$

$$(\theta_1)_{new} = (\theta_1)_{old} - \alpha \cdot \frac{d}{d\theta_1}\left\{\frac{1}{2}(h_\theta(x) - y)^2\right\}$$

→ Stop till get good result or no change in θ_0, θ_1

α - Learning Parameter

$0 \leq \alpha < 1$

→ If value of α is too small then gradient descent can be converge too small.

→ If value of α is very large then gradient descent can be passed the minima and can be fail to converge.

$$\frac{d(cf(\theta))}{d\theta_0} = (h_\theta(x) - y)$$

$$\frac{d(cf(\theta))}{d\theta_1} = \frac{d}{d\theta_1}\left\{\frac{1}{2}(h_\theta(x) - y)^2\right\}$$

$$= (h_\theta(x) - y)\frac{d}{d\theta_1}\left(\theta_0 + \theta_1(x) - y\right)$$

$$\frac{d(\text{cost function})}{d\theta_1} = (h_\theta(x) - y)\, x_i$$

Q $x = [1, 2, 3, 4, 5]$

$y = [3, 6, 7, 11, 15]$

$\theta_1 = 2, \quad \theta_0 = 1$

Learning rate (α) = 0.1

$h_\theta(x) = \theta_0 + \theta_1 x$ $\quad (x=1, y=3)$

If 2×1

$y = 12 = h_\theta(x)$, $y = 3$

$$\frac{d\,cf(\theta)}{d\theta_0} = (h_\theta(x) - y)$$

$$= (12 - 3)$$

$$= 9$$

$$\theta_0 = \theta_0 - \alpha_0 \frac{d}{d\theta_0}(h_\theta(x) - y)^2$$

$y = \theta_0 + \theta_1 x$

Q) $x = [1, 2, 3, 4, 5]$

$y = [3, 4, 2, 11, 15]$

$\theta_1 = 2, \quad \theta_0 = 5$

learning rate $\alpha = 0.5$

$$\frac{d\,cF(\theta)}{d\theta_0} = (h_\theta(x) - y)$$

$$= \frac{d\left(\frac{1}{2} h_\theta(x) - y\right)^2}{d\theta_0}$$

$$= (h_\theta(x) - y) \cdot \frac{d}{d\theta_0}(h_\theta(x) - y)$$

$$= (h_\theta(x) - y) \cdot \frac{d}{d\theta_0}(\theta_0 + \theta_1(x) - y)$$

$$\frac{d\,cF(\theta)}{d\theta_0} = (h_\theta(x) - y)(\theta_0)$$

$$\frac{d(cF(\theta)}{d\theta_1} = (h_\theta(x) - y) x_i$$

$$\theta_0 := \theta_0 - \alpha \frac{d}{d\theta_0}(h_\theta(x) - y)^2$$

$$\theta_{0(new)} = \theta_0 - \alpha (h_\theta(x) - y)$$

$$\theta_1 := \theta_1 - \alpha (h_\theta(x) - y) x_i$$

$x=1, \quad y=3$

$y = \theta_0 + \theta_1 x \qquad \theta_0 = 5$

$\theta_1 = 7$

~~$y = h(x) =$~~

$h_\theta(x) = 5 + 7x \cdot 1$

$h_\theta(x) = 12, \quad y = 3$

$x = 2$
$y = 6$

$h_\theta(x) = 5 + 7 \times 2$

$h_\theta(x) = 19, \quad y = 6$

$x = 3$
$y = 7$

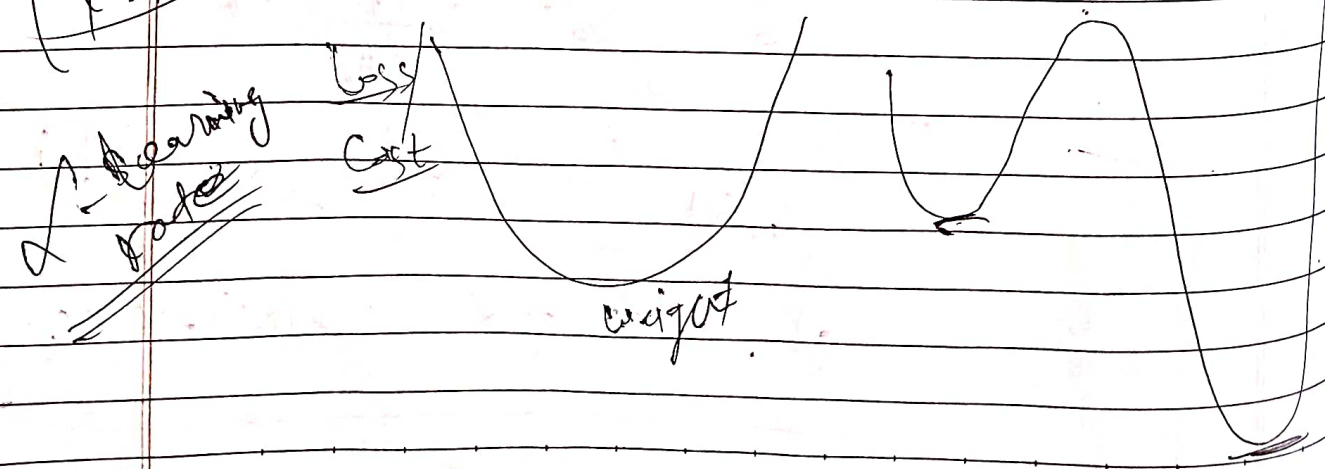
$h_\theta(x) = 26, \quad y = 7$

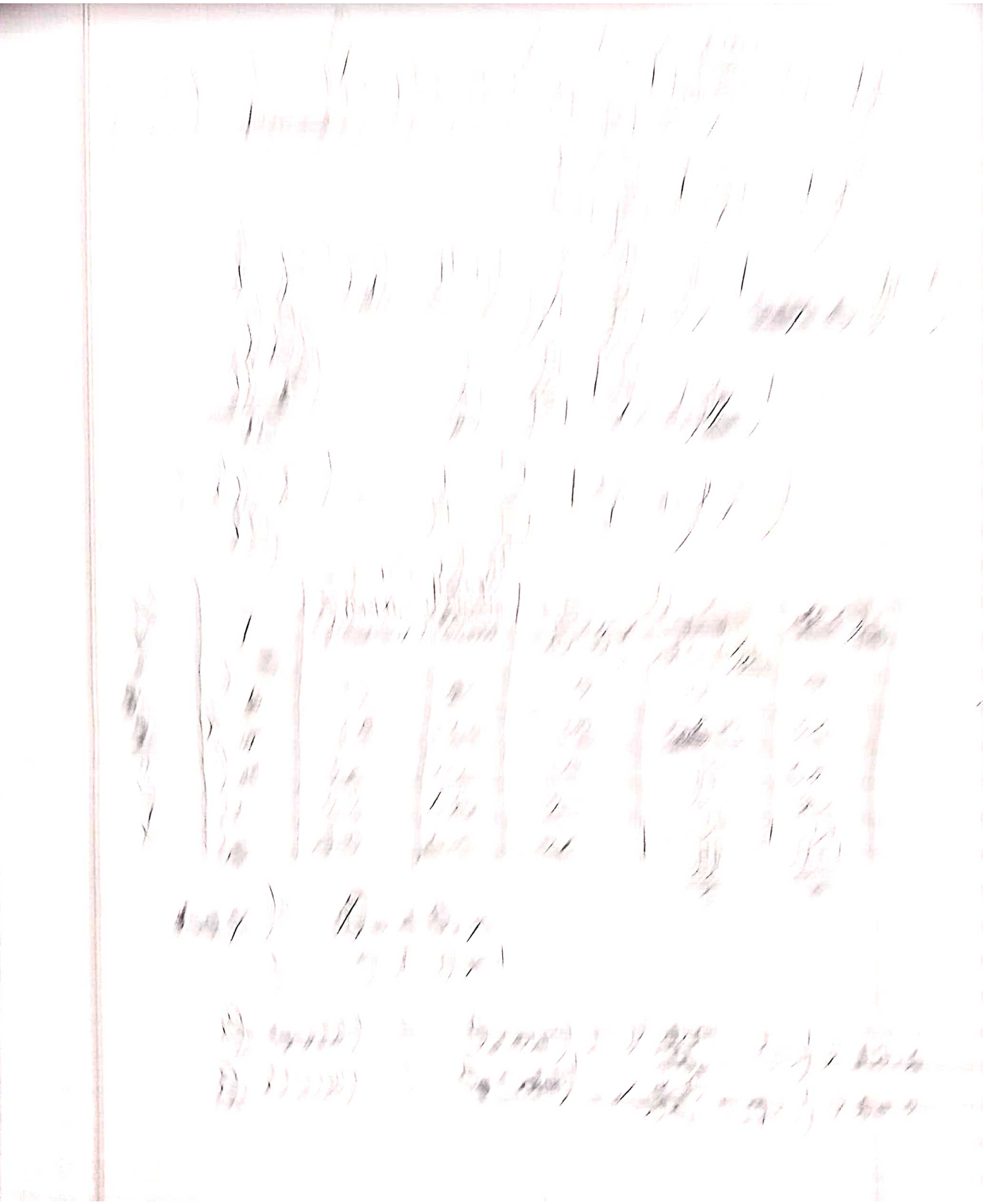
$h_\theta(x) = 33, \quad y = 11$

$x = 4$
$y = 11$

$h_\theta(x) = 40, \quad y = 15$

$x = 2$
$y = 15$





Epoch = 0 04/09/23

x	y	$\hat{y}(h_\theta = \theta_0 + \theta_1 x)$	Cost Square error $(y-\hat{y})^2$	$\frac{d}{d\theta_0}$ Cost	$\frac{d}{d\theta_1}$ (Cost)
1	3	12	81	9	9x1
2	6	19	169	13	13x1
3	7	26	361	19	19x1
4	11	33	484	22	22x1
5	15	40	625	25	25x1
				Sum = 76 / 5	Sum / 5

$$\theta_{0(new)} = \theta_{0(old)} - \alpha \frac{d}{d\theta_0}$$

$$= 5 - \frac{1}{2} \times 15.6$$

$$\theta_{1(new)} = \theta_{1(old)} - \alpha \frac{d}{d\theta_0}$$

$$= 7 - \frac{1}{2} \times 61$$

Epoch = 1

$\theta_{0(new)}$

$\theta_{1(new)}$

Date: Page:

1) Consider the set of points (1,1), (-2,-1) & (3,2), Plot these points & the least squares regression line in the same graph.

x	y
-2	-1
3	2

Matrix Notation

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\text{Cost function} = \frac{1}{n} \sum_{i=1} \left(y_i - (\theta_0 + \theta_1 x) \right)^2$$

$$= \frac{1}{n} \left\| \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \right\|$$

$$= \frac{1}{n} \| (Y - X\theta) \|^2$$

$$\frac{d}{d\theta} \frac{1}{n} \| (Y - X\theta) \|^2 = \frac{d}{d\theta} \frac{1}{n} (Y - X\theta)^T (Y - X\theta)$$

$$\frac{d}{d\theta} \frac{1}{n} \| (Y - X\theta) \|^2 = \frac{d}{d\theta} \frac{1}{n} \left(Y^T Y - 2\theta^T X^T Y + \theta^T X^T \theta X \right)$$

(3) Inverse Method

$$\Rightarrow -2X^T y + 2X^T X\theta = 0$$

equate to 0

$$X^T y = X^T X \theta$$

Normal Equation $\Rightarrow \theta = (X^T X)^{-1} X^T y$

$\Rightarrow$ horse is saddle

17/08/2022

Kaggle (data set) (csv)

reinforcement (Support Vector Machine)

numpy, Pandas, Matplotlib
(read the library file) (plotting graph)

Sklearn, Scipy

Assignment - (i) Load the dataset (ii) graphical (iii)
(i) Shape
(ii) Discription

Date: 19/09/2023

Linear Regression with multiple variable

- X_1, X_2, X_3, X_4 and more.
- Examples of multiple variables (housing data set)

	price	area	bedrooms	bathrooms	stories
d_1	133000	7420	4	2	3
d_2	122500	8960	4	4	4
d_3	1[illegible]25000	[illegible]960	5	3	5

avg(

→ here y = price

$X [x_1, x_2, x_3, x_4]$

x_1 = area
x_2 = bedrooms
x_3 = bathrooms
x_4 = stories

Ridge
Lasso
Elastic

(Lasso - least absolutes shrinkage & selection parameters)

Scatter matrix

→ x^i : input (features) of ith training exam

→ x_j^i : jth features of ith training example.

→ $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$

→ Generalize hypothesis for multiple variables

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_3 + \cdots + \theta_n x_n$

$X = \begin{bmatrix} x^0 \\ x^1 \\ x^2 \\ \vdots \\ x^n \end{bmatrix}$, $\theta = \begin{bmatrix} \theta^0 \\ \theta^1 \\ \theta^2 \\ \vdots \\ \theta^n \end{bmatrix}$

$n \times 1$ $\quad$ $n \times 1$ $\quad$ Date: _____ Page: _____

A

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

$$\begin{bmatrix} \theta^0 & \theta^1 & \theta^2 & \cdots & \theta^n \end{bmatrix} \begin{bmatrix} x^0 \\ x^1 \\ \vdots \\ x^n \end{bmatrix}$$

$$\boxed{h_\theta(x) = \theta^T(x)}$$

Gradient descent Method

~~Learning status~~

~~Regularization~~

Solut

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$

$$\begin{bmatrix} \theta_0 = \theta^1 & \theta^2 & \cdots & \theta^n \end{bmatrix} \begin{bmatrix} x \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$h_\theta(x) = \theta^T(x)$$

Gradient descent Method

Learning rate

Regularization

Solved

~~Regulatio~~ — Shrink (reduce) probability

~~Reato~~

Regularization (Shrinkage in statistics)

λ - Regularization coefficient

λ - control how much fitting the data vs simple hyp

→ Solution to overfitting

→ Reduce number of features

→ manually select features to keep

→ model selection algorithm.

→ Regularization

→ keep ~~All~~ features but reduced magnitude ~~of~~ or value of parameters θ_j.

→ Works well when we have a lot of features

→ Linear Regression with multiple variables

→ Small value for parameter

→ less prone to overfitting

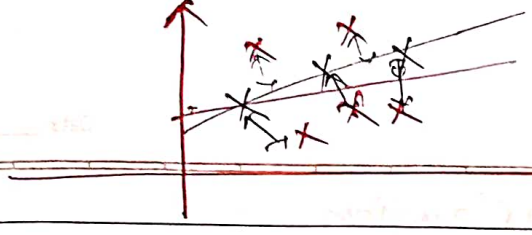
→ Simpler hypothesis.

→ Add Regularization parameter to cost function (θ) to shinks parameters.

→ If λ(lamda) is set to an extremely large value this would result in under fitting

↳ High bias

→ Only penalize theta s from '1', not for '0'



→ for reducing stiffness or generalized model we have to add penality to cost function,

OBJECTIVE

→ Shrink the coefficients or weights of features in the model.

→ Getting rid of high degree polynomial features from the model.

→ The idea of regularization revolves around modifying the loss function/ Cost function $C_0(x)$ in particular, we add a regularization term that pending some specified properties of model parameter.

$$h_\theta(x) = h_\theta(x) + \lambda Reg(\theta)$$

→ There are ~~three~~ two types of regularization

→ Lasso (l_1 norm)
Ridge (l_2 norm) } elastic net (Combination of lasso & ridge).

Date: _____ Page: _____

Lasso Regularization

→ We need to calculate a regularization term that penalize parameter magnitudes for our cost/loss function, we will again use MSE (mean square error)

→ Regularized Cost function/Loss function

$$\text{Cost-function}(\theta)_{\text{lasso}} = \frac{1}{2n}\sum (h_\theta(x^i) - y^i)^2 + \frac{\lambda}{2n}\sum_{j=1}^{m} |\theta_j|$$

(sum of weights) (norm 1)

$$\text{Regularization} = \frac{\lambda}{2}\sum_{j=1}^{m} |\theta_j|$$

$$= \frac{\lambda}{2}\sum_{j=1}^{m} |\theta_j|^2$$

$$= \frac{\lambda}{2}\left|\left|\theta_j\right|\right|_2^2$$

Diagram

MLE

Probability Distribution

Maximum Likelihood Estimation (MLE)

MLE is method that determines values for the parameters of a model. The parameters value are found such that the process described model produced the data that is actually observed.

MLE Principal :

→ Choose parameters that maximize the likelihood function.

Minitab

→ A coin is flipped 100 times. Given that there is 55 heads, find the max^m likelihood estimate for the probability 'p' of heads on a single toss.

→ $n_{C_r}(p)^r \cdot (q)^{n-r}$

Probability of occurance of head :

n, r

$P(55 \text{ heads}) = {}^{100}C_{55}\, p^{55}(1-p)^{45}$

$= \frac{55}{200}$

$55 \cdot {}^{100}C_{55}\, p^{54}(1-p)^{45} - 45 \cdot {}^{100}C_{55}\, p^{55}(1-p)^{44}$

$(1-p)^{44}\, {}^{100}C_{55}\, p^{54}\left((1-p)\cdot 55 - 45 \cdot p\right)$

$(-55p + 55 - 45p)$

$= {}^{100}C_{55}(1-p)^{44}p^{54}\,(-100p + 55) = 0$

$p = 55/100 = 0.55$

Date: Page:

Experiment: flip the coin 100 times & count the no. of heads.

Data: The data is the result of experiment in this case it is 55 heads.

Parameters of interest: We are interested in the value of unknown ~~parameter~~ parameter p.

Likelihood or Likelihood function: This is (data/p

Note: It is a function of both the data & parameter p. In this case the likelihood is p (55 head / p).

$$y = {}^{100}C_{55}\, p^{55}(1-p)^{45}$$

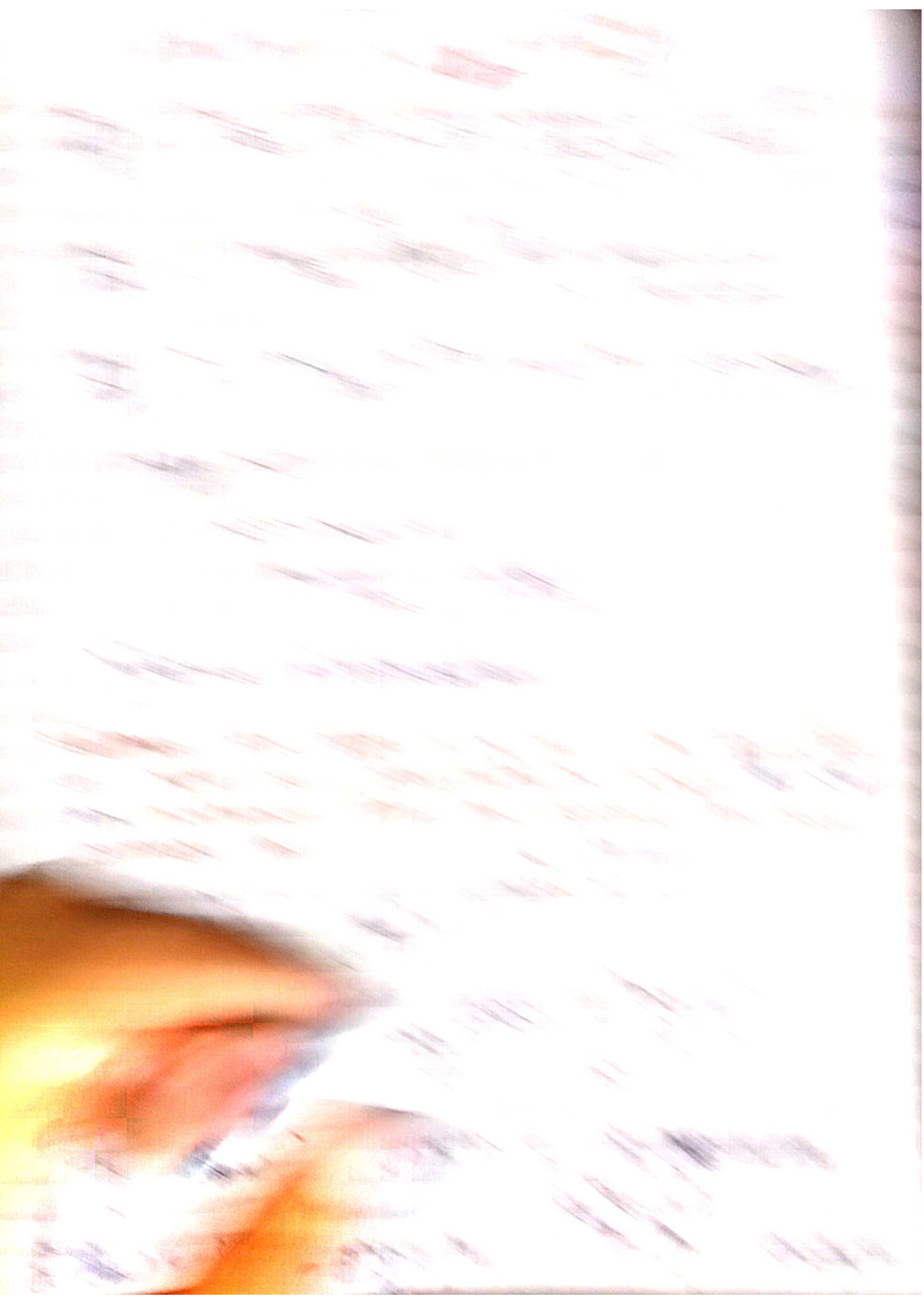
$$\log y = \log {}^{100}C_{55} + \log p^{55} + \log (1-p)^{45}$$

$$\log y = \log {}^{100}C_{55} + 55 \log p + 45 \log (1-p)$$

(differentiate wrt. p)

$$\frac{1}{y}\frac{dy}{dp} = 0 + \frac{55}{p} + \frac{45}{1-p}$$

$$\frac{dy}{dp} = {}^{100}C_{55}\, p^{55}(1-p)^{45}\left\{\frac{55}{p} - \frac{45}{1-p}\right\}$$



Date: _____ Page: _____

$$\frac{dy}{d(\mu)} = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{\sum(x_i-\mu)^2}{2\sigma^2}} \sum \frac{(x_i-\mu)}{\sigma^2} = 0$$

$$\frac{dy}{d(\sigma)} = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{\sum(x_i-\mu)^2}{2\sigma^2}} \left(\sqrt{2\pi}\sigma + \sum \frac{(x_i-\mu)^2}{\sigma^3} \right)$$

$$\boxed{\mu = \bar{x}}$$

$$\boxed{\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$\sigma^4 + \sum \frac{(x_i-\mu)^2}{\sigma^3} = 0$$

$$\sigma^4 = -$$

$x_i = 9, 9.5, 11$

$$\mu = \bar{x} = \frac{9+9.5+11}{3}$$

$$= \frac{29.5}{3}$$

$$\mu = 9.833$$

$f(x_i/\mu,\sigma^2) =$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i)^2 + (\bar{x})^2 - 2(x_i\bar{x})}$$

$$x^2 = \sum x_i$$

$$(n\bar{x})^2$$

$$\boxed{\sigma^2 = \frac{1}{n}\sum(x_i)^2 - n(\bar{x})^2}$$

$$= \frac{1}{3}(81 + 90.25 + 121) - 3 \times (9.83)^2$$

$$9.41 = 290$$

P: $P(A/B) = \frac{P(A \cap B)}{P(B)}$

Marginal Probability :- The probability of an event irrespective of the outcomes of the other random variables eg. P(A).

In the other words, it is known as simple probability which refers to the probability of occurance of a single event ex P(A), P(B)

Joint Probability :- The probability of two (or more) simultenious events eg. P(A and B) or P(A ∩ B)

→ It is often described in terms of event A & B from two dependendent random variables eg. X & Y, the joint probability is often summarized as just outcomes eg:- A & B

Conditional probability :- The probability of two (more) event given the occurance of another

eg: P(A given B) or P(A/B).

Q)

Gender	A_1	A_2	A_3	Total
Male	30	80	90	200
female	20	40	40	100
total	50	120	130	300

(i) find out probability of total no. of female in students

$P(F) = \frac{100}{300} = \frac{1}{3}$

Q) probability that Mit male students with grade A, find out

$$P(M_{A_1}) = \frac{30}{200}$$

$$P(M_{A_1}) = \frac{30}{300} = \frac{1}{10} = 0.1$$

$$P(A_1 \& M)$$

Q) Considered as data if given that student is male then what the probability of A

$$P(M_{A_1}) = \frac{30}{200} = \frac{3}{20} = 0.15$$

$P(A/B)$

$$P(A_1/M) = \frac{P(A_1 \cap M)}{P(M)} = \frac{30/300}{200/300}$$

$$= \frac{30}{200}$$

$$P(A/B) = \frac{P(A \& B)}{P(B)} \Rightarrow P(A/B) \cdot P(B) = P(A \& B) \quad \text{--- (i)}$$

$$P(B/A) = \frac{P(A \& B)}{P(A)} \Rightarrow P(B/A) \cdot P(A) = P(A \& B) \quad \text{--- (ii)}$$

From (i) & (ii)

$$P(A/B) \cdot P(B) = P(B/A) \cdot P(A)$$

Bayes's Theorem

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

Date: _____ Page: _____

$A_1 \cap B$, ~~$A_2 \cap B$~~ $A_2 \cap B$, $A_3 \cap B$

$P(B) = A_1 \cap B + A_2 \cap B + A_3 \cap B$

$P(B) = P(B/A_1) \cdot P(A_1) + P(B/A_2) \cdot P(A_2) + P(B/A_3) \cdot P(A_3)$

$$P(B) = \frac{P(B/A_i) P(A_i)}{\sum_{i=1} P(B/A_i) \cdot P(A_i)}$$

Baye's Theorem → Principled way of calculating a conditional probability without the joint probability the result

$P(A/B)$ is referred to as **posterior probability** and $P(A)$ is refer to prior probability.

$P(B/A)$:- is refer. likelihood.

$P(B)$:- is refer to as evidence.

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{prior}}{\text{evidence}}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B/A) \cdot P(A)}{P(B)}$$

posterior (f)

→ what is P_r that there is fire given that there is a smoke: $P(F/S) = P($

(S)

Date: _____ Page: _____

$A_1 \cap B$, ~~$A_2 \cap B$~~ $A_2 \cap B$, $A_3 \cap B$

$P(B) = A_1 \cap B + A_2 \cap B + A_3 \cap B$

$P(B) = P(B/A_1) \cdot P(A_1) + P(B/A_2) \cdot P(A_2) + P(B/A_3) \cdot P(A_3)$

$$P(B) = \frac{P(B/A_i) P(A_i)}{\sum_{i=1} P(B/A_i) \cdot P(A_i)}$$

Baye's Theorem → Principled way of calculating a conditional probability without the joint probability the result

$P(A/B)$ is referred to as posterior probability and $P(A)$ is refer to prior probability.

$P(B/A)$ → is refer. likelihood.

$P(B)$ → is refer to as evidence.

$$\text{posterior} = \frac{\text{Likelihood} \times \text{prior}}{\text{evidence}}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B/A) \cdot P(A)}{P(B)}$$

posterior (f)

⇒ what is P. that there is fire given that there is a smoke: (S) $P(F/S) =$

P(Fire) — Prior

P(smoke) - evidence

P(smoke/fire) - Likelihood

P(fire/smoke) - Posterior

Q) There are 3 boxes labeled 'A', 'B', 'C'. Box A contains 2 Red & 3 Blue balls Box B contains 3R & 1 Black & Box C contains 1R & 4 Black balls

	Red	Blue
Box-A	2	3
Box-B	3	1
Box-C	1	4

There 3 boxes are identical and have equal probability of getting picked considered as that a red ball is chosen then what is probability that this red ball was picked out of box A.

$P(R/A) = P(A/$

$P(\ /R \quad P(R/$

$P(R$

Date: Page:

$$P(A/R) = \frac{P(R/A) \cdot P(A)}{\sum P(R/A) \cdot P(A)}$$

$$= \frac{2/5 \cdot 1/3}{2/5 \cdot 1/3 + 3/4 \cdot 1/3 + 1/5 \cdot 1/3}$$

$$= \frac{2/15}{2/15 + 3/12 + 1/15}$$

$$= \frac{2/15}{24 + 15 + 12 / 180}$$

$$= \frac{2/15}{81/180}$$

$$= \frac{2 \times 12}{81} = \frac{24}{9}$$

$$\frac{2}{15} \Big/ \frac{81}{180} = \frac{2 \times 180}{81 \times 15} = \frac{4}{27}$$

Name based Classification :

by Classification

PCA, Name