Page No.
Date

DHVT

Unit I

Data

raw or isolated facts from which req info is produced

data science

→ art of uncovering insights & trends in data
→ uses scientific methods, processes, algorithms and systems to extract knowledge & insights from many structural and unstructured data.

problems solved

- classification :- supervised concept, categorizes data into classes
- Regression - method
  - predict continuous outcome (y) from one or more x
- Anomaly detection - obs that differs majorly from rest of data
- Recommendations: offering of relevant suggestions
- Clustering : grouping of data pts having similar attributes
- Reinforcement :-

Industries

- Banking / Finance - fraud detection, customer segmentation, etc
- healthcare - disease detection, genetic analysis, survival analysis
- Retail & E-Commerce sales forecasting, profit prediction etc
- Tele Communications network congestion detection etc.
- Energy & Utilities smart grid security, theft detection, failure models
- Manufacturing - demand forecasting, quality assurance etc.

Page No.
Date

Data Science Lifecycle

→ Business Understanding → ask relevant que
→ clear objectives
→ Data Mining → gather & scrape data
→ Data Cleaning → missing data, duplicate data etc.
missing data :- Nan data
Duplicity → duplicate data → row/column
inconsistent types → numeric as string.
Irrelevant
Outliers
formatting (wrong letter case)
→ Data Exploration
→ Feature Engineering
→ Predictive Modelling
→ Data Visualization

Categories of Data

→ Quantitative
→ no. representation
→ how much
→ how many
→ arithmetic & logical operations
eg exam score

→ discrete : eggs
cannot be represented into decimals, countable
continuous :- only measured
can be represented in decimals
interval & ratio :
numeric
order
difference

→ Qualitative
how & why
categorical

Ordinal
Order data
logical ops
S < M
letter grades
(A, B, C)

Nominal
label data
Marital Status
(Married, Single)

Page No.
Date

Data based on structure

→ Structured
- → has a well defined structure
- → follows a constistent order
- → should be accessed easily

→ Semi-Structured
- → has some structure
- → not in RDBMS
- → lacks rigid schema
- → JSON

→ Unstructured
- no structure
- text/images/videos
- eg document

→ Graph data
- nodes, edges, properties. used to represent & store data
- eg linkedin connections

→ Streaming data
- continuosly generated by diff sources
- req. continuous processing
- youtube, google meet

Page No.
Date

→ Data Collection
process of collecting, measuring & analyzing diff. types of information using a set of standard validated techniques

→ Primary data collection
data collected from first hand sources
best

→ interviews
ask que → respond
high flexibility

→ observations
evaluate behaviour of people in ob controlled & uncontrolled situations
eg. observing random people & walking their pets to open store

→ Surveys & Questionares
→ broad perspective from large groups
→ def.

→ focus groups
similar to interview but in groups

→ Oral History
collecting opinions & personal experiences of people for a event they were involved in

→ Secondary data collection
data that has already been collected by someone else

→ Internet
large pool of free & paid resources
fast & easy
should only source from authentic sites

→ Govt archives
→ authentic & verified
→ not always readily available

→ Libraries
researchers have copies of their research
imp & authentic info
also serve as storehouse for business directories, annual reports etc.

→ Data Preprocessing
data mining technique used to transform raw data into useful & efficient format

→ Data Cleaning + Data Transformation + Data Reduction

→ Data Cleaning
→ data can have irrelevant and missing parts
→ Missing data
→ Ignore the tuples
only suitable on large dataset, multiple values missing
→ Fill the missing values
manually
attr. mean
most probable value
→ Noisy Data
meaningless data that can't be interpreted
due to faulty data collection, data entry errors etc
→ Binning
works on sorted data
divide into segments on equal size - each segment is handled individually
→ Regression
smooth by fitting in regression model
→ Clustering
Similar data in cluster
outliers outside the cluster

→ Data Transformation
transform into appropriate forms
→ Normalization
In order to scale data values in specified range
(-1 - 1) / (0 - 1)
→ Attribute Selection
new attr are constructed from the given set of attr.
→ Discretization
done to replace raw values of numeric attr by interval levels.
→ Concept Heierarchy Generation
attr converted from lower level to higher level in hierarchy. eg. - city → country

→ Data Reduction
- analysis is harder when huge amount of data
aims to inc storage efficiency, reduce data storage & analysis cost
→ Data Cube aggregation
→ Attribute Subset Selection
highly relevant attr should be used, rest discarded
eg: attr having p-value > significance level can be discarded
→ Numerosity Reduction
enables to store data model instead of whole data.
eg. Regression model

→ Dimensionality reduction
reduce size of data by encoding
can be lossy / lossless
→ Wavelet transforms
→ PCA (Principal Component Analysis)

→ Data Discretization

Converting attributes values of continuous data into finite set of intervals with min data loss.

→ supervised discretization - class data is used

- Un " " " - depending how operation proceeds eg: ~~top~~ topdown / bottomup

eg: Age :- 1, 5, 9, 4, 7, 11, 13, 18, 17, 14, 19, 31, 36, 33

↓

Attribute	Age	Age	
	1, 3, 4, 9	11, 14, 17, 13	31, 33

After Discretization - Child Young Mature

→ Histogram analysis

→ Binning

→ Cluster Analysis

dividing the vals of n numbers into clusters to isolate a computational feature of n

→ Decision tree Analysis

topdown slicing is used

select attr with least entropy, run recursion on it.

→ Correlation analysis

discretizing by linear regression

get the best neighbouring interval

then large intervals are combined to develop a larger overlap to form final intervals

21/9/22

DMVT

Unit - 2

→ Descriptive Statistics

Statistics :- concerned with describing, interpretation and analyzing of data.

descriptive / inferential

→ uses analytical methods which provide math to model and predict variation

→ used graphical model to help making numbers visible for communication.

Descriptive → methods of describing ~~data~~ characteristics of data set.

eg : describing weight of product in production line.

→ describing, summarizing, organizing & graphical displays

Outlier - data point i.e. significantly >/< that other data pts in data set.

- may affect calculation of ds.
- detected using graphing the data
- can be normal or may indicate experimental error or incorrect data. (need to decide if we need to exclude them)

Measures used to describe dataset

→ Measures of position

measure central data tendency (where data is centered)

→ Mean

→ works well when distribution is symmetric & no outliers

- $\bar{x}$ (sample)

μ (popn)

→ Median
- half data above it, half below
- reduces effect of outliers
- when data is non-symmetrical
- values should be ordered

→ Mode
- more useful to distinguish b/w unimodal & multimodal distributions (when data has more than one peak)

→ Measures of Spread
→ spread :- how data deviates from position measure
- gives an indication to amount of variation.
→ Range (R)
- highest - lowest values.
- can be misleading when data is skewed / presence of outliers.
- does not make full use of data.

→ Standard Deviation
avg distance of data points from their own mean
- low sd - clustered around mean
- high sd - widely scattered around mean
s (sample) = $\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$
σ (popn)

→ Measures of Shape
- plot
- data will always follow some distribution
- may be symmetrical or non-symmetric

Page No.
Date

eg: uniform, normal, camel-back, bow tied

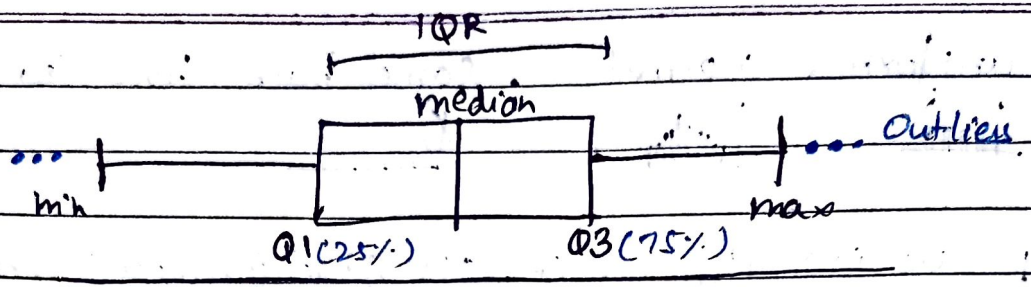
→ helps us identify which descriptive statistics is more useful in a situation
→ symmetrical ? use any (mean - media - ?
skewed ? (median).

→ Skewness
- describes if data is distributed symmetrically around the mean.
- -ve → left skewed
+ve → right skewed

→ Kurtosis
degree of flatness (& or peakness).
if clustered around middle → more peaked
if spread evenly → less more flat

Extra info → IQR (Inter Quartile range)
- divides data into 4 pts. (each - 25%)
- IQR contains middle 50% data
- used when data is not normally distributed.

→ Box Plot
- chart that shows data from a 5-number summary
- used to show skewness / outliers
- used to find spread.



Minimum -
First Quartile (Q1) :- median of lower half
Median - also considered Q2
Third Quartile (Q3): median of upper half
Maximum

+ve skewed :- If max - median > median - min
-ve skewed :- " " < " "

PIVOT Table

df.pivot (index = row, columns = column)

Pivot table is used to summarize and aggregate data inside dataframe.

→ df.pivot_table (index = row, columns = col)
(AVG) /
aggfunc = "sum"

Pivot words ⟷ Pivot table

— summarize large amount of data
— categories.
— drilling down to result from summary
— filtering, sorting etc to focus on info you want.
— provide, concise & printed reports.

* exploring data
* change form layout & field arrangement
* change layout of columns, rows & subtotals
* change display of blanks & errors
* many format

Page No.
Date

Heat Map

- graphically representing numerical data where each data pt. is indicated using colors.
- usual → warm - cool
  (high data pts) (low data pts)
- eg: heat map on website interaction
  red shows high interaction on that portion
  green show low
- Scroll map → based on scrolls
- Click map
- mouse tracking map

- useful for cross examining multivariate data
- good for showing variance across multiple variables, revealing patterns, detecting correlations
- legend is required for it to be succesfully read
- categorical data - color coded
  numerical data - requires colour scale
- can be used to show changes over time

Correlation

- statistical measure that expresses the extent to which two variables are linearly related. common tool for describing simple relationships w/o making a statement of cause & effect.
- r (Sample correlation coeff.) quantifies strength of relationship → -1 to +1
- can't look outside the two variables
- doesn't tell us about cause & effect
- cannot accurately describe curvilinear relationships

p-value - meaningfully conclude that the pop^n correlation coeff is likely

correlation number : r

0 → weak relationship

+ve - both variables tend to be together

-ve - one variables tend to inc while other decrease

p measure of probability used for hypothesis testing

→ ANOVA

→ analysis of variance

→ make multiple comparisons of several pop^n means

→ instead of pairwise, looks simultaneously

- look for 2 variations - variations b/w sample means, variation within each of our sample

→ F-Statistic (variation b/w samples / variation within each sample)

Steps

→ sample means for each of our samples as well as mean for all of the sample data

→ calculate sum of squares of error $\sum (x-\bar{x})^2$: (SSE)

→ calculate sum of squares of treatment

→ square deviation of each sample mean from overall mean.

→ multiply it by no of samples - 1

→ SST

→ calculate degree of freedom

overall : n-1 (no: of data pts. - 1)

treatment : m-1 (no. of samples used - 1)

error : n-m

→ calculate mean square of error (MSE)

$$MSE = \frac{SSE}{n-m}$$

→ calculate mean square of treatment (MST)

$$MST = \frac{SST}{m-1}$$

→ F statistic = $\frac{MST}{MSE}$

→ Linear Regression

→ supervised machine learning algorithm for predictive analysis

→ Linear approach to modelling the relationship b/w a response (dependent) variable & one or more explanatory variable (independent variables)

Simple Linear regression

- single independent variable & corresponding target variable

$$\hat{y} = \beta_0 + \beta_1 x$$

dependent variable, intercept, slope

- multiple Linear regression

2 or more indepent, 1 target variable

can be continuous / categorical

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_i x_i$$

assumptions about data

- homogenity of variance - size of error does not change significantly across values of idv
- independence of observations - idv should not be correlated
- normality - data should follow normal distribution

Applications :- predicting crop yields based on rainfall

marks scored by student based on no of hours study

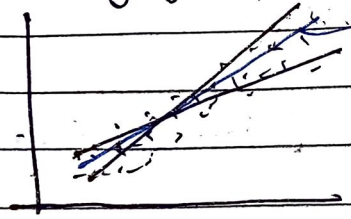
Date

Correlation	Linear Regression
→ specify association b/w two var.	defines numerical relation b/w i/d & dv.
→ not necessary for variables to be different.	variables are different.
→ describes linear relationship	finds the best fit line on data.
→ objective is to find numerical value to express relationship	objective is to find random value based on fixed data

Simple Linear regression

$f(x) = \beta_0 + \beta_1 x$

determination of fit



↳ to find coef of line

↳ use least squares method

RSS - residual sum of squares e = actual − predicted

↳ $e_1^2 + e_2^2 + \cdots e_n^2$

line having min RSS is our best fit for data

multiple linear regression

→ multiple predictors available

$$f(x_1, x_2 \cdots x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

Intercept

Coefficients

Predictors

Page No.
Date

Qualitative predictors

one-hot encoding to convert qualitative data to continuous data. eg: $x_i = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$

eg, for

mlr eg: eg of house prices on House age, total rooms

eg: $\hat{y} = \beta_0 + \beta_1 \times \text{house age} + \beta_2 \times \text{total rooms}$

$\hat{y} = 6989.39 - 72.08 \times \text{House Age} - 0.27 \times \text{total rooms}$

↳ we get a plane

Model Estimation

→ after rejecting null hypothesis, we would like to evaluate model done by R statistic

↓

Root mean squared error

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

measures proportion of variability in observed response variable

Assumptions:-

- linear relationship b/w predictor & target
- no auto correlation in the value of predictors
- constant variance of errors (homoscedasticity)
- residuals must have normal distribution
- little or absence of collinearity among predictors (no ~~multilinearity~~ multi-collinearity)
- presence of outliers & leverage pts can significantly impact.

Advantages of lr

→ easier to implement and interpret the o/p coeff.
→ less complex compared to other algorithms
→ over fitting can be avoided using techniques such as dimensionality reduction etc.

Disadvantages

→ sensitive to outliers
→ assumptions
→ look only on relationship b/w mean of idv & dv.
- mean is not a complete desc of variable.
- lr is " " " of relationships among variables

→ **Model Evaluation using Visualization**

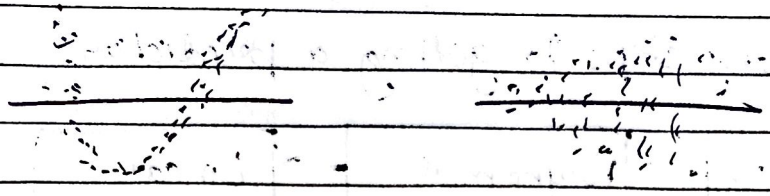
why we regression plot?
gives us good estimate of
→ relationship b/w two variables
→ strength of correlation
- direction of relationship (+/-)

regression plot :- scatterplot + fitted line

→ Residual plot :- measure of how much a regression line vertically misses a data pt
residual values on vertical axis
idv on horizontal axis

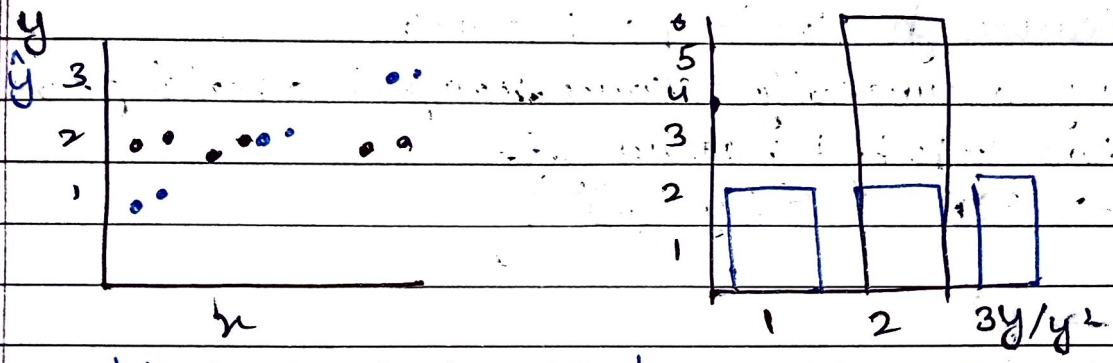
→ randomly spread out around x axis

↳ this shows linear plot is appr.



incorrect models ↑

→ Distribution plot
counts predicted value vs actual value



→ fitted values that result from model vs the actual values.
→ from this we can see mlr follows better than slr.

→ Polynomial regression & pipelines
→ when linear is not best fit
→ transform data into polynomial then use linear
→ describes curvilinear relations
↳ setting higher order terms of predictor variables

→ Quadratic

$$\hat{y} = b_0 + b_1 x_1 + b_2 (x_1)^2$$

Cubic

$$\hat{y} = b_0 + b_1 x_1 + b_2 (x_1)^2 + b_3 (x_1)^3$$

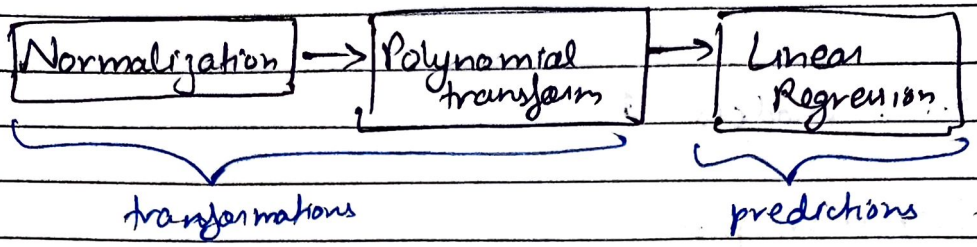
preprocessing
to normalize features when we have multidimensions.

$$b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 + b_4 x_1^2 + \dots$$

Page No.
Date

pipelines

there are many steps to getting a prediction



→ Measures for In-Sample Evaluation

→ way to numerically determine how good a model fits on data

→ two important measures

→ MSE ($\frac{\sum (y_i - \hat{y}_i)^2}{n}$)

→ R^2

→ coeff of determination

→ method to determine how close data is to a fitted regression line

→

$$R^2 = 1 - \frac{\text{MSE of regression line}}{\text{MSE of avg of data}}$$

→ closer to/zero → good/the fit

→ closer to 1 → good fit

→ Prediction and Decision making

how to determine if our model is correct

Look at

→ do predicted values make sense

→ visualization

→ numerical measures for evaluation

→ comparing b/w diff models

Page No.
Date

eg :- Car price based on highway miles per gallon

price = 38123 - 821 * highway-mpg

→ make sense

↳ we might get negative values for price

→ visualization

Simply visualizing with regression might show non linear behaviour.

→ using distribution plots

→ numerical measure

→ MSE might be large

→ R^2 might be closer to 0.

→ comparing

MSE / R^2 values

comparing poly, slr, plr etc.

Page No.
Date

4/12/22 DHVT

Unit-3 - Visualization

→ graphical representation of information and data
→ accesible way to see & understand trends, outliers & patterns in data.

Advantages
→ quickly see trends / outliers
→ easy info sharing
→ dtv → interactively explore opportunities

Disadvantages
→ easy to make inaccurate assumptions when large data.
→ biased or inaccurate
→ corelation doesn't always mean causation

Need? interactive, intuitive, easy to share, personalized

General Types? Chart, table, graph, geospatial, infographic, dashboards
Specific Types? Area Map, Bar Chart, Box Plot, Bullet graph, Gantt Chart, Heat Map, Histogram, Pie Chart

1. Graph Design Principles - for appealing & readable
→ White Space
→ empty space among the elements of graph
- increase readability, focus the reader's attention
macro - spaces around main content
eg - space behind a figure
micro - space b/w digits, axis & text, b/w bars etc.

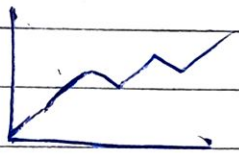
→ Text
- helps ~~reads~~ reader understand the context
- eg :- axis labels, title, annotations
- regarding font size, font family, spacing, paragraph splitting

→ Colour
tool to convey emotions & sensations
characterized by - hue - pure colour, one wavelength
brightness - intensity of colour, amount of white light
Saturation
intensity, amount of shading

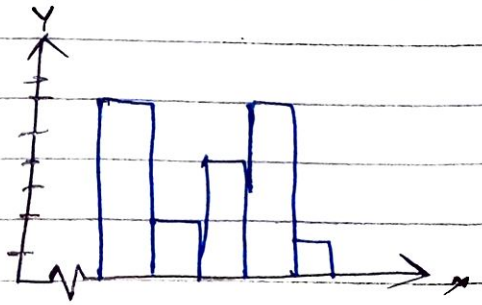
2 Graphs in Statistics

→ Bar Graph
- pictorial representation of grouped data, in vertical / horizontal rectangular bars
- length of bar ∝ measure of data

→ Line Graph
- utilizes points & lines to represent change over time
- line shows relation b/w points



→ Histogram
- displays freq of discrete & continuous data in a dataset using connected rectangular bars.
- no. of obs that fall into interval are represented as rectangular bar

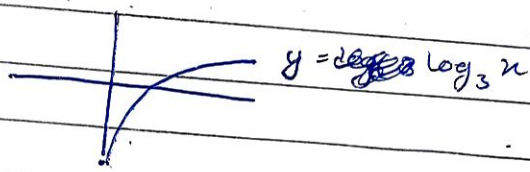


Page No.
Date

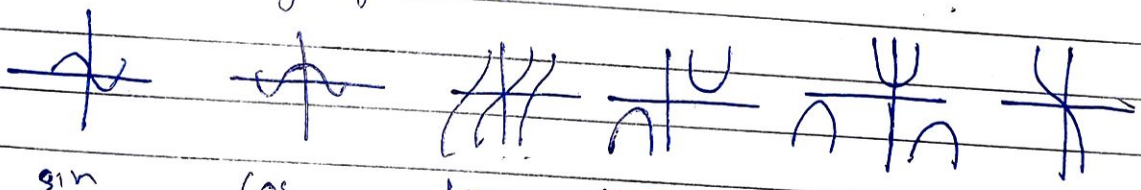
→ Pie Chart
- used to represent the numerical proportions of dataset.
- dividing circle into sectors
  - represents proportion of particular element as whole.

→ Exponential Graph
representation of exponential functions
$y = 3^x$ - increasing
$y = 3^{-x}$ decreasing

→ Logarithmic graphs
inverse of exponential



→ Trignometric graphs
plotted for trigo functions



→ freq distribution graph
used to show freq outcomes in a particular sample

Page No.
Date

3 Interpreting data visualizations
→ identify what information the chart is meant to convey.
→ " info on each axis.
→ " range covered by each axis
→ look for patterns and trends.
→ look for averages / exceptions
→ look for bold / highlited data.
→ read specific data
→ ensure data is credible

Common Visualization mistakes
→ truncated (shortened) Y axis - broken scale
→ misleading cumulative graphs
→ ignoring conventions
- failed calculations
- visualization type is all wrong
- displaying too much data
- ~~by~~ trying too hard to be original
- making reader work too hard
- purposeful bias :- deliberate attempt to influence data
data omissions / adjustments
? - selective bias :- slightly more discreet "

- using percentage change in small sample ($\frac{19}{20}$: 95%)

- correlation implying causation
- decoration
- ignoring pop^n size makes accurate comparisons impossible

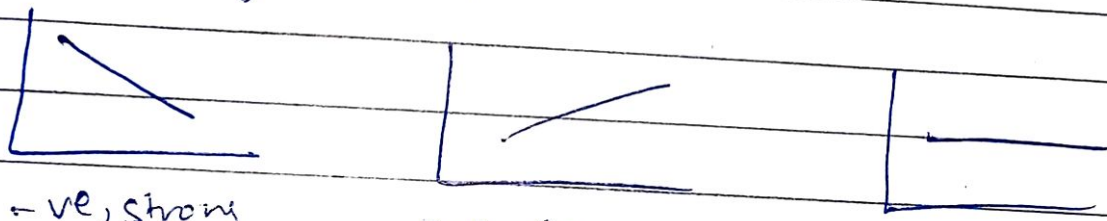
4. Graph Selection Matrix
yog google karlo

5. Correlation, bar & paired graph → Excel
mutual connection b/w two or more sets of data

$$r = \frac{\text{Covariance}}{SD(x) \times SD(y)} = \frac{n(\Sigma x_i y_i) - (\Sigma x_i)(\Sigma y_i)}{\sqrt{(n\Sigma x_i^2 - (\Sigma x_i)^2)(n\Sigma y_i^2 - (\Sigma y_i)^2)}}$$

= CORREL (array 1, array 2)
↓
insert cell range for array (eg A2:A6)

Correlation chart
↳ usually scatter plot



-ve, strong +ve, strong zero

→ select bivariate data x & y in excel sheet
→ Insert → Scatter / Bubble chart
→ add a linear trendline to show correlation

can also change chart types, title, labels etc.

6. Integrity

Graphical integrity refers to how ~~actually accurate~~ accurately visual elements represent data.

e info can vary widely, desire & tendency to
→ scale data disproportionately so that it fits
→ evenly spread values

can result in false impression

6 Principles - Edward Tufte

→ representation of nos should match their true proportions
→ labeling should be clear & detail
→ designs should not vary from some ulterior motive.
→ well known units for money
→ # of dimensions should be same as data
→ should not imply unintended context.

B Visual Integrity

→ what's presented accurately represents what's in the data
→ no design choices distort / obfuscate the inherent facts analytical findings
→ include info on data provenance
cite sources
do not affect credibility
→ clearly define data variables
representation makes sense if accompanying text explains what it involves.
prevent misinterpretation

→ make sure the data being used is complete
eg:- omission of outlier X
manipulating axis thresholds X
leaving out dependent variables that are correlated to ones included in chart. X

→ make sure data is consistent
present in a way that is correctly interpreted by viewers.
eg:- overlaying multiple axes on a single chart & including lines pegged to a different axis. is wrong

→ be consistent on scale too
lie factor :- highlight a finding by scaling the results to make it look more prominent

→ free from visual noise
inconsistent

→ dont filter out data so it cant be viewed
sometimes design of visualization, limits data can be viewed

7 VISUAL PERCEPTION

Ability to interpret the surrounding environment by processing information that is contained in visible light.
Why in visualization? to aid in good decision making drawing insights

→ Visual perception is selective - selectively pay attention, not see everything

→ Our eyes are drawn to familiar patterns.
We see what we expect to see.

→ Our working memory is very limited
we can hold a very limited amount of information in our memory

Visual perception : act of seeing a visual or an image
handled by ~~cortex~~ visual cortex
fast & efficient

Cognition :- act of thinking, processing information, making comparisons etc
cerebral cortex
slower & less efficient

→ data vis. shifts the balance b/w perception & cognition to use our brain's capabilities to its advantage
↳ more use of perception, less of cognition

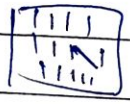
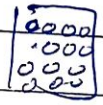
8 Memory
how human memory works? - Iconic, Working, Long term
interact w visualize
memorise/remember

Iconic / Sensory memory
we see a visual, info remains in iconic memory for a tiny period of time.
we process & store info automatically in this time only
↓
preattentive processing
↳ even before paying attention
detects several attributes

Page No.
Date

* preattentive processing
find/count no. of 5's, in
automatically occurs in brain prior to conscious awareness
very fast < 250 ms

~~feat feats~~
features :- contrast to surroundings
eg colors, orientation, size, motion etc

in graphs? lengths, colors etc

working memory / short term memory
we use when we are actually working with the visual info that is of interest to us in processed here.
stays for about a min.
capacity :- 5-6 similar items
capacity can be increased by chunking
grouping similar items

Chunking :- when info is done in form of visuals that show patterns, more info can be chunked together

---

9

Myths about data visualization
→ we visualize because some people are visual learners
→ " " for people who have difficulty understanding numbers.
→ " " to grab attention with eye catching but inevitable less informative
→ best data visualizers are trained in graphic arts
→ provide best means of storytelling contained in data

Page No.
Date

Need of Visual Data

→ seeing the big picture
  ↳ overview
→ easily & rapidly comparing values
→ seeing patterns among values
→ comparing patterns

10 Types of Charts

a) Change over time
show data over a period of time
use cases → stocks
health stats
chronologies
eg:- line, bar, stacked bar, candlestick, area, timeline, horizon, waterfall

b) category comparison
compare data b/w multiple distinct categories
use cases → income across countries
popular venue times
team allocations

eg bar, grouped bar, bubble, parallel coordinates, multi line, bullet

c) Ranking
show item's posn in ordered list
use cases → election results
performance stats
eg. Ordered bar, ordered column, parallel coordinates

d) part-to-whole
how partial elements add to total
use cases : budgets
revenue
eg. stacked bar, pie, donut, stacked area, treemap, sunburst

e) correlation
show correlation
use cases :- income & life expectancy
eg. Scatterplot, bubble, column/line, heatmap

f) distribution
show how often each value occur in a dataset
use cases :- pop^n distribution.
income distribution
eg Histogram, Box plot, Violin, Density

g) Flow
show movement of data b/w multiple states
use cases - fund transfers
vote counts / election results
sankey, gantt, chord, network

h) relationship
how multiple items relate to eachother
use cases : social networks
word charts
eg network, Venn diagram, chord, Sunburst

Page No.
Date

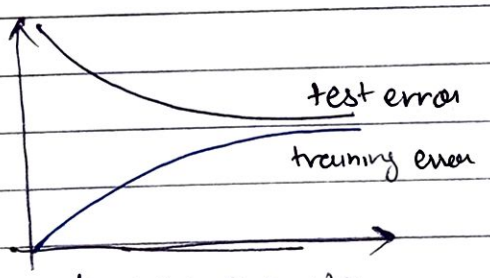
Unit 5

→ Generalization
generalization assesses a model's ability to process new data and generate accurate predictions after being trained on a training set.

Overtraining / undertraining will prevent a model from (overfitting / underfitting) | generalizing

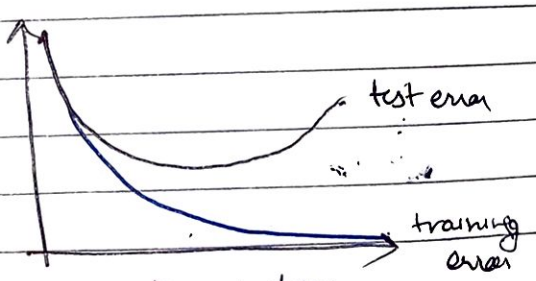
Reasoning
overfitting - model may generalize unintentional / coincidental regularities in training data which are not present in test data.

effects with # of training examples, # of parameters



→ test error decreases as more examples helps in generalization / avoid coincidental regularities

→ training error inc. as small sets are easy to memorize



capacity
→ ~~test error~~ has non monotonic influence, would like to create powerful enough to learn actual regularities but not powerful enough to exploid accidental regularities

Page No.
Date

Measuring

optimising performance of training set is not enough

- divide data :-
  - → training set
    collection of examples on which the network is trained.
  - → validation set
    fine tune by peparameters like no. of hidden units & learning rate.
  - → test set
    evaluate generalization performance

→ Regularization

process of shrinking / regularizing the coeff towards zero.
used to improve performance by imposing penalty in ML.

L1 norm → calculated by adding absolute values of vectors
L2 norm → calculated by taking square root of the sum of the squared vector values.

Lasso Regression :- Coeffs are penalized to the pt where
→ they reach zero in the Least Absolute Shrinkage & Selection Operator Regression
→ uses L1.
→ comes in handy when lot of variables can be used as feature selection

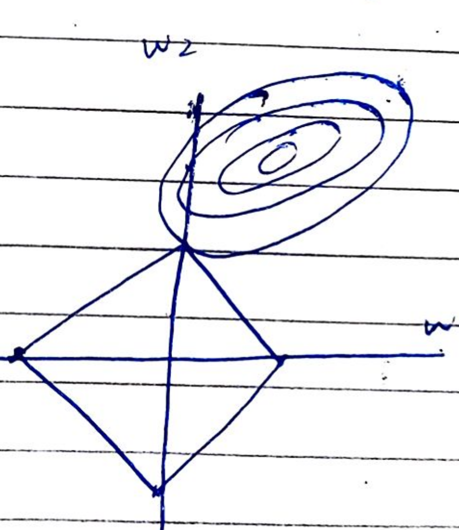
R/d

Regularization with modified loss functions

- LASSO (L1)

$$\text{minimise} \left\{ \underbrace{\sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{ij} \right)^2}_{\text{ordinary least squares}} + \underbrace{\lambda \sum_{j=0}^{p} |w_j|}_{\text{regularization term}} \right\}$$

→ penalizes regressors by shrinking their weight
- regressors that contribute little are more penalized
- λ is weighting factor for regularization to tune overfit ⇔ underfit



→ reduces overfitting
- eliminate insignificant features
- selects only significant regressors

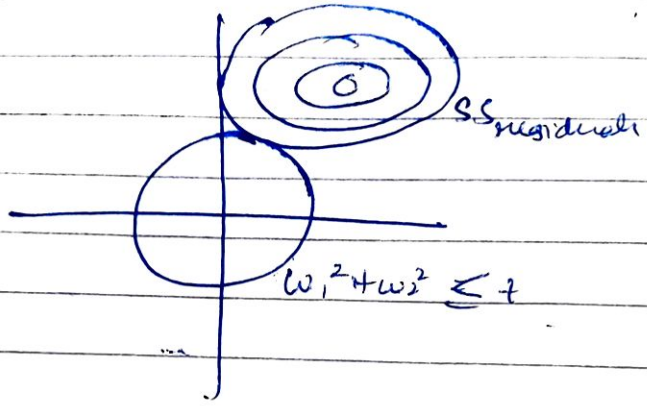
$$(\Sigma)(y_i - (w_1 x_{1i} + w_2 x_{2i}))^2$$

Date

— Ridge (L2)

$$\text{minimize} \left\{ \sum_{i=1}^{M} \left(y_i - \sum_{j=0}^{p} w_j \times x_{i,j} \right)^2 + \lambda \sum_{j=0}^{p} w_j^2 \right\}$$

Regularization term



→ L2 leads to near zero coeffs
— handles multiple correlated regressors better

— Elastic Net Regularization
- combines L1 & L2
each has its own weighting factor

$$\text{Minimize} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} w_j x_{j,i} \right)^2 + \lambda_1 \sum_{j=0}^{m} |w_j| + \lambda_2 \sum_{j=0}^{m} w_j^2 \right\}$$

λ_1 & λ_2 allow
- balance of — attribute elimination
- handling multiple correlated regressors

— proper tuning

Page No.
Date

Ridge Regression

→ when variables in a model are multicollinear
→ minimizes no. of inconsequential independent variables but does not eliminate them.
→ uses L2.
→ L2 penalty = ∝ square of beta coeffs

Choosing right regularization

→ ridge
  → when all independent vars were included
  → collinearity / ~~copodency~~ codependency b/w variables

→ lasso
multiple selectors and we want feature selection

→ Elastic-net
lot of variables

→ Model Evaluation Metrics
In classification problems, we use 2 types of algos (depending on kind of o/p it creates) :-

→ Class output :- algos like SVM / KNN
probability o/p :- logistic regression, random forest, gradient boosting etc.

1. Confusion Matrix

N×N matrix

N is no. of classes being predicted.

generally used only with classification models

Accuracy:-

Predicted values \ Actual Values	+ve	-ve
+ve	TP	FP
-ve	FN	TN

TP:- True +ve - you predicted +ve & its true

TN: true -ve - " " -ve " "

FP: false +ve - Type I Error

you predicted positive and its false

FN:- false -ve :- type 2 Error

you predicted negative but its false

Accuracy:- from all classes, how many were predicted correctly

↳ $\frac{TP+TN}{total}$

Precision (+ve prediction value)

proportion of +ve cases that were correctly identified

↳ $\frac{TP}{TP+FP}$

-ve pv Sensitivity / Recall :- proportion of actual +ve cases which are correctly identified

↳ $\frac{TP}{TP+FN}$

Specificity proportion of actual -ve cases which are correctly identified

↳ $\frac{FN}{FP+FN}$

Page No.
Date

2 F1 Score

when we try to get best precision & recall

harmonic mean of " "

$$F_1 = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

$$F_1 = 2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Why not mean? because HM punishes extreme values more

eg precision = 0 , recall = 1

am = 0.5

hm = 0 ✓

→ if we want to give weight to either one of them, we use

$$F_\beta = (1 + \beta^2) \frac{p \times r}{(\beta^2 \cdot p)(r)}$$

3. Gain and Lift Charts

Concerned to check rank ordering of probabilities

→ calculate probability of each obs
→ rank these probabilities in decreasing order
→ build deciles with each grp having atleast 10% of obs.
→ calculate response rate for each decile - good / bad

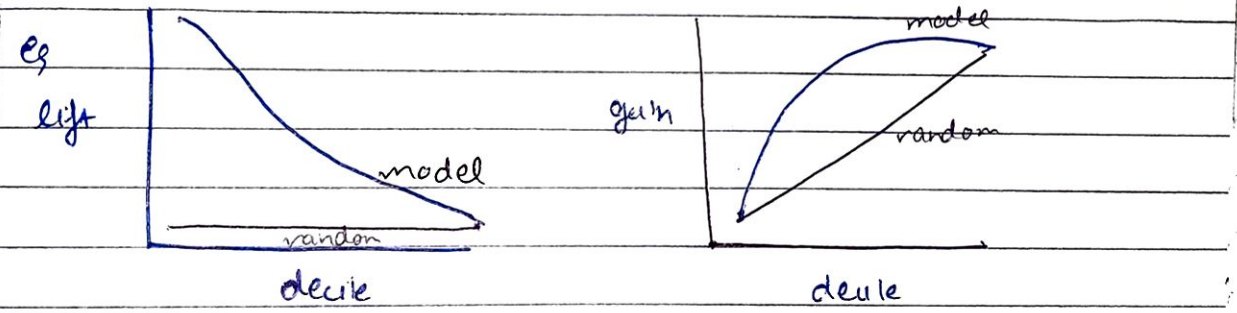
→ to measure how much better our model prediction compared to without the model

- evaluates model performance in a portion of population

- decile - simply putting our data into 10 bins
- we obtain no. of cases (no of data in decile)
  no. of responses (no. of actual true data in decile)

→ Gain - ratio btw the cumulative number of the no. of responses up to each decile divided by total no. of the obs in data.

lift - gain ratio percentage to the random percentage at a decile level
measure how much better we can expect to do with the predictive model compared to without it

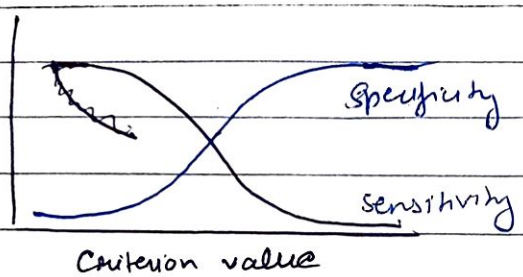


4. Kolomogorov Smirnov chart - KS chart
- measure of degree separation btw true & -ve distributions
- KS is 100, if scores partition into positives & negatives
- if model cannot diff btw true & -ve, it is as if it selects randomly so KS - 0.
- KS lies from 0 - 100

5. Area Under the ROC curve (AUC - ROC)

ROC → Reciever Operating Characteristic

for a probabilistic model, we get a diff value for each metric hence, for each sensitivity, we get a diff specificity.



ROC is the plot b/w sensitivity & 1 - specificity

	TPR	FPR
some thumb rules	↑	↓
0.9 - 1 - excellent (A)	$\frac{TP}{TP+FN}$	$\frac{FP}{FP+TN}$
0.8 - 0.9 - good (B)		
0.7 - 0.8 - fair (C)		
0.6 - 0.7 - poor (D)		
0.5 - 0.6 - fail (F)		

points to remember :-

→ for a model which gives class as o/p will a represented as single pt in ROC plot. such models cannot be compared with each other as judgement needs to be taken on single metric & not multiple metric.

→ In case of probalistic model, we get single no. AUC-ROC but we still need to look at the entire curve to make conclusive decisions

Page No.
Date

Advantages

Why ROC & not lift?

→ lift is dependent on total response rate of popⁿ, hence if that changes, lift curve will also change.

→ ROC is independent of total response rate of popⁿ. this is because its numerator & denominator of both x & y will change on a similar scale in case of response rate shift.

6. Log Loss

AUC ROC does not take into account the model's capability to predict higher probability that are more likely to be true

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1-y_i) \cdot \log(1-p(y_i))$$

$p(y_i)$ is the predicted probability of the class

$1-p(y_i)$ is the " " " -ve "

$y_i = 1$ for +ve class & 0 for -ve class (actual values)

→ negative mean of log of corrected predicted probabilities

→ lower the logloss better the model

→ while the AUC is computed with regards to binary classification with varying threshold, log loss actually takes certainty of classification into account.

7. Gini Coef

used in classification

→ gini coef = 2 × AUC − 1

→ its purpose is to normalize AUC so that a random classifier scores 0, and perfect classifier scores 1

→ range :- [−1, 1]

8. Concordant - Discordant ratio

eg:- 3 students A B C

likelihood of passing

A - 0.9

B - 0.5

C - 0.3

total pairs? AB, BC, AC

after year ends, A & C passed, B failed

choose pairs with one responder, other non responder

AB BC

Concordant pair :- P(responder) > P(non-responder)

non- " " :- P " " < " "

50% concordant ratio

concordant ratio > 60% is good

primarily used to assess model's predictive power & not for how many customers to target

9 RMSE

Root Mean Squared error

most popular metric

assumption → errors are unbiased and follow a normal distribution

$$\sqrt{\frac{\sum_{i=1}^{N} (\text{predicted}_i - \text{actual}_i)^2}{N}}$$

→ "√" empowers this to show large no. of deviations
→ "square" displays more robust results (magnitude of error)
→ avoids use of absolute error values ; highly undesirable
→ in more samples, reconstructing error distribution using RMSE is reliable
→ highly affected by outliers
→ as compared to absolute error, RMSE gives higher weightage & punishes large errors

10. RMSLE

used when we don't want to penalize big difference b/w predicted & actual

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

→ if a & p are small, RMSE ~ RMSLE
→ if one is big, use RMSLE
→ if both are big, use RMSE, RMSLE becomes negligible

11. R-Squared / Adjusted R-Squared

In RMSE we do not have a benchmark to compare

$$R^2 = 1 - \frac{MSE(model)}{MSE(baseline)} = \frac{MSE \text{ of predictions against actual}}{MSE \text{ of mean prediction against actual}}$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(\bar{y}_i - \hat{y}_i)^2}$$

better the model, higher R^2

Page No.
Date

Adjusted R^2

On adding new features, R^2 either increases or remains the same

R^2 does not penalize for adding features that add no value to model

$$\text{adjusted } R^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right]$$

k : no. of features
n : no of samples

if feature isn't valuable, R^2 remains same ~~& 1 denominator~~
so overall we subtract a greater value from 1 so adjusted R^2 decreases

from ppt

R^2

total variance in actual response

$$SS_{total} = \sum (y_i - \bar{y})^2$$

error

$$SS_{residual} = \sum (y_i - \hat{y}_i)^2$$

variance in response given by model

$$SS_{model} = \sum (\hat{y}_i - \bar{y})^2$$

$$R^2 = \frac{\text{explained variance}}{\text{total variance}} = \frac{SS_{reg\,model}}{SS_{total}}$$

$$= 1 - \frac{\text{unexplained var}}{\text{total var}} = \frac{SS_{residual}}{SS_{total}}$$

issues? biased - add regressors, we get better R^2
may add insignificant predictors
may start modelling noise
model overfits

Adjusted R-Squared ($\bar{R}^2$)

Mean Variances

$$MS_{total} = \frac{\sum (y_i - \bar{y})^2}{n-1} \qquad n-1 : df$$

$$MS_{model} = \frac{\sum (\hat{y}_i - \bar{y})^2}{p} \qquad df = p = \text{no of regressors}$$

$$MS_{residual} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}$$

$$df_{total} = df_{model} + df_{error}$$

$$\bar{R}^2 = 1 - \frac{MS_{residual}}{MS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}} \frac{(n-1)}{(n-p-1)}$$

R^2 vs $\bar{R}^2$

→ formula	
→ increases as more regressors are added	increases only if regressor is coeff significant
→ biased estimate	unbiased estimate
→ does not penalize non-significant terms	penalizes
- always +ve	can be -ve
- not suitable for statistical test of significance of wts	suitable for statistical test of significance of wts.

2 CROSS VALIDATION

train model on one subset of data
Validate model on complimentary subset

disadvantage :- we loose a good amount of data from train
the model
can be high bias

overcome

50-50 split

train on first 50, validate on other
train on other 50, validate test on first 50.
reduces bias

2-fold cross validation

Page No.
Date

k-fold Cross-validation

eg. 7-fold validation
→ divide popn into 7 equal samples
→ train on 6, validate on 1
→ next iteration, do validation on diff sample

→ reduces selection bias, reduces variance in prediction power
→ avoids overfitting
→ once we have 7 models, we take avg of error terms to find which model is best

trade off for k.
small k :- high selection bias but low variance in performance
large k :- small " " " high " " "

→ Overfitting & Underfitting in ML

main goal of model is to generalize well
so we need to ~~check~~ check overfitting & underfitting

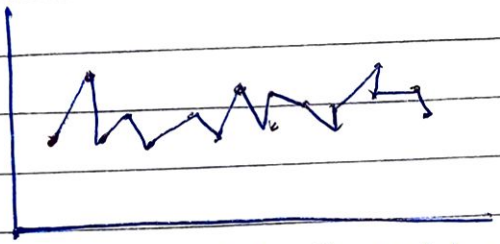
signal :- refers to true underlying pattern of data
noise :- unnecessary and irrelevant data that reduces
bias :- prediction error i.e. introduced in the model | performance (error rate) due to oversimplifying the algorithms.
variance :- perform well with training data, but not with testing data.
↓
diff b/w error rate of training & testing data

Page No.
Date

Overfitting

→ when our model tries to cover all the data pts. / more than required data pts.
→ model starts caching noise & inaccurate values present in dataset, reduces efficiency & accuracy of model.
→ has low bias & high variance

eg:



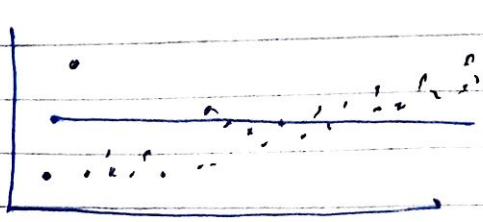
goal of model is to get best fit, but here there is no best fit line

avoid overfitting :- Cross validation
training with more data
removing features
early stopping the training
regularization
ensembling

Underfitting

→ when model is not able to capture underlying trend in data
→ may fail to find best fit

eg



avoid :- increase training time
" no. of features

Page No.
Date

Goodness of fit

defines how closely the result of predicted values match true values

good underfitted mode — good model — overfitted model

(stop at good pt.)

find by → resampling

→ validation dataset