Machine Learning (Jitendra Bedi)

Agenda

Problem Formulation

Text data (its collection, cleaning, getting into data structure)

Naive, Naive bayes, KNN

SVM

Practical machine learning → Regularization technique (RIDGE / LASSO) — To check if model is working well with new data

Bias Variance

SMOTE

Resampling methods

Cross validation

K-fold validation

bagging / boosting — To deal with unbalanced dataset, if we have 100 occurances of nonfraud and 5 occurances of fraud, Our model/Brain would perform better to predict non-frauds rather than frauds.

Captive - Amex, Barclays, HSBC

Analytics / Consulting - Mu-Sigma

Product companies - builds product on a problem, sells

Problem formulation - first check what should be the deliverable for a problem → Imp

Problem 1 - You work in a captive bank, CEO says, I want to launch a product / cards with some ~~dist~~ discounts. He wants to roll out to 5000 initial existing customers. ~~Fede male~~ Male female does not matter, However, it should be targeted to millenials (young generation)

Intent / Target customers

Algorithm to identify top 5000

Transaction history

Earning Potential

Competitive analysis, Demographics

Problem 2 - Karnataka govt is formed. New CM wants to give 5000 iyaval subsidity cards for gaining goodwill

→ In first case deliverable is a sorted list of 5000 customers
→ In second, we can only deliver a model which would mark the person as recieven of LPG card or not
→ 1st is 'sure' problem, 2nd is 'Unsure' problem

Text data Analysis

Process of distilling actionable insights from text

eg. Xiaomi (~~Remdi~~ Redmi) phones. You want to read social media to know the sentiment about xiaomi.

To collect data from twitter, we need to have twitter development account (using your own twitter account) Prof gives 4 min video

install packages bit64, twitteR, ROAuth

API - ~~inte~~ a program to have interface with different applications (handshaking).

```
api_key <- "0md_____"
api_secret <- "scF_____"
access_token <- "949-_____"
access_secret_token <- SntW_____"
```

```
Setup_twitter_oauth(api_key, api_secret, acces_token, access_secret_token)
xiaomi_tweets <- userTimeline("XiaomiIndia", n=2000)
xiaomi_tweets_df <- twListToDF(xiaomi_tweets)
```

dim (xiaomi_tweets_df)
write.csv (xiaomi_tweets_df, file = paste(
"xiaomiIndiaTweets1.csv"))
xiaomi_tweets_df ← read.csv ("XiaomiIndiaTweets1.csv",
StringsAsFactors = FALSE)

↗ It would save all the tweets in above csv. Also it would not consider string as factor. It would consider all tweets same instead of going to levels. Required for unstructured data.

To collect data from RSS feeds

Go to TOI, click on RSS feed, click on cricket, a new ~~tab~~ window would open, copy the URL, paste in R code

install packages RCurl, XML, stringr
xml.url ← "https://timesofindia.indiatimes.com/rssfeeds/4719161.cms"
rssdoc ← xmlparse (getURL (xml.url))
rsstitle ← xpathSApply (rssdoc, '//item/title', xmlvalue)
rssdesc ← xpathSApply (rssdoc, '//item/description', xmlvalue)
rssdate ← xpathSApply (rssdoc, '//item/pubdate', xmlvalue)
rssLink ← xpathSApply (rssdoc, '//item/Link', xmlvalue)
rssdf ← data.frame (rssdate, rsstitle, rssdesc, rsslink)
write.csv (rssdf, "timesofIndia.csv")

You can collect RSS feeds from 100s of websites and store them in a database. It becomes your data and perform data analysis on it. You can even a job which picks latest tweets

You can analyse which website has latest cricket news, rank them (like Trivago)

Text data Analysis

Text mining workflow :-

Problem definition
1) Problem definition & specific Goals

Unorganized state
2) Identify text to be collected
3) Text organization
4) Feature extraction (attributes)
5) Analysis

organized state
6) Reach an insight, recommendation or output

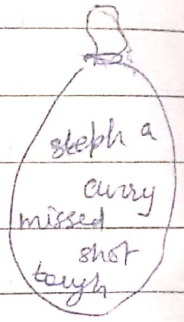
2 ways of Text organization

1) Semantic Parsing
2) Bag of words

sentence: steph Curry missed a tough shot

noun phrase: steph Curry
verb phrase: missed a tough shot

named entity: steph curry
verb: missed
article: a
adjective: tough
noun: shot



Bag of words - collect all words in a bag and try to make sense out of it

N-grams

N=1 This is a sentence — unigram — This, is, a, sentence

N=2 This is a sentence — bigram — This is, is a, a sentence

N=3 This is a sentence — trigram — This is a, is a sentence

Bag of words

Bag of words text mining represents a way to count terms, or n-grams, across a collection of documents (documents means words)

text ← data analytics is the new sensation of the modern times, often misunderstood as only graphs and dashboards. It is time for businesses to think beyond charts and start leveraging the science behind the data.

Manually counting words in this sentence is a pain using R we can do it.

```
library (qdap)
freq_terms ← freq_terms (text, 4)
plot (freq_terms)
```

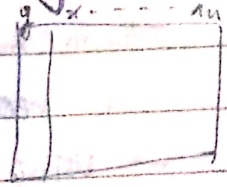
It would find 4 most frequent terms.

It would give a bar chart with count of each word in the sentence.

Using this you can find out which word is used most in the sentence.

Converting raw text to table structure

As raw data is in unstructured form, to pass it into our model we need to convert it into a tabular form so that we can have y and x defined



At this point we would find all x, out of which one can be considered as y in future

Load data (xiaomi twitter example)

```
# import text data
tweets <- read.csv("Xiaomi India Tweets.csv", StringsAsFactors = FALSE)
# View the structure of tweets
str(tweets)
# Print out how many number of rows in tweets
nrow(tweets)
# Isolate text from tweets
xiaomi_tweets <- tweets$text
```

Make the Vector a VCorpus object

datatype

Corpus – Collection of documents, R considers it as

Document – Collection of words. Here each tweet is a document. Every word in the tweet is a term

2 type of corpus –

Permanent 1) P-corpus - requires memory on the system

Volatile 2) V-corpus - virtual, remains in temporary memory

V-corpus is resource friendly, Memory is released as session is closed (Memory efficient)

→ To make volatile corpus, R needs to interpret each element in our vector of text (xiaomi_tweets) as a document.

→ Using TM's package Source function, we will extract (source) our text data contained in a vector. As an output we would get Source object.

library (tm)

plain text document collection of docs

text of tweets only

xiaomi_source ← VectorSource (xiaomi_tweets)

xiaomi_corpus ← VCorpus (xiaomi_source)

→ To create VCorpus from our source object, VCorpus() is used. VCorpus object is a nested list, or list of lists. At each index of the VCorpus object, there is a PlainTextDocument object, which is essentially a list that contains the actual text data (content) and some corresponding metadata (meta)

In R

Make V-corpus using Source

xiaomi_corpus ← VCorpus (xiaomi_source)

Print data on 15th tweet in xiaomi_corpus

~~vector~~ ← xiaomi_corpus [[15]] → extract 15th tweet and show metadata extract 15th document and gives it (tweet) metadata

Print content of the 15th teweet in corpus.

xiaomi_corpus [[15]][1]

[1] is for extracting content of twee and to extract ~~vector~~ content of vector we use [], bracket

→ To clean the corpus, use pre processing functions as part of TM package

tolower(), removePunctuation(), removeNumbers(), stripWhiteSpace(), removeWords() ← removes specific words which are specifically defined.

Pass above VCorpus to tm_map function to get a clean corpus

→ Cleaning with qdap. – extra cleaning functions.

bracketX() – Remove all text within bracket, Its (so) Cool becomes Its Cool

replace_number() – 2 becomes Two (number to word equivalent)

replace_abbreviation() – Sr. becomes Senior

replace_contraction() – shouldn't becomes should not

replace_symbol() – $ becomes dollar (common symbols to word equivalent)

→ Stop words - frequently used but provides little information. I, the, he, she'll. (174 stop words in the common list of TM package.

→ To add more to common stopword list and creating our own.

all-stops ← c("word1","word2", stopwords("en"))

use remove-words() fn to remove all from text

Stopwords("en")

removeWords(xiaomi-tweets, stopwords("en"))

new-stops ← c("xiaomi","redmi", stopwords("en"))

data ← removeWords(xiaomi-tweets, new-stops)

Word stemming and Stem Completion

→ Useful preprocessing step.

→ stemDocument() in TM package to go to words' root.

eg. stemDocument(c("Computational", "computers", "computation"))

→ root = Compu

eg - Jio catches public sentiments using Social Listening Engine. However we need to find/note the root of words used in many different tweets As any +ve person would use similar words to say +ve about Jio

In R - Pre-processing single words

Create Complicate

complicate ← c("Complicated", "Complication", "complicatedly")

Perform word stemming

stem-doc ← stemDocument(complicate)

↓ gives the stem of ~~what~~ all words in complicate vector

Brexit tweets analysed tweets before and after Brexit
Created seperate dictionaries based on age, youngesters, middle aged, Elders. (young, young)

```
# Create the completion dictionary
comp_dict <- "Complicate"
# Perform stem completion
complete_text <- stemCompletion(stem_doc, comp_dict)
# Print complete text
complete_text
```

This helps in finding the frequency, how many times users used complicated sentiment for my brand.

3 kind of search engines/algorithms

1) Direct - matches apple to apple
2) Fuzzy - you give threshold (3 out of 5), 3 matchs in a string, mark this.
3) Context Related - needs historical data. Somebody in past used this, so new one matches with that.

→ Applying Pre-processing steps to a sentence

Here, we could stem easily as it was a vector. But in a sentence, ↑ text_data "In a complicated haste, Tom rushed to fix a new complication, too complicatedly", we would not be able to stem words easily as it is one string (one unit character vector). So we needs to split it into strings and then unlist.

```
rm_punc <- removePunctuations(text_data)
n_char_vec <- unlist(strsplit(rm_punc, split = ' '))
stem_doc <- stemDocument(n_char_vec)
complete_doc <- stemCompletion(stem_doc, comp_dict)
Complete_doc
```

```
In  a  complic  haste  Tom  rush  to
"" ""  "Complicate" ""  "" ""  ""
fix  a  new  complic  too  complic
"" "" "" "complication" "" "complicatedly"
```

→ Applying preprocessing steps to a corpus (collection of documents)

Creating a clean corpus function, where corpus would be passed

```
clean_corpus <- function(corpus){
  corpus <- tm_map(corpus, stripWhitespace)
  " <- "(", removePunctuation)
  " <- "(", content_transformer(tolower))
  " <- "(", content_transformer(replace_abbreviation))
  " <- "(", removeNumbers)
  " <- "(", removeWords, c(stopwords("en"),
                           "xiaomi", "redmi"))
  return(corpus)}
```

content transformer function to be used as 'tolower' func. is a function in base R. To call it in tm_map, we need to use it,

```
# apply customized function to the tweet_corp
clean_corp <- clean_corpus(xiaomi_corpus)
# print out a cleaned up tweet
clean_corp[[227]][1]
# print same tweet in original form (without cleaning)
tweets$text[227]
```

NLP packages/ techniques { Word Semantics
Word to vector

TDM vs DTM

Now we would build data structure using clean corpus

	Tweet 1	Tweet 2	Tweet 3	...	Tweet N
Term 1	0	0	0	0	0
Term 2	1	1	0	0	0
Term 3 (Term 3 has appeared 1 time in tweet 1)	1	0	0	0	0
...	1	0	3	1	1
Term M	0	0	0	1	0

	Term 1	Term 2	Term 3	...	Term M
Tweet 1	0	1	1	0	0
Tweet 2	0	1	0	0	0
Tweet 3	0	0	0	3	0
...	1	0	0	0	1	1
Tweet N	0	0	0	1	0

more frequently used

Term document Matrix (TDM) — Transpose ⟷ — Document Term Matrix (DTM)

(any word) Terms as rows — Tweet/Documents as rows

Tweets/document as column — Terms as columns

Intersection: Tells frequency of terms in a particular tweet — Intersection tells how many Tweets have a particular term

Task Problems — If a tweet has "good" in it, it would be counted once. If a tweet has "not good", it would also be counted as once. To deal with these situations, we might have to use bigram, trigram instead of unigram.

```
xiaomi_tdm ← TermDocumentMatrix(clean_corp)
xiaomi_dtm ← DocumentTermMatrix(clean_corp)
xiaomi_m ← as.matrix(xiaomi_dtm)
dim(xiaomi_m)
xiaomi_m[100:103, 2390:2393]
xiaomi_m[148:150, 2587:2590]
xiaomi_t ← as.matrix(xiaomi_tdm)
```

Sparsity — zeros, pass matrix which has more zeros in it. (0 matrix)

For a text data to be used in ~~text~~ mining, we need to convert data into some matrix.

DTM gives us structure with $x_1, x_2 - - x_n$. However we do not know y. One of these x_i ($x_1 - - x_n$) would be our Y. So we would pass this through Sentiment Analyzer (SA). SA would find out the sentiment from each tweet (in terms of 0/1, +ve/-ve, happy/sad etc) using the frequency of each term in a tweet. This new result (0/1, +ve/-ve) would become our Y column.

Finding Frequent terms with tm

```
# Calculate the rowSums
term_frequency <- rowSums(xiaorui_m)
# Sort term-frequency in descending order
term_frequency <- sort(term_frequency, decreasing = TRUE)
# View the top 10 common words
term_frequency[1:10]
```

fans	eduardodudu	get	amp	now
236	166	131	123	112

```
# Plot bar chart of 10 most common words.
```

Finding frequent terms with qdap

```
library(qdap)
frequency <- freq_terms(tweets$text, top=10, at.least=3,
                        stopwords = "Top200Words")
plot(frequency)
frequency2 <- freq_terms(tweets$text, top=10, at.least=3,
                         stopwords = tm::stopwords("english"))
plot(frequency2)
```

Distance Matrix and Dendogram - Clustering

→ To perform word cluster analysis — use dendograms on TDM (for unsupervised) or DTM (for supervised)

* Using TDM/DTM, call dist() to compute the differences between each row of the matrix, which would help in clustering.
* call hclust() to perform cluster analysis on dissimilarities of distance matrix
* Visualize the word frequency distances using a dendogram and plot().

→ Important pointers

* We can apply above analysis to text. However, we need to limit number of words in TDM using removeSparseTerms from tm
* With large number of words/Terms, it is difficult to analyze. Also many intersections of TDM would have 0 values
* 0 values do not help us in any analysis so we need to adjust the sparsity of TDM/DTM.
* Sparse TDM/DTM means containing mostly zeros.
* Good TDM has between 25 to 70 terms
* Lower the sparse value, more terms are kept
* Sparse value closer to 1, fewer terms are kept
* Sparse value is % cutoff of zeros for each term in the TDM.
* Will use existing tdm to remove SparseTerms

3,

```
dim(xiaomi_tdm)                    - 4401  1561
tdm1 ← removeSparseTerms(xiaomi_tdm, sparse = 0.95)
```

It would remove all those terms from the analysis whose sparcity is greater than the threshold

3,

Output

<< TermDocumentMatrix (terms: 8, document: 1561) >>

Non-/sparse entries : 1506/11482

sparsity : 92%

maximal term length : 13

weighting : term frequency (tf)

As this output gives only 8 terms (guideline is to have b/w 25 to 70 we try with cutoff = '975

It gives us output with terms 38. which we can use for further analysis (38 most useful terms)

Imp Suppose we have a tdm with 1000 tweets and more than 3000 terms, I cant easily interpret a dendogram which is this cluttered So we need to play with sparse parameter to reach to optimal number of terms to be used and to increasing sparsity %age

→ If we create dendograms (cluster tree) using above terms, we would get some meaningful information

→ However, most of the times, dendograms based of textual data does not give very good info

→ Eg. An average value about population tells something, but does not tell everything like, min, max, std dev,

→ Non sensical clusters - which gives no sensible cluster, mostly happens with text data

Sensical clusters - gives valuable information (very rare with text)

→ To get distance using dist(), change TDM/DTM into matrix (as.matrix()) and then to data.frame()

In R

```
tdm2 ← removeSparseTerms(nomi_tdm, sparse = 0.975)
tdm_m ← as.matrix(tdm2)
```

Kind of association in Market Basket Analysis

```
tdm_df <- as.data.frame(tdm_m)
tweets_dist <- dist(tdm_df)
hc <- hclust(tweets_dist) ; plot(hc)
```

To improve the look and feel of dendogram, use dend Extend

```
hcd <- as.dendrogram(hc)
hcd <- branches_attr_by_labels(hcd, c("eduaubdedubu",
    "eduaubdedubua", "merepassgifhai"), "red")
plot(hcd, main = "Better Dendogram")
rect.dendrogram(hcd, K=2, border = "grey50")
```

Using Word Association

Objective is to find out words which are highly associated with a particular word, for ex, Xiaomi - Android, xiaomi - chinese, xiaomi - cheap.

→ We calculate correlations with every word (0 to 1), give a cutoff of correlation ; to find associations

Kind of associations in Market Basket Analysis

```
associations <- findAssocs(xiaomi_tdm, "smart", 0.2)
```

I have taken 0.2 as threshold as it is very difficult to have high association in text data. However, we can try with 0.3 and 0.4

```
associations_df <- list_vect2df(associations)[, 2:3]
ggplot(associations_df, aes(y = associations_df[, 1])) +
  + geom_point(aes(x = associations_df[, 2]), data =
    associations_df, size = 3)
```

text)

Classification Methods

(1) Naive rule
(2) Naive Bayes
(3) K-nearest neighbour

- All are data-driven
- Not model driven
- makes no assumption about data

→ When using linear regression, assumption is linear relation b/w y and x. Atleast 1 β is non zero, only then the model is going to work.

→ With logistic regression, assumption is log(odds) would follow a linear relation.

Naive Rule

- Gives class boundary in order to classify
- Classify records as "Majority class"
- Not a 'real' method, ancient rule, no use of predictors (x)
- Introduced so it will serve as a benchmark to compare other algorithms

Eg:- If I have a dataset of 100 student's results (80 pass, 20 fail, so majority is pass). Now I need to classify the result as p/f for 101 student, I pass that record to naive rule algorithm, it would classify that result as "Pass" (irrespective of other details)

Naive Bayes

- It uses predictor values to classify the record
- To classify Pass/fail for a new record, compare the predictor values of this record with similar values in the dataset
- Assign that class to your new record.
- Requires categorical variables. Numerical variables must be

Imp

Imp If data is categorical, NB is also an option to classify them. Others, Decision Tree, Logistic, SVM, NB, ANN,

Continuous Page
Date / /

binned and converted to categorical
- Can be used with very large datasets
- Eg. Spell check - computer attempts to assign your misspelled word to an established class (i.e. correctly spelled words)

		M	S	SSc	
If new record is P, P, P	1	F	P	P	F
- Using Bayes, classified as 'F' as majority class is F	2	P	P	P	P
	3	F	P	P	F
- Using NB, it would be classified as P. Does not check for exact match P,P,P in one row. Instead, algorithm tries to understand/analysis P for M, S, SSc in all rows.	4	F	P	P	F
	5	P	F	P	F

Exact Bayes classifier
- Relies on finding other records that share same predictor values as records to be classified.
- Drawback - even with large datasets, it may be hard to find other records that exactly match with new record.
- To overcome above problem, we use 'Naive Bayes'
  - NB assumes independence of predictor values (within each class). However, usually variables are correlated to each other
  - Uses multiplication rule
  - Finds some probability that records belong to class 'C' given predictor values, without limiting the calculation to records that share all these same values.
  - If we already have record P, P, F and the new record is F, P, P, NB would put it in same class as it would check, out of 3, 2 are P, 1 as F. order does not matter
  - Works well with large datasets (specifically categorical)

Drawbacks :-

- Requires large number of records
- Problematic when a predictor category is not present in training data
  as it assigns 0 probability of response ignoring ~~informing~~ information in other variables.

R code for NB using Stock market data

```
library(ISLR)
attach(Smarket)
train = (Year < 2005)
Smarket.2005 = Smarket[!train,]
Direction.2005 = Direction[!train]
library(e1071)
snb = naiveBayes(Direction ~ Lag1 + Lag2, Smarket,
                 subset = train)
snb.pred = predict(snb, market.2005)
table(snb.pred, Direction.2005)
mean(snb.pred == Direction.2005)
```

	Direction.2005	
Snb.pred	Down	Up
Down	28	20
up	83	121

Another option is QDA — Quadratic data Analysis

You can compare NB and QDA

used when relation b/w y and x is not linear

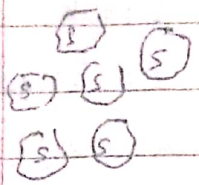
it is quadratic in nature.

K-Nearest Neighbors (classifier)

Being used in testing driver less cars. Good for classifying images

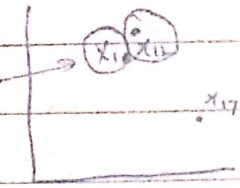
Red colour

Driver less car would first look at the sign and colour of signboard, instead of checking the Text



If multiple signs of different red colour shades

- for a record to be classified, identify nearby records
- Nearby means records with similar predictor values $x_1 - x_n$
- classify the record as whatever the predominant class is among the nearby records.
- Main point is two points are nearest, they have similar kind of characteristics, so they can be classified in same category.
- In KNN, we find out 4, 5 ... n nearest neighbor based on our problem, then try to find out best possible class.



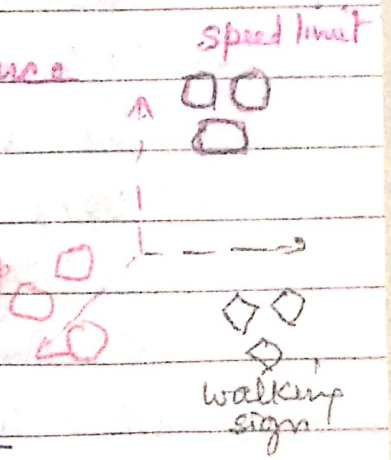
Measuring Similarity with distance

for eg. distances are measured based on RGB values, Based on which similar signs are classified together

To measure nearby, we use

euclidean distance

$$\sqrt{(x_1-u_1)^2 + (x_2-u_2)^2 + \dots (x_n-u_n)^2}$$



I have to classify, whether new x is Red or Black with K=5, I would check 5 nearest neighbours, 3 are black, 2 are red, so I have high confidence that my new x is Black.

Recommendation - Keep K odd to handle scenarios (2 Red, 2 Black, not sure which one to assign), start with K=1, then 3, 5 ... and find out best results

Bigger K is not always better. In some cases small K works better.

Typically we choose K which has lowest error in validation data. K can be from 1 to n (n being all the records), However K=n becomes naive rule

One recommendation = $K = \sqrt{n}$

Low K vs High K

- Low values of K (1, 2, 3 ..) capture local structure in data (but also captures noise)
- High values of K provide more smoothing, less noise, but may miss local structure (local patterns)
- Using KNN for numerical prediction - Instead of majority vote determining the class, use average of response values. May be a weighted average where weight decreases with distance (near records have more weight, far record has less weight)

Imp Recommendation - Before using KNN, normalize the data (so that everything is on same scale)

define a min-max normalize function

```
normalize <- function(x) {
        return((x - min(x))/(max(x) - min(x)))
        }
```

Advantages

- Simple
- No assumptions required about normal distribution
- Effective at capturing complex interactions among variables without having to define a statistical model

Shortcoming

- Required size of training set increases exponentially with # of predictors, p (Because expected distance to nearest neighbor increases with p, as with large vector of predictors, all records end up "far away" from each other)
- In a large training set, it takes long time to find distances to all neighbors and then identify nearest one.
- This constitute course of dimentionality (Use PCA/FA to reduce variables to fewer dimensions)

In R

```
library(ISLR) ; attach(Smarket)
train = (Year < 2005)
Smarket.2005 = Smarket[!train, ]
Direction.2005 = Direction[!train]
# K nearest neighbor
library(class)
train.X = cbind(Lag1, Lag2)[train, ]
test.X = cbind(Lag1, Lag2)[!train, ]
train.Direction = Direction[train]
```

Set seed(

Try with different values of K

knn.pred = Knn(train.X, test.X, train.Direction, k=1)

table(knn.pred, Direction.2005)

mean(knn.pred == Direction.2005) ← Gives accuracy

error is 1 − accuracy

For IRIS dataset

library(caret)

data(iris)

In caret library Knn is used as Knn3

fit ← Knn3(Species ~., data = iris, K=5)

predictions ← predict(fit, iris[,1:4], type = "class")

table(predictions, iris$Species)

SA out for Restaurant Reviews

Sentiment analysis is already done on a restaurant's reviews (Good (1) Liked or bad (0) Not Liked). Information collected from zomato.

We are going to process this data apply text clean up, preprocess, ~~mining~~ mining techniques and then apply classifier techniques.

Using tsv, tab seperated file to tackle problems with ',' in data. (csv would break the comment to new comment as soon as it encounters a tab)

Natural Language Processing

dataset = read.delim("Restaurant_Reviews.tsv", quote = '', stringAsFactors = "FALSE")

library(tm) ; Library(SnowballC) ;

corpus = VCorpus(VectorSource(dataset$Reviews))

corpus = tm_map(corpus, content_transformer(tolower))

corpus = tm_map(corpus, removeNumbers)

⑤ - Bernoulli : need to study this in SA, some pictures

Comlin Page
Date / /

corpus = tm_map(corpus, removePunctuation)
corpus = tm_map(corpus, removeWords, stopwords())
corpus = tm_map(corpus, stemDocument)
corpus = tm_map(corpus, stripWhiteSpace)

here we are using default stopwords and default dictionary

# Creating a bag of words model
dtm = DocumentTermMatrix(corpus)
dtm = removeSparseTerms(dtm, 0.999)
dataset_n = as.data.frame(as.matrix(dtm))
dataset_n$Liked = dataset$Liked

encoding target feature as factor
dataset_n$Liked = factor(dataset_n$Liked, levels = c(0,1))

# splitting dataset into training and test set
library(caTools)
set.seed(123)
split = sample.split(dataset_n$Liked, SplitRatio = 0.8) (80/20)
training_set = subset(dataset_n, split == TRUE)
test_set = subset(dataset_n, split == FALSE)

# fitting Random forest classification to the training set
library(randomForest)
classifier = randomForest(x = training_set[-692],
y = training_set$Liked, ntree = 10)

# Predicting the test set results
y_pred = predict(classifier, newdata = test_set[-692])

# Making confusion matrix
cm = table(test_set[,692], y_pred)
label = training_set[,692, y_pred]

we can try with logical regression here as output is 0 and 1 and find the accuracy.

SVM - Support Vector Machine

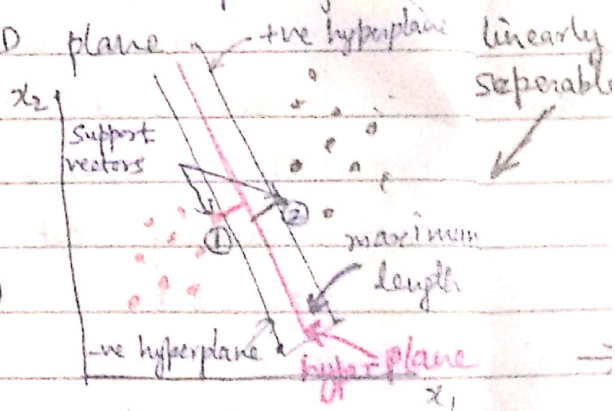
It is a classifier

Here we see that there are two seperate groups, red and black, but how do we put a classification boundary for two groups.

There could be many ways to put that boundary as shown. Now how to confirm which lines would seperate the two colour groups the best.

SVM seperates using maximum margin concept

Red and black points are not just the data points. Infact each point is a vector which represents values of multiple parameters $x_1, x_2 - - x_n$ for that record, $C(x, x_2 - x_n)$. We are plotting each vector as a point on 2-D plane



SVM would find out a plane/ partition which gives maximum distance (margin) from vector 1 and vector 2.

distance of plane from vector 1 and vector 2 should be equidistant. Vector 1 and 2 are called support vectors, which are nearest to the plane (at equidistance) and give up maximum margin.

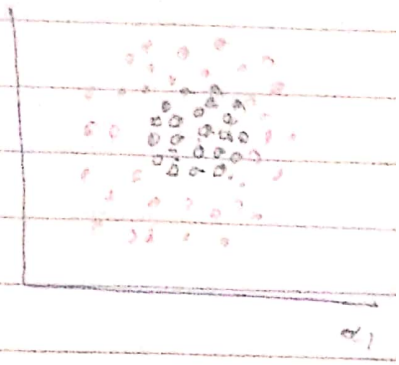
Central line = maximum margin hyperplane or maximum margin classifier

Apple - orange example

Why SVM — For boundary line scenerios (i.e you are having a tie whether a vector point

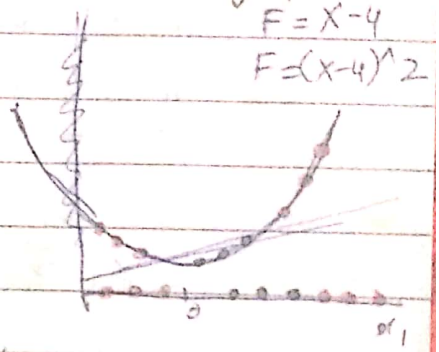
SVM - Can be used for classification like fraud/no fraud.
SVR - can be used for regression problems

belongs to group 1 or group 2), support vectors works very well as compared to other classifier methods.



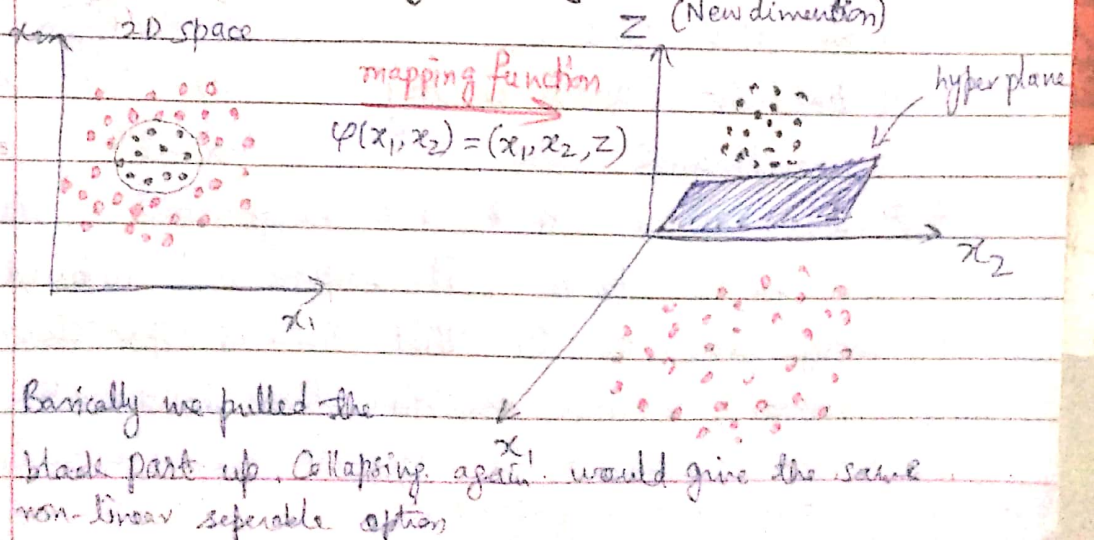
How would you create a decision/classification boundary using linear line. Here we will have to use a higher dimension for this case

→ One way is Some transformations can be done on data like subtracting a constant (which is not helping much here) so we are squaring it to make it parabolic instead of keeping all points on one axis. (quadratic). Now we can use a line to seperate black ones from red (linearly) but it is computationally extensive.



Linearly seperable

→ Another way is, use a mapping function on the 2D space, which seperates 2 groups using a hyperplane, by adding another dimension Z.



Basically we pulled the black part up. Collapsing again would give the same non-linear seperable option

Red has different z value, black has different value but we thats why we are able to seperate based on z using a hyperplane.

Mapping to higher dimension space can be highly compute intensive. Solution to this is

Kernel Trick

Gaussian RBF Kernel invented it

K - Kernel function σ - spread

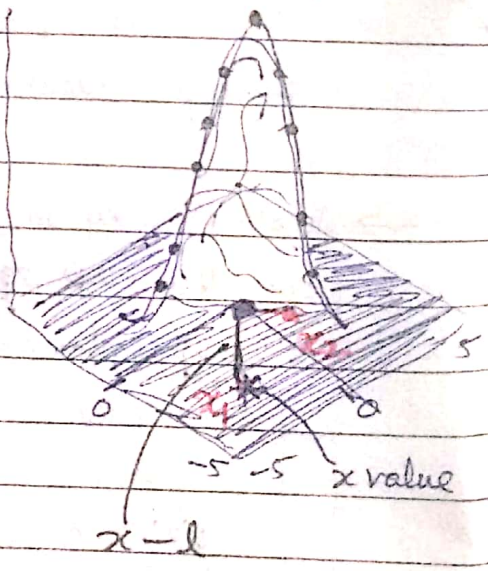
x - a vector

L - landmark

$$K(\vec{x}, \vec{l^1}) = e^{\frac{-\|\vec{x} - \vec{l^1}\|^2}{2\sigma^2}}$$

vertical axis is the result of above function. earlier all the x values are in 2D (shaded) plane. Using Kernel function, we get values for x input in 0 to 1 range. It would fall somewhere on the cone.

Bold dot (0,0) is my landmark.



$e^{\text{large number}} = 0$

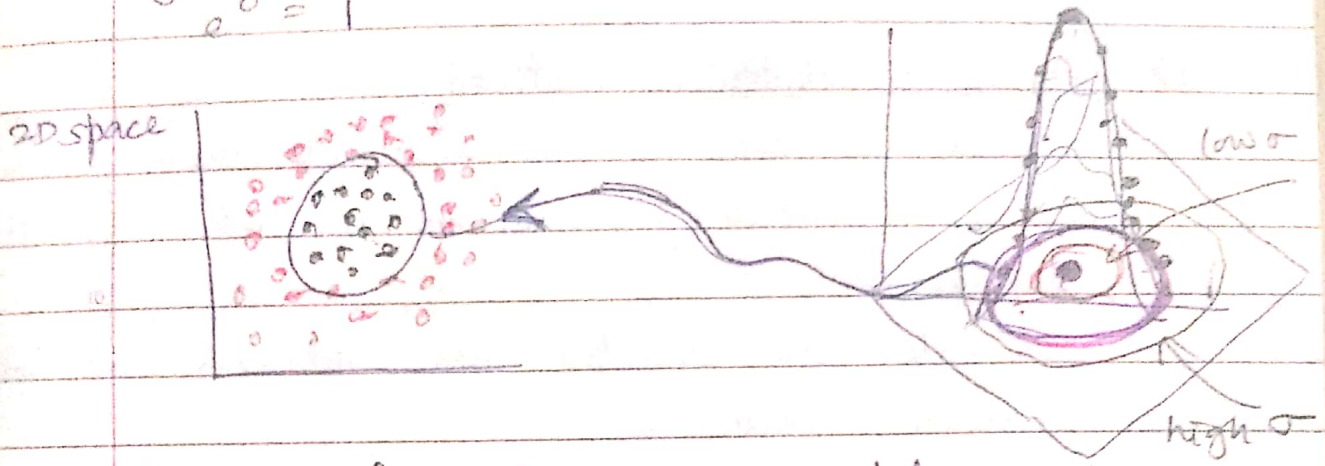
$e^{\text{small num(-)}} = 1$

→ Lets assume $x_1 - l$ is a large number, its square is further large, its exponential would be very close to 0. That means for those x_1 values, which are far from landmark, K value would be close to zero (would not gain much height)

→ If $x_2 - l$ is small, its square would be further small, exponential would be close to 1, so it would nearly gain height close to 1.

→ However, Landmark point would gain maximum height, as $x - l = 0$

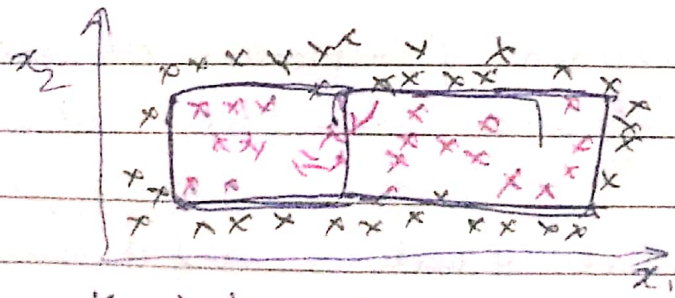
$e^0 = 1$



Bottom of this one would work as decision boundary, without the need of new dimension.

→ If spread of the data increases, decision boundary would increase

→ If spread of the data decreases, decision boundary would decrease.



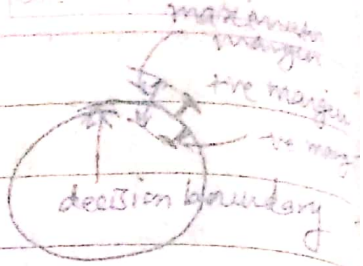
We can use 2 Kernel functions to determine decision boundary

$K_1 + K_2$ <=> - Same value — pink

$K_1 + K_2$ >≠< - Same value — Black

Sigmoid Kernel function can also be used. Here also we can check support vectors and maximum margins around them circular.

decision boundary



Missing Data imputations

1) Regression Substitution method

- Use complete data points to calculate the regression of the incomplete variable on the other complete variables
- Substitute the predicted mean for each unit with a missing value
- Use information from the joint ~~juridi~~ distribution of the variables to make the imputation
- Regression mean imputation can generate unbiased estimates of means, associations and regression coefficients in a much wider range of settings than simple mean imputations.
- However, one problem remains. The variability of the imputations is too small, so the estimated precision of regression coefficients will be wrong and inferences would be misleading.

In regression, we calculate y based on many x. In case of missing values, that value becomes y and all other values become x which helps us in finding missing value (Jacknife is a resampling crossvalidation technique. this is to find missing values)

y	x_1	x_2	x_3	x_4
1	1	5	4	5
2	7	10	9	9
6	•	2	6	15
5	11	8	8	
11		12	2	25

To get •, I can use $x_1 \sim (y, x_2, x_3, x_4)$ where I would pick only first 2 rows. As these 2 rows only has complete information

So we can run multiple regression equation here

$$x_1 = \beta_0 + \beta_1 y + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Regression Imputation example with fitness data -

- Using Fitness data set
- Variable oxygen has complete data.
- Variable RunTime has 3 observations missing.
- Variable RunPulse has 3 observations missing together with RunTime (4, 11, 14) and five on its own (5, 8, 18, 19, 25)
- So we can develop 3 lines of regression:

1) RunTime on Oxygen to predict missing observations 4, 11, 14
2) RunPulse on Oxygen to predict missing observations 4, 11, 14
3) RunPulse on Oxygen and RunTime to predict missing observations 5, 8, 18, 19, 25.

(2) K-Nearest Neighbor Approach

- Another way to find missing data is K-NN.
- It is quite simple in principle but effective and often preferred over some of the more sophisticated methods described above.
- Nearest neighbors are records with similar data patterns. Average of K-nearest neighbors' completed data are used to impute the value of a variable.

MICE package, MCAR Package

that is missing its value.

- K can be set by the analyst (5 to 10 is adequate)
- Advantage of KNN, it assumes data is missing at Random (MAR), missing data depends only on observed data, which means KNN approch is able to take advandage of multivariate relationships in the completed data
- Disadvantage is, it does not include a component to model random variation, consequently uncertainity in the imputed value is underestimated

(3) Multiple Imputation

- Parameter estimates (missing values) using this approach is nearly unbiased.
- It involves use of random components to overcome the problem of underestimation of standard errors
- Here we impute multiple complete dataset and then combines the results of multiple analysis using fairly simple rules (average etc)
- Basically, missing numbers are imputed using an appropriate model, multiple times, then multiple complete data set are used for final version.

Bias & Variance

Main goal of supervised model - Prediction

Prediction error - reducible + irreducible error

reducible: due to unfit model (which we can minimize)

irreducible: noise (which we can't minimize), irregularities, not sure where it is coming from

- One technique to noise reduction - smoothing, however it reduces very less noise. Not under control of analyst
- Reducible error - Bias, Variance
- Bias happens due to wrong assumptions. Someone who has domain knowledge or working on a dataset for long, takes things for granted or ~~measures~~ takes some wrong assumptions (whole modelling is based on some thought which at the end goes wrong) eg. person ~~who~~ in delhi has high chance of going defaulter or a person with permanent house has less chances of going default
- Complex models (more variables or more constraints) lead to high bias
- Variance - spread of data. Error due to sampling due to the sampling of the training set
- Model with high variance fits training set closely (overfitting). Variance is model's sensitivity to specific sets of training data.

n = number of records (say 10,000)
p = number of predictors

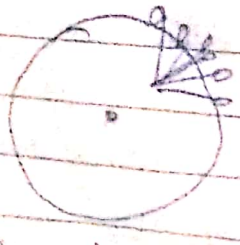
(1) if $n \geq p$ ($p = 10$)
Output will have low variance, low bias as sufficient n is there to capture all scenarios. $\frac{n}{p}$ is very high. Also, my o/p is not dependent on only 1 or 2 p but 10 p. So, I can comparatively less bias

(2) if $n \approx p$ (say $n = 100$, $p \approx 100$)
Output will be high variance

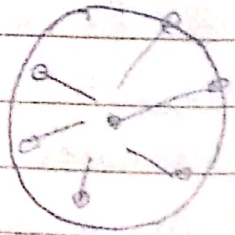
(3) if $n < p$ (say 2 records, 100 predictors)
variance will go to infinite scale. So more data better predictions

Bias Variance trade off

Many a times, we may have to deal with bias variance tradeoff, i.e. we may have to choose b/w low bias - high variance (less restrictions, more variance) or high bias - low variance (more restrictions, less variance)



high bias low variance (model is consistent but inaccurate on average)



high variance, low bias (inconsistent but accurate on average)

High bias because it is biased on one observation
If my email has capital letter, I will just look for that
More than 10 capital letters → spam
Too General

So I have to do bias variance tradeoff (optimum point) where model is neither highly biased nor highly variant,

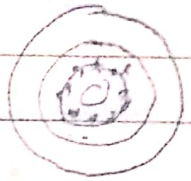
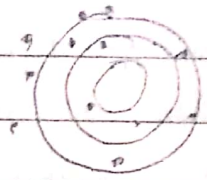
with more data, variance is controlled

Overfitting - Too specific (High variance)

Underfitting - Too general (~~low variance~~) (High bias)

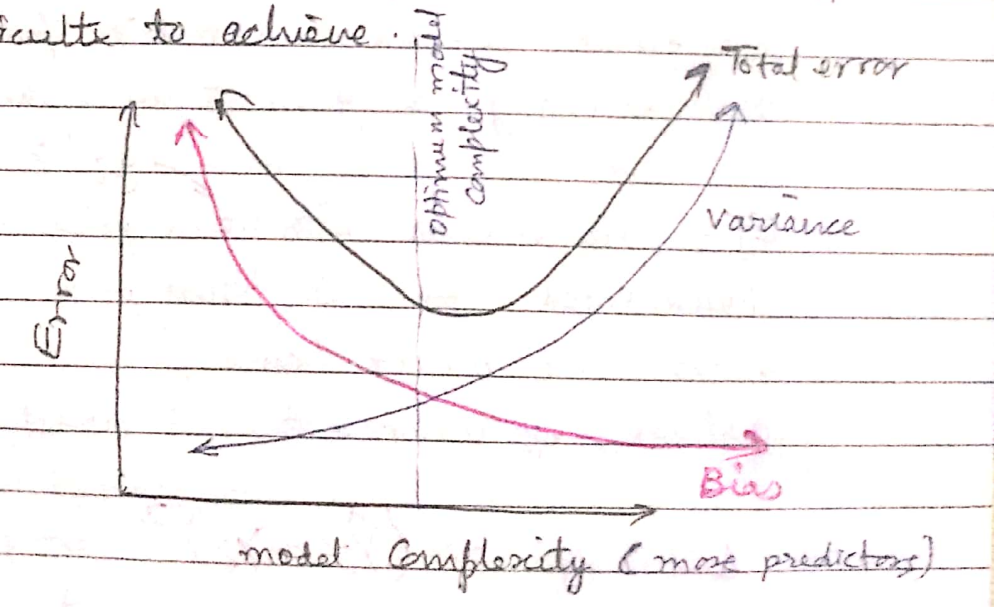
(decision is impacted by one input)

→ model fits training data a lot better than test set

	Low variance	High variance
Low bias		
high bias		

variance increases, spread of data increases

Low bias, low variance, desired state but difficult to achieve.



Spam mail example (count of capital letters and explanation marks)

optimum point - decrease in bias = increase in variance

Distribution of data should be similar in train & test (plot graph techniques) [Ex: Mobile ph camera & DSLR Camera]

2 methods to overcome overfitting

i) Reduce the model complexity (decrease # of variables in prediction)

ii) Regularization techniques

↳ Ridge and Lasso

Ridge and Lasso regression (Regularization)

- Every ML problem is an Optimization problem where the intent is to either find maximum or minimum of a specific function.
- This function is called loss function or cost fn. For ex minimizing error fn, maximizing accuracy fn in case of regression/classification
- Loss function is the main metric for evaluating the accuracy of your trained model.

Eg - Data is given with house prices for different house sizes, for this house price prediction model, we predict house prices based on their features, by using a trained model ← $\hat{y}_i$

y_i ← (ground truth)

loss $L_i = (\hat{Y}_i - Y_i)^2$ ← for a specific data point

Loss function $L = \sum (\hat{Y}_i - Y_i)^2$

↑ Mainly used in logistic regression algorithms. Also called as Quadratic loss or least Squares

minimize

Objective is to ~~define~~ this loss function as much as possible to be close to the ground truth.

Every ML problem defines its own loss function according to its goal (RMSE, MAPE - etc)

- More features (variables) may reduce the loss However, need to be careful about overfitting

Concept of Regularization

i/P variables

- We normally keep the same number of features, but reduce the magnitude of coefficients.
- Basically we fit a model involving all p predictors Estimated coefficients are shrunken towards zero.
- This shrinkage (also known as regularization) has the effect of reducing variances.
- Depending on type of shrinkage, some coefficients may be estimated as exactly zero.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

β_s help in predictions and reducing least square estimates

β_s which are contributing heavily on Y, those β_s would be retained. β_s which do not contribute much on Y will be shrinked (by factor λ)

Ridge (L2 technique)

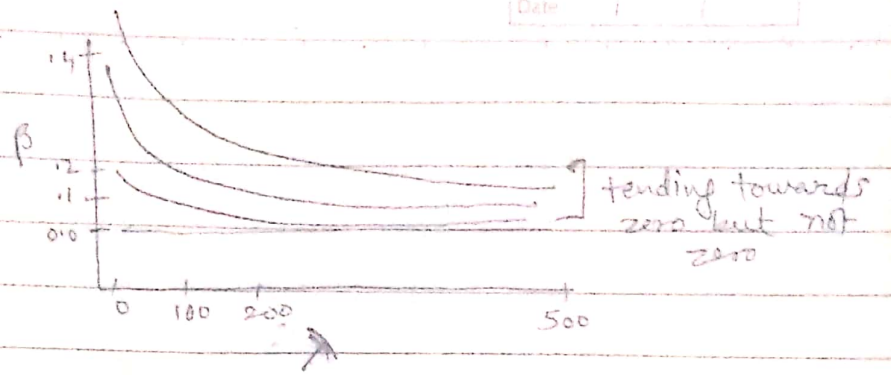
Regularized ML model is a model that its loss function contains another element that should be minimized as well.

$$L = \Sigma(\hat{Y}_i - Y_i)^2 + \underbrace{\lambda \Sigma \beta^2}_{\text{Shrinkage Penalty (Regularization term)}}$$

- Objective is to reduce loss function by punishing it for high values of coefficients β.
- We need to simplify a complex model as much as possible.
- Goal of an iterative process is to minimize the loss function
- By punishing β values, we add a constraint to minimize them as much as possible.
- There is a gentle tradeoff b/w fitting the model, but not overfitting it.
- Ridge is an extension of Linear regression (basically a regularized linear regression model)
- λ, scalar, $>= 0$, that should be learned as well, using cross validation method.

Imp → Ridge enforces β to be lower but does not enforce them as 0.
- It will not get rid off irrelevant features but minimizes their impact on the ~~trained~~ trained model.

Basically we are putting an upperbound on β.



Advantage of Ridge

1) It shrinks parameters and hence prevents multicollinearity (two variables collinear, we reduce impact of one to great extent which in turn reduces its relation with other)

2) Reduces model's complexity by coefficient shrinkage

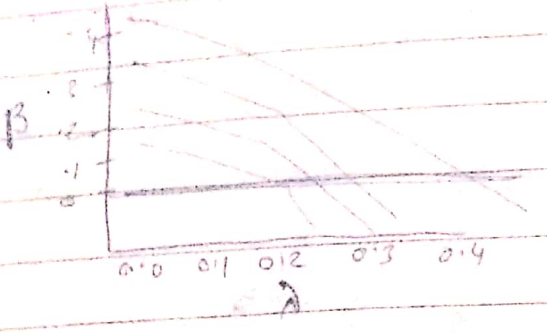
3) Uses L2 regularization technique.

Lasso (Least absolute shrinkage and selection parameter) - L1 technique

$$L = \Sigma (\hat{Y}_i - Y_i)^2 + \lambda \Sigma |\beta|$$

- In lasso, regularization term is in absolute value
- This small difference has huge impact on overfitting underfitting trade off
- Lasso overcomes the disadvantage of Ridge regression by not only punishing high values of β coefficients but actually setting them to zero, if they are not relevant.

Imp- Hence, you might end up with fewer features included in the model than you started with, which is a huge advantage.



Dimentionality reduction is to get fewer dimensions which are not correlated
Lasso is a feature engineering model/ technique which removes those variables which have 0 contribution

L1 and L2 regularization

Date / /

1 — Longley dataset example in R — which provides a well-known example for a highly collinear regression.

Longley Economic Regression data

package glmnet

Re-Sampling Methods

Need of cross validation –

Suppose we create a regression model, based on r^2 (coefficient of determination) we determine its quality of fit. On the training dataset, it gives good prediction accuracy.

But how can we convincingly say that predictions would be accurate with new samples of dataset?

→ One way of validation – Once you run regression and predict on a dataset, perform similar study to collect similar data again which replicates original data as closely as possible, use this new data to predict validity of MLR equation.

its impractical as it requires to conduct a new study to obtain 'validation data', Also it is difficult in ~~repitching~~ replicating the original study

→ Alternative approach is Cross validation which is a more practical approach, as with existing data, we test in all possible ways.

- In CV, original sample is split into 2 parts, one is called training (derivation) sample
- Other is test (validation + test) sample.
- Generally 50/50 for large datasets or 70/30

Ways to split sample

↳ Divide data randomly, thereby eliminating any systematic differences (Out of a sample of 100, 20 records are of different flavour. Any data can go to any sample.

→ Define matched pair of subjects in the original sample and assign one member of each pair to derivation and validation sample (out of 20, 10/10 goes into both)

→ Model is built on one part and validated for accuracy on other part.

R^2 offers no protection from overfitting. However, CV allows to have cases in testing set from the cases in test set, Hence offers protection against overfitting.

Ideal procedure

1) divide dataset into 3, training, test & validation
2) Build optimal model on training set and use test set to check for its predictive capability. Also known as Cross validation step.
3) See how well the model can predict on validation set
4) Validation error gives an unbiased estimate of the predictive power of the model.

Drawbacks of CV

- Validation estimate of test error can be highly variable as it depends on which observation is included in training / validation set.
- As we are using fewer observations in training set as compared to complete data set, validation set

Splunk - to analyze machine logs
SOR - One single source of data, true information (fact/truth)
submit call, update call, sync call.

error rate may tend to overestimate

Leave One Out Cross Validation (LOOCV)

→ Used when we have very small dataset
→ Similar to Jack Knife
→ A set of n data observations is repeatedly split into a training set containing all the observations except one which works as the validation set.
→ First training set contains all but 1st observation
→ 2nd training set contains all but 2nd observation and so on.
→ Test error is then estimated by averaging the 'n' resulting MSE
→ Similar to Jackknife. JKn is used for modelling. LOOCV is used for sampling (if data set is small).

If data set is large we dont need it. LOOCV can result in overfitting also as we are training the model for all situations. But required to build at least a ~~small data~~ good model with small dataset.

K-fold CV

- An alternative of LOOCV
- n observations are divided into k groups (or folds) of approximately equal size (for ex 5 fold)
- 1st fold is treated as validation set, rest all data as training set
- In second training set, 2nd fold is considered as validation set, rest all as training set
- Test error is calculated by averaging K estimates of MSE for each validation set.

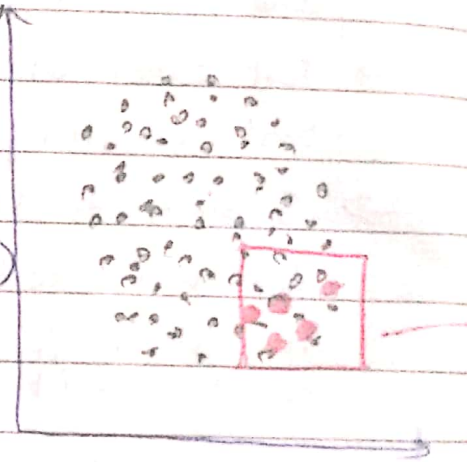
20	20
20	20
20	20
20	20
20	20

K = 10
for 200 observations

$$CV_{(K)} = \frac{1}{K} \sum_{i=1}^{K} MSE_i$$

SMOTE (Synthetic Minority Oversampling Technique)

- Used for unbalanced data, where the favoured class is less/small in number (minority class) eg: prediction of frauds, or people having certain decease (out of 10,000 only 50 would belong to favourable record)
- In above case predicting the probability of non-default (0) is easier/accurate than predicting the probability of default (1)
- Unbalanced dataset may give Biased predictions and Misleading accuracy.
- Examples
  - Credit card frauds (0.5 vs 99.5 %)
  - Manufacturing defects
  - Rare disease diagnosis
  - Natural disasters (earthquakes)
  - Enrolment to premier institutes



To resample unbalanced dataset

→ Either increase the minority class
→ Or decrease the majority class
→ Random Under Sampling
  - Randomly remove majority class observations
  - Helps balance the dataset
  - Discarded observations could have important observations
  - May lead to bias

Total observation 1000
Fraudulent = 10 or 1%, Normal = 990 or 99%
Reduce normal to 90
Fraudulent = 10 or 10% now

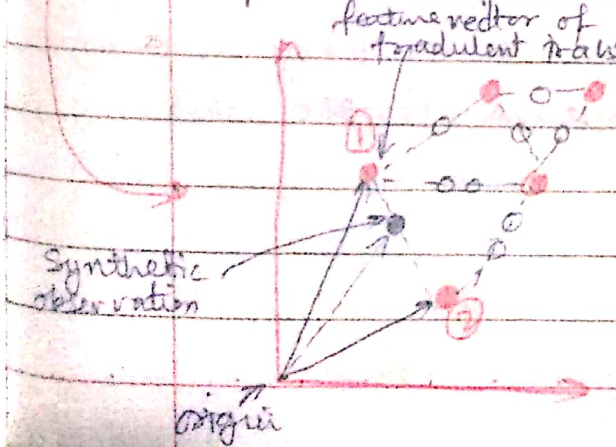
→ Random Over sampling
- → Randomly add more minority observations by replication
- → No information loss
- → Prone to overfitting due to copying same info

increase fraudulent by 100
so new fraudulent = 110 or 10% out of total 1100.

Doing Under/Over sampling manually is Extrapolation. Using algorithm, SMOTE

SMOTE (DMwR package in R)

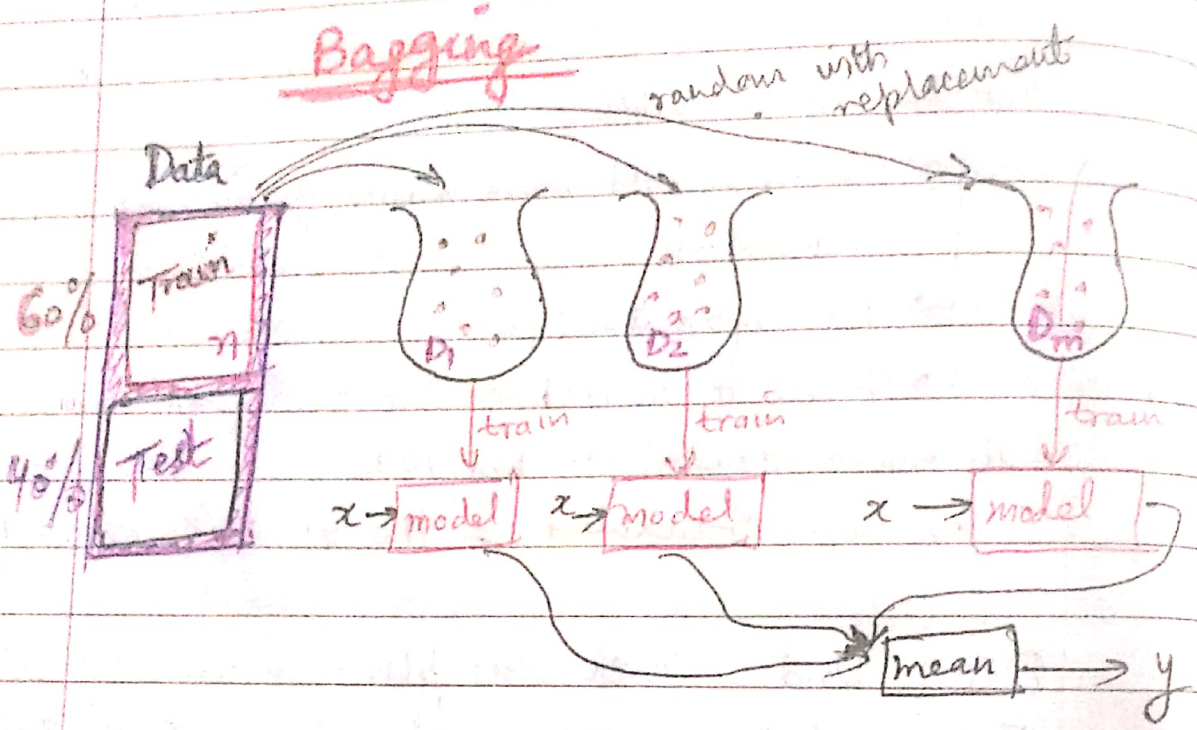
→ Creates new "synthetic" observations
→ Identify the feature vector and its nearest neighbor
→ Take difference b/w the two
→ Multiply the difference with a random number b/w 0 and 1.
→ Identify a new point on the line segment by adding the random number to feature vector
→ Repeat the process for identified feature vector



① is a vector of multiple independent variables (plotted on a 2d plane, distance from the origin is found and based on that distance its nearest neighbour is found.)

Then the difference b/w the distance of two neighbours is calculated and a synthetic point is created b/w the two points

Bagging

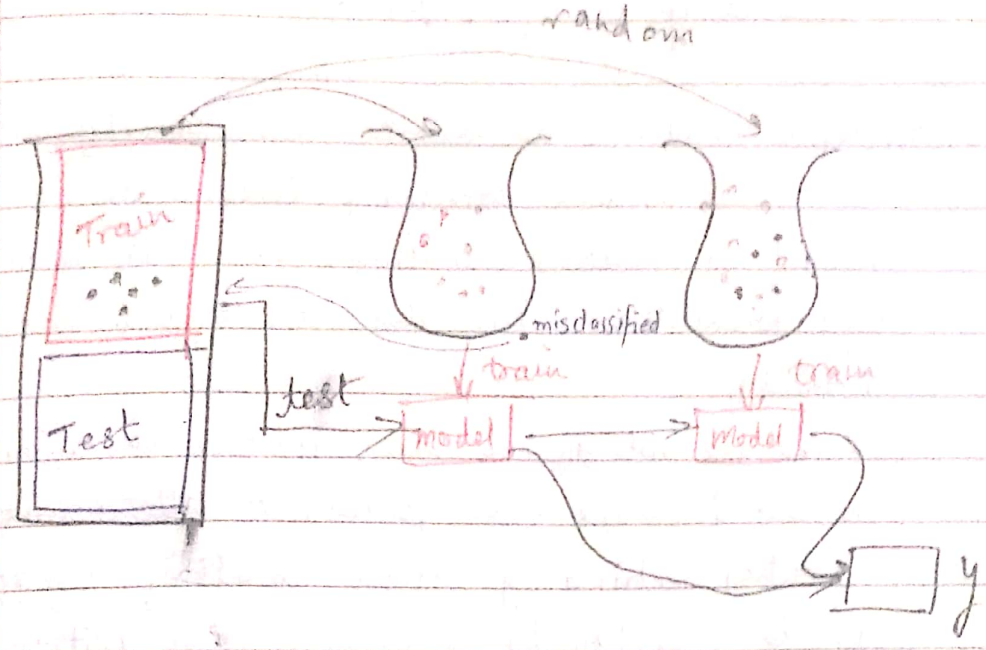


n Observations are placed in m bags randomly. 1 observation can go into multiple bags. Also a particular observation may not go in any bag, as observations are placed randomly with replacement.

Now we prepare our model, train it based on data in each bag, collect the output from each bage, average it to get Final output.

Same bagging re sampling technique was used in Random Forest also.

Boosting



In boosting again data samples (bags) are created with random observations, However, difference is, when model is created and trained against data of bag 1, we analyze its output (accuracy / error). Those data points which are misclassified (error) go back to the training data set with increased weightage.

For ex, after 1st bag, out of 10, 8 are rightly classified, 2 are wrongly classified. These 2 go back to the training set with higher weightage so that their probability to be selected in 2nd bag is more.

So we are boosting our model to learn from the errors again and again

Bag 1 classify a red dot as black (so is an error) this black dot record goes to training dataset from where it would be picked in 2nd bag.

In 2nd bag, a red may again be classified as black or a black may again be classified as red.

Gradient boosting gives very high accuracy (very popular)

- Think residuals as mistakes committed by our model.
- Tree based models (considering decision tree as base model for gradient boosting) are not based on such assumptions, but think about this assumption logically.
- if we are able to see some pattern of residuals around 0 we can leverage that pattern to fit a model.

Imp - Intution behind gradient boosting algorithm is to repetitively leverage the patterns in residuals and strengthen a model with weak predictions and make it better

- Once we reach a stage whether residuals do not show any pattern that can be modelled, we can stop modelling them (Otherwise it might lead to overfitting)
- So we minimize loss function such that test loss reaches its minima
- Basic assumption of linear regression is that sum of its residual is 0 (i.e. residuals with +ve and -ve values should be spread randomly around zero)

Check R codes for 2nd residency, Jatinder covered Ridge/Lasso/SVM/Smote/Churn using xgboost etc

In last example. Ridge / Lasso / SVM,
we started with 24 variables and now we are trying to have best subset by (optimized number of variables) by following methods like R^2, adj R^2, RSS, CP value

Tidiver used generic tune function to tune SVM model (which can be used for any model)

2

Without ~~Gridsearch~~,

Prof Narain Gupta
MDI Gurgaon

Optimization Techniques

No of applications like finance, marketing, operations, projects related IT application

7 steps of Problem solving: (First 5 steps are process of decision making)

Structuring the problem
1) Identify and define the problem
2) Determine the set of alternative solutions
3) Determine the criteria for evaluating alternatives

Analyzing the problem
4) Evaluate the alternatives
5) Choose an alternative (take decision)

6) Implement the selected alternative
7) Evaluate the results

Optimization is most useful when there is scarcity of resources

When we have plenty of time, cost, resources, we are not bothered about optimization.

Similarly for PM, in order to complete a project, we need to work in 3 main constraints time, scope and cost.

Basically Optimization is Resource optimization (money, machine, man power, time - etc)

In today's world for cost effectiveness and competitiveness, it is required to achive a goal within certain constraints (like time, cost etc)

Linear Programming

- Programming means choosing a course of action.
- LP means choosing a course of action when the mathematical model of the problem contains only linear functions.
- Maximization or minimization of some quantity is the objective in all LP problems.
- All LP problems have constraints that limit the degree to which the objective can be pursued.
- A feasible solution solution satisfies all the problem constraints.
- An optimal solution is a feasible solution that results in the largest possible objective function value when maximizing (or smallest when minimizing).
- A graphical solution method can be used to solve a linear program with two variables.
- LP Problem - if both objective fuction and the constraints are linear.
- Linear functions - in which each variable appears in a seperate term raised to the first power and multiplied by a constant (can be 0)
- Linear constraints - Linear functions that are restricted to be "less than or equal to", "equal to", ", ≥" a constant

Ex-1 100 pound aluminium, 80 pound steel
2 p alm, 3 p steel for each bike (deluxe)
→ Solution $\frac{100}{2} = 50$
$\frac{80}{3} = 26.6$

26 bikes can be made as steel is exhausted

Ex-2 100 p aluminium, 80 p steel
4 p alm, 2 p steel for professional bike
→ Solution $\frac{100}{4} = 25$
$\frac{80}{2} = 40$

25 bikes can be made as aluminium is exhausted

Ex-3 Deluxe - $10, Professional - $15
Need to maximize profit
100 p of aluminium, 80 p of steel
Delux - 2p alm, 3p steel
professional - 4p alm, 2p steel
→ Solution ① 25 professional
profit = 25 X 15 = $375
near optimal ② P-20, D-10, profit = 20X15 + 10X10 = $400
optimal ③ P-17, D-15, profit = 17X15 + 15X10 = $405
Hit n trial approach

Structured approach

X - no of units made at optimal for Delux
Y = no of units made at optimal for professional
Max (10X + 15Y) — Criteria/Objective fn
Aluminium consumed = $2X + 4Y \leq 100$ } Constraints
Steel consumed = $3X + 2Y \leq 80$
X, Y should be non-negative.

This structured way where criteria as well as constraints are linear called Linear Programming (LP)

X, Y - decision variables

$2X + 4Y \leq 100$ → RHS
↓
LHS arithmatic operator

$10, $15 - Profit coefficients

	A	B	C	D	E	
	Var	Solution options	obj coeff	Aluminium	Steel	market demand
Common structure in excel for LP	X	15	10	2	3	1
	Y	17.5	15	4	2	D
	Consumed		=Sumproduct (B*C)	= Sumproduct (B*D)	= Sumproduct (B*E)	← LHS
	Constraints			≤	≤	← operator ≤ ≥
	Available			100	80	← RHS →

maximize ↑ Constraints

+ X, Y - non-negative

Set as objective in solver to maximize
by changing cells B2, $B $3
put constraints

or minimize in case of cost

Select Solving method as (Simplex LP) or check Assume linearity and Assume non-negativity in options for (2003, 2007 versions)

Solver gives X = 15 with $412.5
Y = 17.5

So in one production cycle we can manufacture 17.5 professional and (so 17 complete, 1 half) in next cycle we can complete this professional so in next cycle 18 profession would come outs of manufacturing unit

IBM CPlex - Solver tool - 25 lakhs p.a subscription cost
Most powerful optimizer in the world, solves 10 lakhs variables, 10 lakhs constraints in seconds

Integer Programming - when instead of decimal values of X and Y, we mention that X and Y should be integer (for ex. in cycle case, if it is recurring production, it is better to produce 17.5. but if it is 'one-time' production, we want to produce complete cycles which maximize the profit)

In solver add condition that X, Y as "int"
It gives X = 14, Y = 18, profit as $410.
↳ optimal solution in one time production case.

RHS → Suppose we add a condition that maximum market demand for X is 6.
One more constraint - $x \leq 6$ needed.
But, its better to put as $1x + 0y \leq 6$
so that the model is generic (to avoid hardcoding)

$$aX + bY \leq 6$$

Take problem of marketing domain
Media Selection

As a marketing manager, you have to decide where to invest marketing budget. Budget cannot directly be converted into sales but assigning/allocating it to right buckets can improve sales.
eg. spending marketing budget for advertising about a play school (whole india or within 3km by 3km area)
Advertising instruments - hoadings, buses/auto, Social media, Print media, campaign, radio, exhibitions, events at a

place.

How to decide which instrument to pick of TV, which channel, news/cartoon, if news, what time, afternoon/evening/sunday

How to optimize ??

⇒ Objective is to maximize reach, frequency and quality of exposure.

⇒ Restrictions arise during consideration of company policy, contract requirements and media availability

→ SMM company budget = $282,000 for adv.

→ Have to adv. during 1 wknd in Nov (Fri, Sat, Sun)

→ Options available: daytime, evening news, sunday game time

→ Mixture of 1 minute TV spots is desired.

Ad type	Estimated audience Reached with each ad	Cost per Ad
Daytime	3000	$5,000
Evening News	4000	$7,000
Sunday Game	75000	$100,000

Constraints

- at least one ad of each type
- 2 game-time ad slots available
- 10 daytime, 6 evening news spots available daily
- Atleast 5 ads per day should be there
- Spend on Friday ≤ $50,000
- Spend on Saturday ≤ $75,000
- No limit given for Sunday.

Objective is to maximize the reach in given budget

DFR - Daytime Friday (DSA, DSU), Daytime Sat (Sun)
EFR - Evening Friday (Sat, Sun), ESA, ESU
GFR - Gametime Friday (Sat, Sun), GSA, GSU

Maximize $300[DFR + DSA + DSU] + 4000[EFR + ESA + ESU] + 75000[GSU]$

Constraints

$DFR + DSA + DSU \geq 1$ ①
$EFR + ESA + ESU \geq 1$ ②
$GSV \geq 1$ ③
$GSU \leq 2$ ④
$DFR \leq 10$ ⑤
$DSA \leq 10$ ⑥
$DSV \leq 10$ ⑦
$EFR \leq 6$ ⑧
$ESA \leq 6$ ⑨
$ESU \leq 6$ ⑩
$DFR + EFR \geq 5$ ⑪
$DSA + ESA \geq 5$ ⑫
$DSV + ESU + GSV \geq 5$ ⑬
~~DSR + EFR ≤ 50000~~
$5000\,DFR + 7000\,EFR \leq 50{,}000$ ⑭
$5000\,DSA + 7000\,ESA \leq 75{,}000$ ⑮
$5000(DFR + DSA + DSV) + 7000(EFR + ESA + ESU) + 100{,}000\,GSV \leq 282{,}000$ ⑯
DFR - - GSV should be non negative ⑰

Now refer to already done excel sheet.

What happens behind a Solver

LP formulation

Max $5x_1 + 7x_2$ — Objective function

s.t. $x_1 \leq 6$

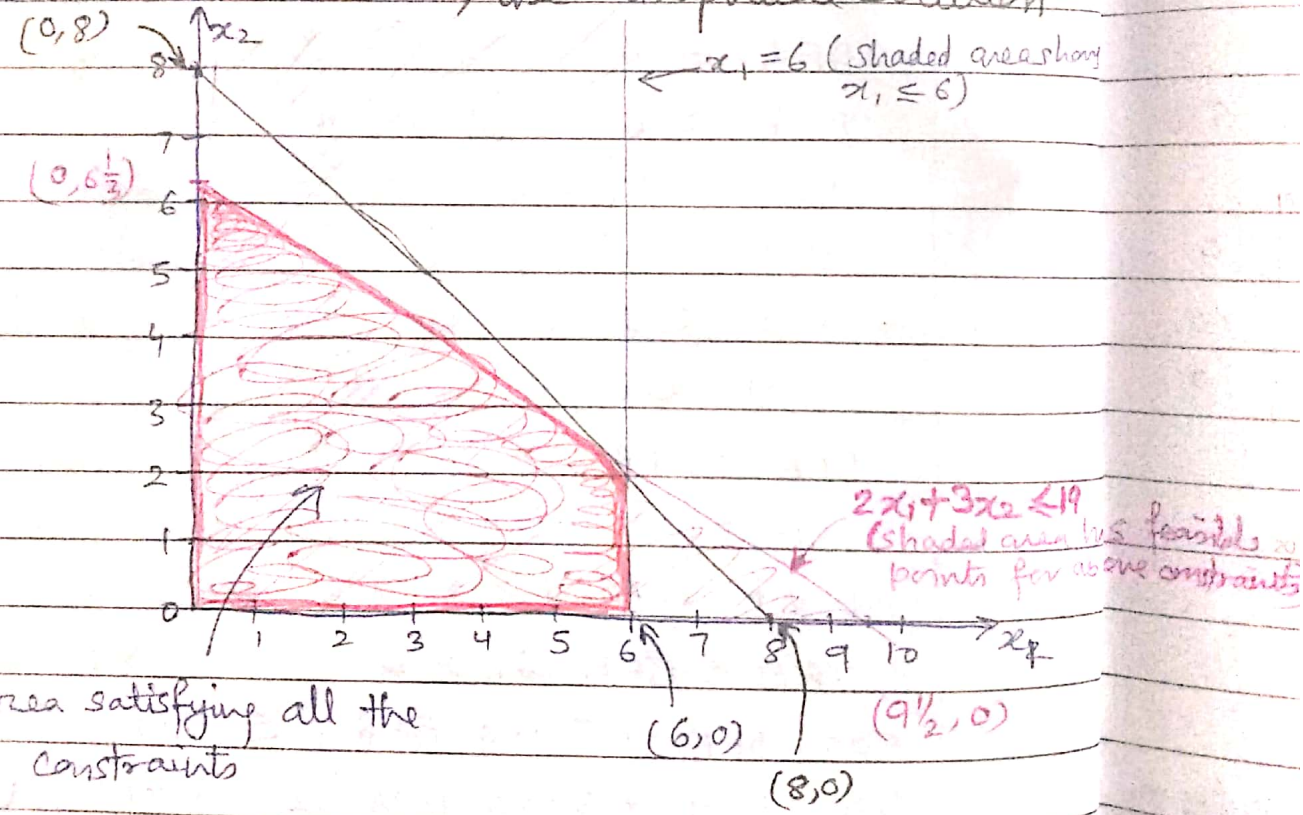
$2x_1 + 3x_2 \leq 19$ } Regular constraints

$x_1 + x_2 \leq 8$

$x_1 \geq 0$ and $x_2 \geq 0$ — Non-negativity constraints

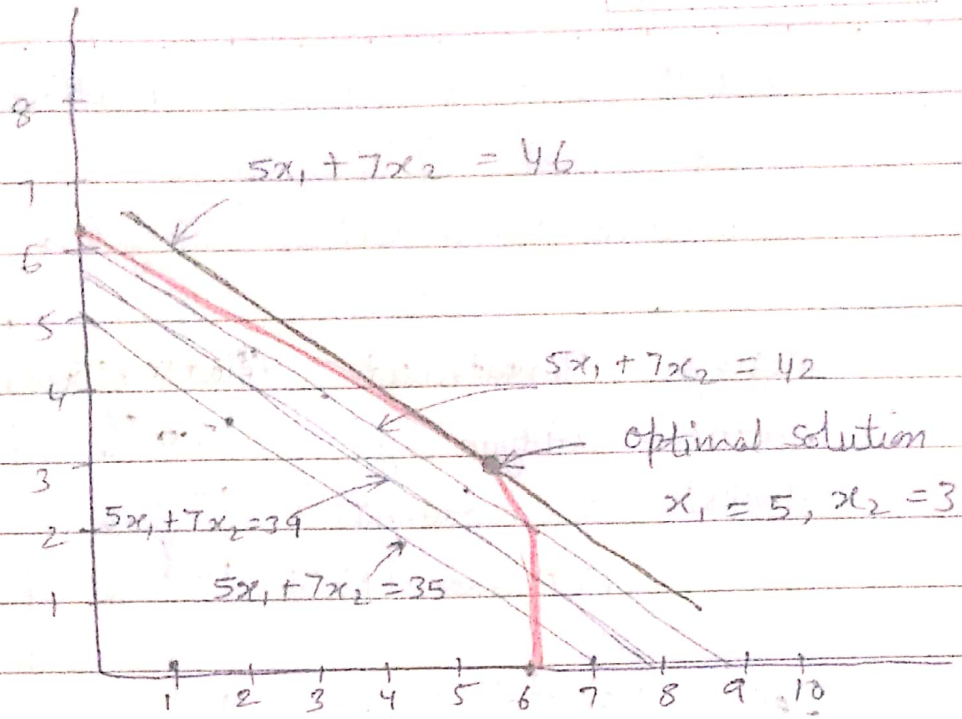
Optimal solution — $x_1 = 5$ $x_2 = 3$ and obj fn = 46

How do we reach this, use Graphical solution



area satisfying all the constraints

Now need to put objective constraints



Intersection of 2 equations

$x_1 + x_2 = 8$

$5x_1 + 7x_2 = 46$

In 2 dimension, lines are there
However in multidimensions, planes are used, which cuts the area not satisfying constraints

Portfolio Selection (Asset allocation) Problem

Asset allocation - determines how to allocate investment funds across variety of asset classes, like stocks, bonds, mutual funds, real estate.

Portfolio models - determines % of funds that should be assigned in each class.

Objective - to create a portfolio that maximizes return and minimizes risk.

Need to construct optimal portfolio for client -
- Cash to invest $400,000
- 10 different investments, falling into 4 categories

Category	Investment	Aftertax return	Liquidity factor	Risk factor
Equities (Stocks)	Unidyde Corp	15.0%	100	60
	CC's restaurants	17.0%	100	70
	First general REIT	17.5%	100	75
Debt (Bonds)	Metropolis Electric	11.8%	95	20
	Unidyde Corp	12.2%	92	50
	Lewisville Transit	12.0%	79	22
Real Estate	Realty Partners	22%	0	50
Money	T-Bill account	9.6%	80	0
	Money Mkt Fund	10.5%	100	10
	Saver's certificate	12.6%	0	0

Constraints -
1) weighted avg liquidity factor for portfolio ≥ 65
2) Weighted avg risk factor for portfolio ≤ 55
3) Investment in Unidyde stocks or bonds $\leq 60,000$
4) Investment in any one category (except money category) $\leq 40\%$ of investment
5) Inv in any one investment (except money market fund) $\leq 20\%$ of inv
6) Inv in Money market fund $\geq \$1000$
7) Inv in Saver's certificate $\leq \$15,000$
8) Inv in debt $\geq \$90,000$
9) Inv in T-Bill account $\geq \$10,000$

X_1 — X_{10} – money invested in 1--10 instruments

Objective function – $15X_1 + 17X_2 + 17.5X_3 + 11.8X_4 + 12.2X_5 + 12X_6 + 22X_7 + 9.6X_8 + 10.5X_9 + 12.6X_{10}$

Constraints

① $\frac{100}{400,000} X_1 + \frac{100}{400,000} X_2 + \frac{100}{400,000} X_3 + \frac{95}{400,000} X_4 + \frac{92}{400,000} X_5 + \frac{79}{400,000} X_6 + \frac{0}{400,000} X_7 + \frac{80}{400,000} X_8 + \frac{100}{400,000} X_9 + \frac{0}{400,000} X_{10} \geq 65$

$\frac{x_1}{\Sigma x}$ is a non linear eq. (x_1^2 is also non-linear)
so take $\leq x$ to RHS

So C1 $100X_1 + 100X_2 + \cdots + 100X_9 + 0X_{10} \geq 65 \times 400,000$

C_2 $100X_1 + 100X_2 + \cdots + 100X_9 + 0X_{10} \leq 55 \times 400,000$

C_3 $X_1 + X_5 \leq 60,000$

C_4 (a) $X_1 + X_2 + X_3 \leq 0.4 \times 400,000$

(b) $X_4 + X_5 + X_6 \leq 0.4 \times 400,000$

(c) $X_7 \leq 0.4 \times 400,000$

C_5 (a) $X_1 \leq 0.2 \times 400,000$

(b) $X_2 \leq 0.2(400,000)$

(c) same with X_3 to X_{10} except X_9

(9 constraints here)

C_6 $X_9 \geq 1,000$

C_7 $X_{10} \leq 15,000$

C_8 $X_4 + X_5 + X_6 \geq 90,000$

C_9 $X_8 \geq 10,000$

C_{10} $X_1 + X_2 + \cdots + X_{10} \leq 400,000$

Now refer to already done excel sheet

Imp takeaways —

after solving in solver

Surplus → Subtract LHS - RHS for ≥ constraints

Slack → Subtract RHS - LHS for ≤ constraint

Wherever the value comes +ve is not helping in finding the optimum solution.

Even if we remove these constraints, solution would not be impacted.

However, this cannot be found out before solving.

In this solution, 64,355 is the yearly return

If slack or surplus is 0, these constraints are working and called Binding constraint.

changing anything in these constraints can relax or ~~worsen~~ worsen the problem.

↑ Known as Sensitivity analysis. What if. or sensitivity or Scenario analysis to check what if I to change something (what is the effect)

Revenue Management for Airline Industry

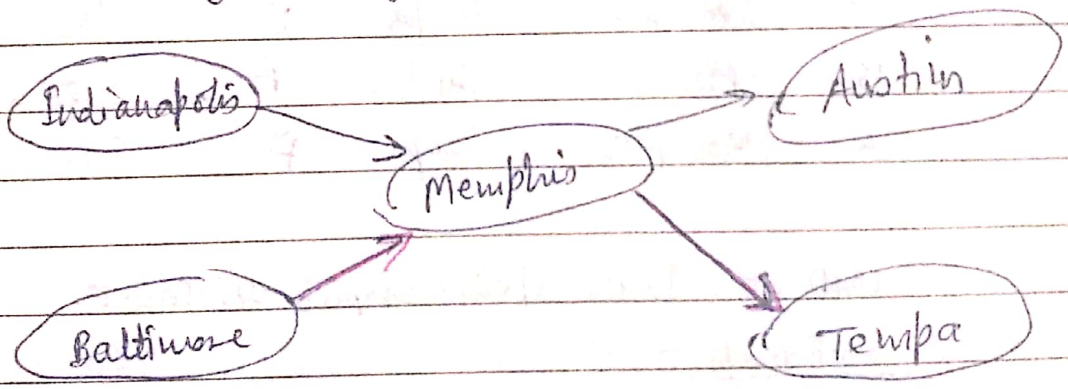
- It involves manging the short term demand for a fixed perishable inventory in order to maximize revenue profit.
- This methodology was first used to determine how many airline seats to sell to an early-reservation distount fare and many to sell (IPL match, movie theatre etc)

= Application areas include hotels, apartment, rentals, car rentals, cruise lines and golf course, Dynamic Pricing (changing based on demand and supply)

- How many seats to block for economy / business
- To have operating break-even, seats are initially sold in less. Once break even is achieved dynamic Pricing (for profit making) starts

No shows are already predicted and tickets are sold in high rate accordingly. (over-booking)

Leapfrog Airways - 2 planes (WB828) based in Indianapolis and Baltimore, stopover Memphis (1hr)
- Seating capacity 120



- Two fare classes, Discount fare D
  full fare F
- follows ODIF - origin destination itenary fare

Objective is to find out how many seats it should allocate to each ODIF

ODIF	Origin	Destination	Fare Class	Fare	Demand
1	Indianapolis	Memphis	D	175	44
2	IP	Austin	D	275	25
3	IP	tempa	D	285	40
4	IP	Memphis	F	395	15
5	IPolis	Austin	F	425	10
6	IPolis	tempa	F	475	8
7	Baltimore	Memphis	D	185	26
8	Bmore	Austin	D	315	50
9	Bmore	Tempa	D	290	42
10	Bmore	Memphis	F	385	12
11	Bmore	Austin	F	525	16
12	Bmore	Tempa	F	490	9
13	Memphis	Austin	D	190	58
14	Memphis	tempa	D	180	48
15	Memphis	Austin	F	310	14
16	Memphis	Tempa	F	295	11

IMD = Indianapolis Memphis Distance

Constraints –

~~① IMD + IMF ≤ 120~~

① IMD + IAD + ITD + IMF + IAF + ITF ≤ 120

② BMD + BAD + BTD + BMF + BAF + BTF ≤ 120

③ IAD + IAF + BAD + BAF + MAD + MAF ≤ 120

④ ITD + ITF + BTD + BTF + MTD + MTF ≤ 120

⑤ IMD ≤ 44 ⑥ IAD ≤ 25 ⑦ ITD ≤ 40 ⑧ IMF ≤ 15

⑨ IAF ≤ 10 ⑩ ITF ≤ 8

⑪ BMD ≤ 26 ⑫ BAD ≤ 50 ⑬ BTD ≤ 42

⑭ BMF ≤ 12 ⑮ BAF ≤ 16 ⑯ BTF ≤ 9

⑰ MAD ≤ 58 ⑱ MTD ≤ 48 ⑲ MAF ≤ 14

⑳ MTF ≤ 11

Product Mix

- Floatways tours has budget $420,000
- Two vendors, sleekboat and Racer
- Requirement is to buy 50 boats (atleast)
- Equal number of boats from each vendor to maintain goodwill
- Total seating capacity required is atleast 200

Boat	Builder	Cost	Maximum seating	Expected Daily Profit
Speedhawk	Sleekboat	$6000	3	$70
Silverbird	Sleekboat	$7000	5	$80
Catman	Racer	$5000	2	$50
Classy	Racer	$9000	6	$110

Objective is to maximize the profit

DV - how many boats of each type to be purchased

Max - $70x_1 + 80x_2 + 50x_3 + 110x_4$

Constraints - $6000x_1 + 7000x_2 + 5000x_3 + 9000x_4 \leq 420,000$

$3x_1 + 5x_2 + 2x_3 + 6x_4 \geq 200$

$x_1 + x_2 + x_3 + x_4 \geq 50$

$x_1 + x_2 = x_3 + x_4$

$[x_1, x_2, x_3, x_4] \geq 0$

Distribution Channel optimization

eCommerce, Supply chain, Logistics
↳ Shows different kind of linear Programming

Delivery Cost per Ton (from Plant / to Mumbai/Delhi)

	Mumbai	Delhi	
Plant 1	24	30	50 Ton (total availability at Plant 1)
	25 Ton	25 Ton	
	demand at mumbai / delhi		

Total cost = 24 × 25 + 30 × 25
= $ 1350

We need to minimize total cost so Plant 1 will supply to Mumbai first (as cost is less), completes its 25 Ton demand and then supply to Delhi and completes its 25 Ton demand.

Now in below case,

Delivery cost per Ton

	Mumbai	Delhi	
Plant 1	24	30	50 Ton (total Capacity of Plant 1)
	25 Tons	45 Tons	

Answer would remain the same, 25 Ton would go to Mumbai first as delivery cost is less there. Rest 25 ton would go to Delhi.

Another situation below

Delivery cost per Ton

	Mumbai	Delhi	
Plant1	24	30	50 Tons
	45 Tons	25 Tons	

Demand at Mumbai & Delhi

In order to minimize total delivery cost, 45 Tons (as per mumbai demand) would be supplied to mumbai first and then rest 5 Tons to Delhi.

Total cost = $45 \times 24 + 30 \times 5 = \$1230$

Another Situation ~~is~~ below

Delivery Cost Per Ton

	Mumbai	Delhi	
Plant1	24	30	25 Tons
Plant2	30	24	25 Tons
	25 Tons	25 Tons	

As Plant1 has capacity of 25 Tons, it would supply to Mumbai all its 25 Tons (to match demand) This would keep delivery cost minimum. Then Plant 2 would deliver its 25 Tons to Delhi.

Total cost = $24 \times 25 + 24 \times 25 = \$1200$

Delivery Cost per Ton

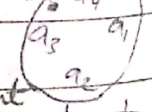
	Mumbai	Delhi	
Plant 1	24	30	25 Tons
Plant 2	30	40	25 Tons
	25 Tons	25 Tons	

if we supply to Mumbai first and then Delhi
Cost = 24 × 25 + 40 × 25 = 1600
which is very high
However, if we take 30 × 25 + 30 × 25,
we get 1500 which is better solution.
As multiple choices were there, so scope of optimization was also there.

Delivery Cost per Ton

	Mumbai	Delhi	Kolkata	
Plant 1	24	30	40	25 Tons
Plant 2	30	40	42	35 Tons
	25 Tons	25 Tons	10 Tons	

Difficult to solve with 2 × 3 matrix
Real life situations 20000 × 30,000 matrix
Example.
Ambulances reaching patients example



which ambulance would cater to this patient

D.V. – How much Plant 1 and 2 supply to different cities (origin to destination)
Transportation optimization.

	1	2	3
1	x_{11}	x_{12}	x_{13}
2	x_{21}	x_{22}	x_{23}

x_{ij} : how many tons plant i would deliver to Mumbai

Constraints

Supply constraints
$$\begin{cases} x_{11} + x_{12} + x_{13} \leq 25 \\ x_{21} + x_{22} + x_{23} \leq 35 \end{cases}$$

Demand constraints
$$\begin{cases} x_{11} + x_{21} \leq 25 \\ x_{12} + x_{22} \leq 25 \\ x_{13} + x_{23} \leq 10 \end{cases}$$

Objective $- 24x_{11} + 30x_{12} + 40x_{13} + 30x_{21} + 40x_{22} + 42x_{23} \longrightarrow$ Minimize

If total supply is less than demand, all of it should be exhausted (supply constraint becomes = in equation)

If total demand is less than supply, all of it should be ~~exhausted~~ completed (demand constraints becomes = in equation)

Refer to excel sheet "Distribution Optimization"

$$\text{Min} \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$ c_{ij} - cost coefficient

$$\sum_{j=1}^{n} x_{ij} \leq S_i \quad i = 1, 2, \dots, m \quad \text{Supply}$$

$$\sum_{i=1}^{m} x_{ij} \leq d_j \quad j = 1, 2, \dots, n \quad \text{demand}$$

However both can't be $\leq$ at the same time either Supply will have = or demand will have =

Optimization Techniques are Applied Mathematics

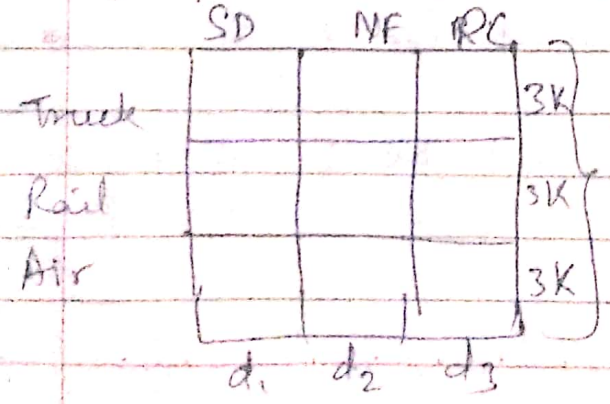
Transportation Problem

- Acme Block company has to complete orders of 80 tons concrete at 3 locations
- Northwood - 25 tons
- Westwood - 45 tons
- Eastwood - 10 tons
- 2 Plants of acme, each produces 50 tons per week (Supply)

- Navy has 9000 pounds of material in Albany (Supply)
- has to ship to 3 installations
  - Sandiego
  - Norfolk
  - Pensacola
- Their demand, 4000, 2500, 2500 pounds
- Govt regulation ~~dem~~ requires equal distribution of shipping among 3 carriers (truck, Railroad, airline)

shipping cost at diff locations by diff carriers / Pound

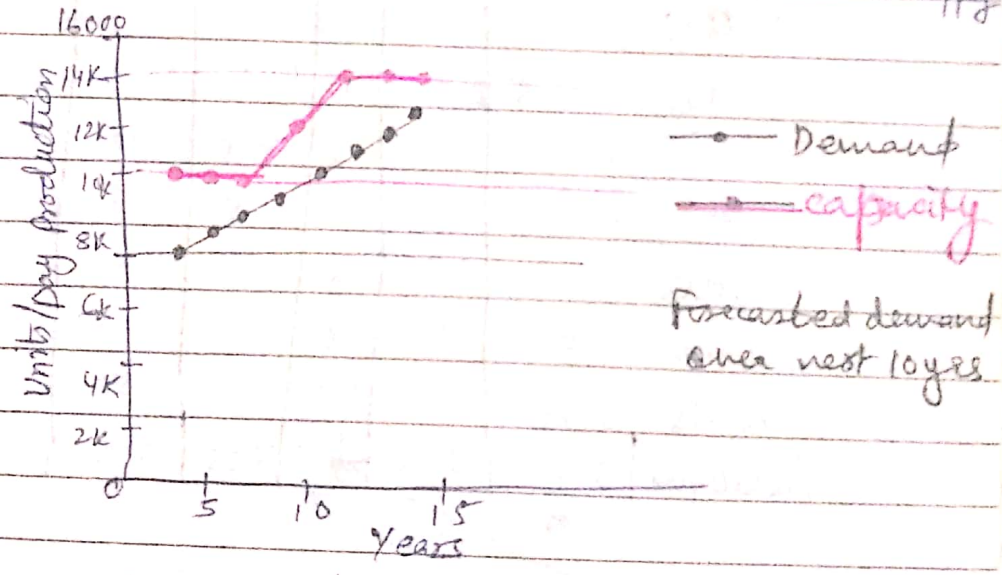
Mode	Destination: Sandiego	Norfolk	Pensacola
Truck	$12	$6	$5
Railroad	20	11	9
Airplane	30	26	28



INFORMS - Certified Analytics professional @ CAP
informs.org
HBSP - Harvard case studies

Designing Optimal Capacity strategy

- 4 manufacturing plants and 7 distribution centers for a paper manufacturing company.
- Case is about paper cartons
- For the new manufacturing facility, lead time to produce is 2 years
- Per cartons cost would be $10
- At present all 4 manuf plants are operating with lot of downtime, lot of maintenance and availability of plants is only 90%
- Want to create some extra cushion in supply



At present capacity is running at 10,000, even if we operate at 100% capacity, it is 10,000
Demand is growing from 8,000 to 13,000
In 10 years, if demand is going to raise, capacity (supply) of plant is also raising from 10 to 14. (Still more than demand)
So gives some cushion before a plant run out of supplies.

— So company has to decide, where a plant would manufacture, if it does and how much

Production costs and capacities

Facility	Production Cost per carton	daily capacity cartons/day
Toronto	$14	2,500
Denver	$19	1,500
Los Angeles	$13	3,500
Seattle	$17	2,500

Product distribution costs and 5/10 yr demand (Daily cartons)

Distribution centers	Toronto	KC	LA	Seattle	GJ	5 yrs	10 yrs
Toronto	0.75	2.50	4.50	4.25	5.25	1000	2000
KC	2.50	1.00	2.50	2.75	3.25	7500	1000
LA	4.50	2.50	0.50	2.25	1.75	2500	3000
Seattle	4.75	2.75	2.25	0.75	2.50	1500	2000
Chicago	1.50	1.50	3.75	2.50	3.75	1500	2000
Atlanta	3.00	2.25	3.00	3.50	3.50	750	1000
Guadalajara	5.25	3.25	1.75	3.75	0.50	2000	3000

Transportation costs from one DC to other last 2 columns, demand at different centers per day (5 yr/10 yr forecast per day)

Expanding plant's capacity is ruled out as already they are already going through maintenance and other issues.

How about setting a new plant at Guadalajara
- Cost is $30 million in 2 yrs timeframe

Question - Is it worth to go for a new plant spending 30 millions.

We evaluate investments on the basis of returns. How much more profits we will make with this extra plant.

You would like to open a manuf plant where your distribution center is also there.

- Find out total cost without additional cost
- Find out total cost with additional plant
- → Difference would decide whether new plant is worth it or not

If Toronto supplies to Toronto, cost would be

0.75 + $14 = $ 14.75

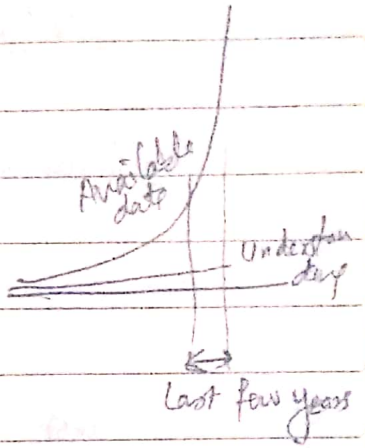
from denver to Toronto

19 + 0.75 = $19.75

per day

Tableau (Data Visualization)

No pie chart, No 3-D chart
↳ Issues ← Area of circle
Angle of pie



Gap b/w available data and its understanding

Visual Analytics

(1) Understand cognitive perceptions ← right choice of size, shape etc
- Sensory, reflex, Pre-attentive attributes

(2) Eliminating visual clutter - cramped, unnecessary, flashy
- Add colours for a meaning, data inc vs norm data inc

(3) Choose an appropriate display mechanism
- Right choice of charts
- Combination of dimensions & variables, categorical & continuous, drives the choice of dashboard

(4) Design dashboards - Not a dumpyard
- have proper scientific layout
- have right amount of interactivity.

(5) Explore visually -

(6) Analyze visually

(7) Create Storyboards

Static vs Dynamic Dashboards

Tableau Products

1) Tableau desktop
   - Tableau personnel (Local files)
   - Tableau Professional (Local files & Server based files)

Not an analytics/stats tool but connect to many like R. Creates reports based on structured data. Professional has connections to many data sources. Does not work on unstructured data (like images)

2) Tableau Server
   - Data kept in secure environments

~~Tableau files~~, when you share reports (interactive/live reports), 40-50 users of the report. Server would be installed on your private. Save reports, save databases, share reports etc. We can save files as an URL. Automated reports (getting updated automatically) from desktop, you have to hit refresh button.

(3) Tableau Online
   - Toned down version of server
   - On cloud.
   - cheaper but you don't own the server
   - Hosted tableau server

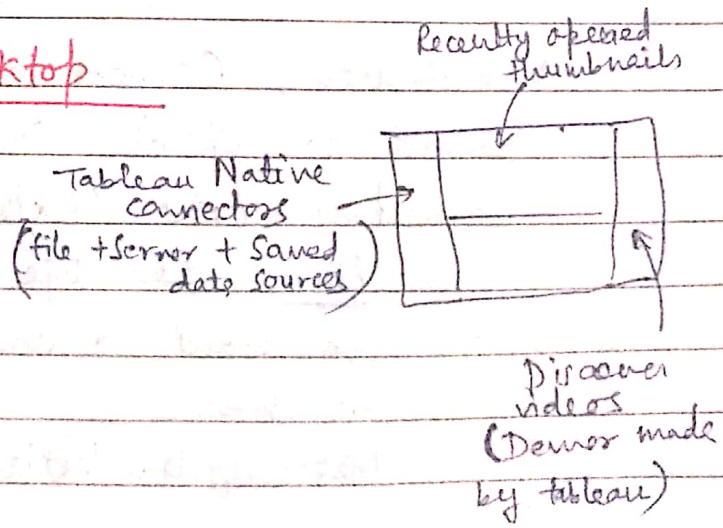
(4) Tableau Reader
   - like adobe reader. Reads Desktop files

- When you share Tb reports, other person should either have Tb desktop installed or Tb server or Tb reader atleast to read the file.
- If free, Person can interact with the data / filter / interactive feature.
- Can not change anything in the report.
- Checkpoint - If report is developed in Tb version 2018.2, ask the reader to install 2018.2 reader to avoid any compatibility issues.

(5) Tableau Public -

- free hosted service to publish reports.
- No security, Only done for marketing purposes
- You cant save any file on local m/c
- It would save on Tableau Public Server only
- Use for learning purpose

Tableau Desktop



Sample Superstore file

→ Documents → my tableau repository
→ datasources → 2018.2 → en-US-10
→ Sample - Superstore

This is what Sid needs (Order sheet)
Return sheet is also there
First row should have headers
Data type, maintain consistencies.
Imp For tableau to pick up a date, it should have all 3 components (day, month, year)
Order does not matter
Otherwise mimic the date by putting some value in missing component

Country, State, city - Geospatial fields
→ 3 letter code with these fields should also work like "IND"
→ East, west, North, South, not geospatial as we can assign lat long values.

→ In tableau, Connect to Sample - Superstore and connect to specific "Orders" sheet.
→ Connection Live / Extract
Live - Real life data to come in, when you want to commission a live dashboard on server
Makes system little slow but data is live

Extract - Not live but you can keep refreshing
Faster.

→ Filters can be added at 1st page itself.. Suppose I want to work on southern date only, set a filter here itself using 'Add'

→ Can also remove certain columns from Tableau Prep. More of ETL, does some cleansing and data preparing job. From above filter you can just filter rows.

Dimensions - Categorical, like Customer ID, dates, Row ID (If ROW ID is measures, move it to dimension)

Measures - measurable number, like discount, quantity, Sales etc

Check if something is not correctly defined as a measure or dimension, swap it

🌐	geo spatial field, postal codes
Abc	String / Categorical
📅	Date / Time fields
#	Number field / Measures or dimensions

Tableau has in-built lat-long value set. So if you put the spelling of geospatial fields correct, it would automatically classify those fields as 🌐. All Countries, tier 1, tier 2 cities are mapped.

Automatically generated fields

→ In Dimensions, there is a field called "Measure Names" — It is a combination of all measure fields

→ In Measures, latitude (generated) and longitude (generated)

→ Measure values also the same, contains all other measures

→ Number of records — say # records in country "IND"

→ Rows, Columns – fundamental for any chart
→ labels on each bar
– Color, Size, text, Detail etc are preattentive attributes.
– law of pragnance – Brain likes to see everything (objects in the most simple form
→ i.e why the like to see data in sorted form.
– Swapping – if labels are not wrapped / shown properly in columns, swap them in rows.
– Wrapping labels – too much work
embedding on bar – Not a good idea
abbreviations – will make more complicated
Colours – Diff. bars with diff colours, will add visual clutter.

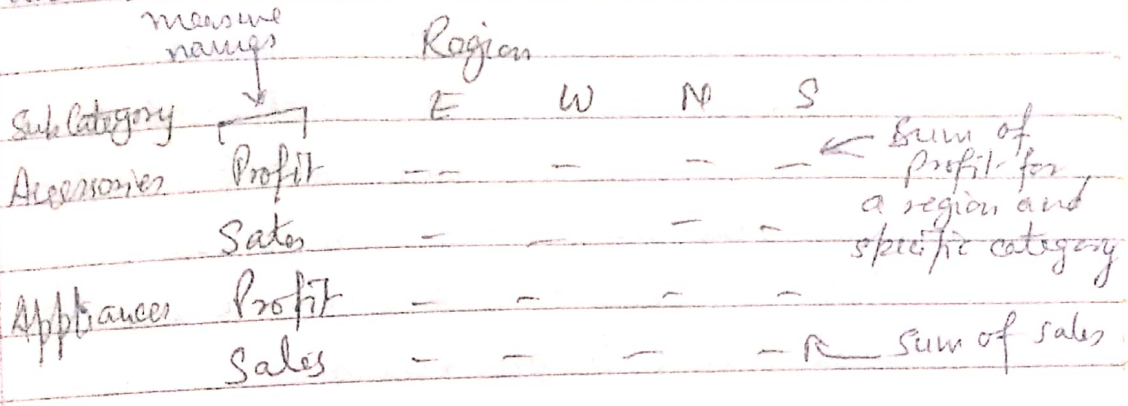
Category vs Sum(Sales)

→ No need to sort and swap when you are dealing with Time Series date / chronology
→ It has to be displayed the way it is (low/high as it is. no need to sort ascending or descending)
→ Want to see the irregularities there.

3 thumbnails in tableau
– New sheet
– New dashboard
– New Story

Column → Region
Rows → Sub-Category
Abc columns in table → Sales
Abc columns in above Sales table → Profit

It would give a table with Sub-category vs Region and Sales & Profit filled for each cell one below the other.



As there are 2 measures being used in central part of the report, "Measure Names" is added automatically & filters where out of all measure names, only Sales and profit are selected.

As we are using 2 measures here, tableau understands that ~~it~~ the intent is to do multivariate analysis, hence adds "measure Names" automatically in rows along with Sub-category.

Measure Values are added automatically in left hand side "Pragmatic attributes' space, with icon T as we want to make a pivot kind of table with Text values (sales and profit) in middle part of table.

Though Measure names and measure values stores all measures, but only those would be selected for graph here which we drag and drop in the middle portion for analysis.

To avoid scroll, moved measure names from rows to columns.

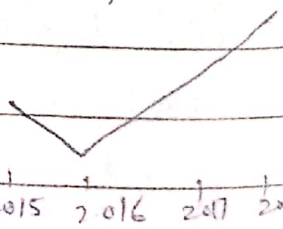
In previous chart, it is difficult to answer questions like in which region, which category has maximum sales but least profit or max sales with high profit etc

So instead of putting Sales and Profit in central area which is giving text/Pivot table, Sum of Sales in Columns and Sum of Profit in colour card. Here size of bar easily tells us about Sales performance and colour of bars tells profit performance.

Colour and size amplifies cognition (sensory impact)

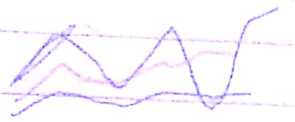
HLOD - Highest level of detail
LLOD — lowest level of detail
for ex order date, year, month and date gives lowest levels of details

Columns - ⊞ Year (order date)
Rows - Sum of Sales
This would give year of year performance of Sales
It shows Dipped in 2015 and then started increasing



expand year in columns once, it would give quarter on quarter sales, for every year
However, if we further expand quarters into months, it causes visual clutter
→ To fix scroll issue, Set Entire view

Another way to see Year on Year sales is to superimpose different year charts on one another for easy comparision



Drag Year(order date) to colour card
month(order date) to columns
and Sum of Sales as rows.
chart type as automatic
With superimposing plots, it is easy to find cyclic / seasonal patterns.

Select your dimension and measures using ctrl key and click "show me", it would show what chart choices are possible with those variables.

For ex, with State and Sales, we use Global map chart

for setting proper country maps, click on unknowns / map tab → edit locations and set country (either hardcoded for a specific country or use "From field" as country to include multiple countries from data set)

It would automatically ~~select~~ put state and country in "Detail" card, Sum(sales) in Colour Card and tableau inbuilt latitude/longitude in Columns and Rows (as we choose "map" chart)

It would give color legend also.

Pattern — Coastal and larger cities having maximum sales

However, to check which cities are contributing to max sales in Coastal states, we need to plot city vs sales with "Symbol map"

Because from the previous map, we know Coastal States have max sales, But need to confirm if there is any specific factor or large cities in US which leads to high sales

Sum(Sales) in ◎ Size card
State, Country, city in Detail card
Longitude, Latitude in Colums/rows
Earlier it was darker the colour more the sales, here, greater the size of dot, more the sales

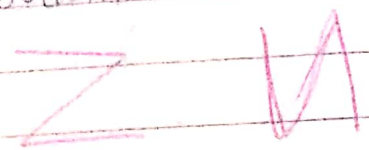
For Dashboards, best screen size is with no scrolls. So in size choose "Automatic". It would fit the way it is oriented.

However, important point to mention here is if you are the creator of dashboard (or created on laptop ~~or~~ in landscape mode) and someone opens it in laptop/tab (landscape mode), dashboard would work well. But if other person opens in mobile (Portrait mode) it would not work well. So we need to know our audiences and set dashboard size accordingly. Instead of automatic, we can choose Fixed size → custom → various sizes

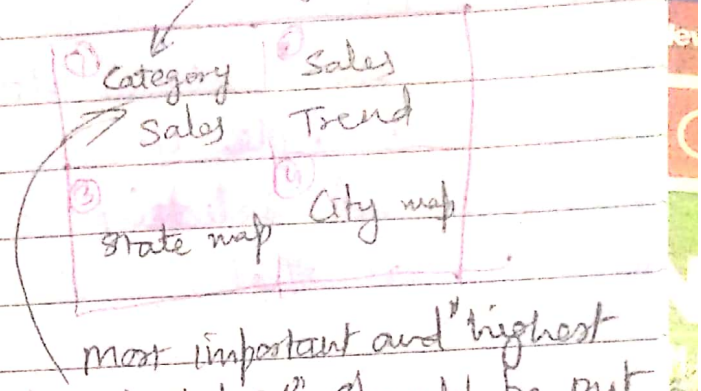
Dashboard Design Principles

Prime area (important) No logo, titles, legends, filters

Mostly for english people, left to right, top to bottom.



Z principle



most important and "highest level of detail" should be put here (high level detail, than subsequent levels in quadrants)

But for certain parts of world say middle east culture, it has to be revers Z i.e 

Brushing and Linking

Every single item in the dashboard acts like a slicer

Say if we choose category = "Technology" in first chart, other 3 charts are going to be filtered for technology.

If in chart 3, I choose state = "Michigan" other 3 charts are going to show info related to michigan only.

To do this choose  use as filter for all the 4 charts so all 4 charts are now working as filters and interacting with each other.

Say I select Technology in 1st chart, I can get clear insight which state is selling / not selling technology products

Use "Esc" to come back to non filtered charts

clear sheet (next to swap) to clear sheet in one go

So we can do all slicing/dicing, rolling up/down in same dashboard

Use 'Lasso' function to select/ cover multiple cities in an area ◯ instead of just selecting 1 and filtering charts based on that.

Story Telling

→ Engaging audience when you do things right
→ Telling boring things in a more structured and interactive manner

Hierarical fields and making custom hierarchy

→ Date fields have natural hierarchy
they can easily be grouped by months and years (intrinsic)

Custom
→ Country → State → cities are also hierarical
→ Category → sub category → Product ID/name

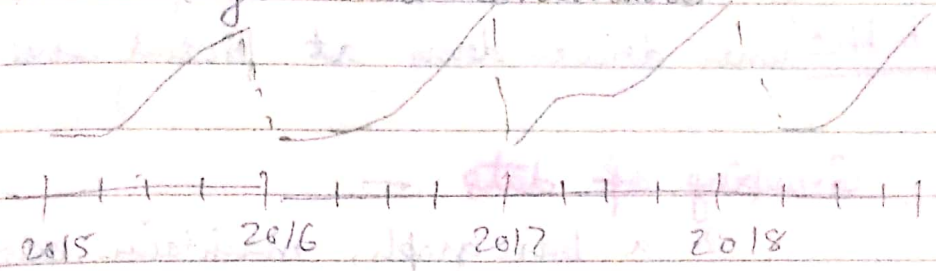
Discreet vs Continuous data

measures are generally continuous as we cannot bin them. We can create artificial bins to somehow fit continuous data like Sales into bins. Region (E, W, N, S) is having clear bins, so it is a dimension

usually
Dimensions – Discreet (Blue)
Measures – Continuous (Green)

Order date - dimension

So by default tableau take date as discreet field. So change it to continuous



Quarter discreet Q_1, Q_2, Q_3, Q_4

Continuous - Q1 2015, Q2 2015, Q3 2015, Q4 2015 followed by Q1 2016

Overall trend or forecasting → Continuous

Cyclic pattern → discreet

So date fields (hierarical) can be made to behave discreet or continuous.

Custom hierarchy

A simple test to check hierarchy and its order (Parent - child) is to place Parent dimension first in rows and then child dimension (Sub-category) next to parent in rows.

further place Product Name next to sub-category, The resulting table would show clear hierarchy if present.

Drag Sub-category over Category in dimensions. Tableau gives option to name this custom hierarchy for ex Product hierarchy.

Dimensions would be reorganized under a ∨ sign.

Confirm with business if you find some new hierarchy and not clear whether it is real, to check if newly developed hierarchy makes any sense or not

- Gives options to tell story in sequential manner
- Gives options of drill down at lower level easily

"Packed bubble" * Shared multiple challenges with bubble chart when drilled down at product level hierarchy
→ many pie charts in one, diff to be precise

Grouping of data —

In a bar graph, maintain existing category and create new groups.

Ctrl key + (select labels + fastners) + (Right click)
group members (Ctrl icon)

→ merges 2+ category into 1 (Rename, edit)

→ It creates new dimension Sub-category (group) as well

Data filters (filters on multiple conditions)

- Dimension filters
- Measures filters
- Date filters (as date can be discreet / continuous so separate filter)

Tabs in filter box (Dimension filter) - discreet
check box filter

- General
- Wildcard, contains "Asia"
- Condition, special conditions based on fields
- Top, top 10/ Bottom 10 etc

Show filter to show it in right side column

Tabs in filter box (Measures filter) - continuous
Slider filter

- All values 15 45
- Sum, Average, median, count, distinct, min, max
- std dev, variance, Attributes

Tabs in date filter

- Here, based on selection, filter would work as discreet/continuous and respective tabs would be shown
- For continuous, select "Range of Dates"

Date filter slider gives calender option as well to pick date ranges

Clicking will show calender

3/1/2018 ——□—— 30/12/2018

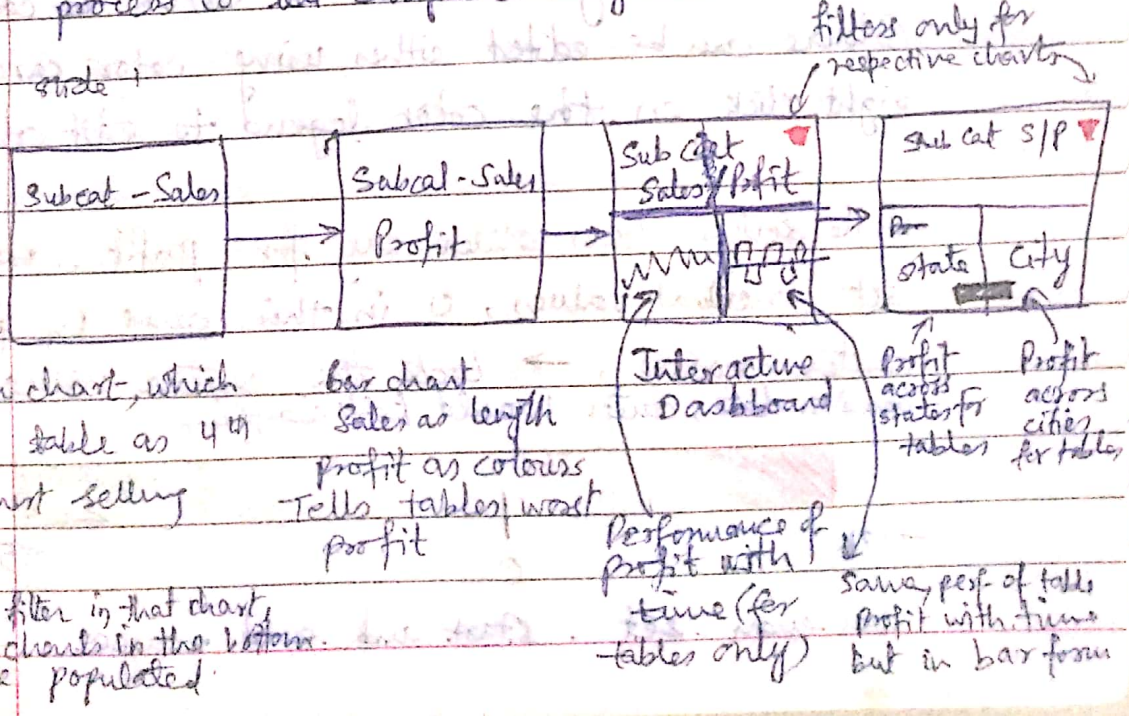
Make same filters applicable to all worksheets

Select filter → Right click → Apply to worksheets → All using this data source (other options based on requirements)

Story Board

Initial analysis shows tables 4th highest in terms of sales but in bottom line in terms of profit.

Will make a 4 stage process, an interactive process to tell complete story



Line vs Bar – Line shows trend

– Bar gives more precision over line

Combination of both line and bar in the same dashboard would help us telling the story

Need to bring interactivity as there is "Table performance for Profits"

As other filters are giving sum(profit) for tables, we want to use another filter in last 2 geospatial maps which where we can filter only profit making transactions (or even loss making transactions, Sum(Profit) adds both profit and loss transactions). This filter would help find out area which are really contributing to profit.

May require colour blindness friendly palatte (blue orange) instead of red-green

One in every 10 male in US/UK is colorblind

Colours can be edited either using color card or right click on the color legend to edit color palette

While setting colour slider say for profit, we can set neutral values, 0 in this case by fixing center value. → Lock the center value, otherwise based on data, center would keep changing

can even set start and end range

Line chart - Profit trend

Columns - Continuous, Month(order date)

Rows - Sum(Profit)

Color card - Sum(Profit)

Marks - Automatic (Bar for bar chart)

Show trend line - Right click trend chart, mark trend line

To implement ~~filter~~ filter b/w state and city maps, drag profit to filters and instead of Sum (selected by default), select "All values"

So instead of aggregating profit for cities (+ve plus -ve values), it would use individual transactions

Apply this filter to selected worksheet

Now for Dashboard (screen 3), Set top chart as filter "Only for that sheet". Dont apply it to other sheets as we want top 1 to control bottom 2. We do not want bottom sheets also to act as filters.

To see why chairs are performing very good in Seattle, go to tool tip on city map and insert newly created "Deep dive into cities" sheet. (line chart) — This would insert a chart in a sheet only. As you hover on various cities, it would Profit trends for that city

2 measures, need to find relationship b/w them

Scatter Plot

① Correlation would come out but most importantly it would show trend.

ex - -ve advertisement — decreasing sales
+ve adv — increasing sales

② Outliers can be spotted very easily
③ You can get an idea of where your clusters are forming.

1 measure, 1 dimension — bar chart (best)
if 1 dimension is date — line chart can be good
dot chart (good option)

Multiple data sources / files can be added in same session. They can interact or may not interact.
No size limitation, works on OS configuration

Calculated fields

Can be a calculated dimension (string)
or calculated measure (number)
or calculated date (dates)

Calculated fields remain in tableau session only
Analysis tab → create calculated field
String concat - [Category] + " - " + [sub-category]
Substring - To get first name, let we need to find the position of - W/o first name and last name

Once the position of space is found, we can use it to seperate first name.

findingspace = FIND([Customer Name], " ")

Though this new field finding space gives number, it should be a dimension as it is giving discreet value and not a continuous value.

Now to get 1st name,

Left([Customer Name], [Finding space] - 1)

To get last name, first find total length of customer name and the subtract the position of space to get # of characters in last name.

Then use Right function to get last name.

or use mid function

Null in dimensions - data quality issue, null, NA wrongly put strings etc

Null in measures - for these dimensions, no data, or we can say 0 values.

For some analysis, we may need to remove these null values and convert them to 0.

- Create Calculated field - "Corrected target$"

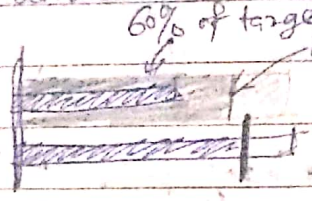
ZN([Targets])

returns expression if not null, otherwise returns zero. Now plot your corrected target instead of targets

Another application - survey results on a scale of 1 to 5, with avg value 3.5. So easy to find out which dept fared well and which metric ≥ 3.5, shown as green. for < 3.5 shown as red. used Parameters to find bottom 10 metrics to be worked upon

Bullet graphs

Useful in a situation when we have targets and we want to compare sales for specific dimension (say state, area, ——)



Swap reference line fields by clicking say horizontal axis and select swap.
This 60 : 80 : 100 % band for targets to achieve can be configured to other value
Also, instead of 2 bands, we can do 1 band, 5 band, no band, etc (various combinations)
x-axis → edit reference lines → 60% 80% of avg corrected targets
Play with Computational values.

To have colour coded bars where Sales is achieved vs where Sales is not achieved.
Calculated fields "Variance"
[Sales] - [Corrected targets]
Apply this to the color card.

Analysis based on days taken to ship the product in various regions, say the average time claimed by the ecommerce website is 3.5 days. Now analyse which areas are doing good & in terms of shipping time.

days to ship = Ship date - Order date.
Kind of date diff function of excel

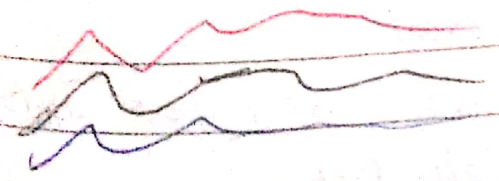
days to ship = datediff ('day', orderdate, shipdate)

VPs to distribute states to E, W, N, S manager based upon their portfolio size/load

Reclassification

→ One way is to go back to raw data and reclassify those records by mentioning Texas in South region instead of central
→ Use Case function in tableau
→ Create calculated field "Reclassified Regions"

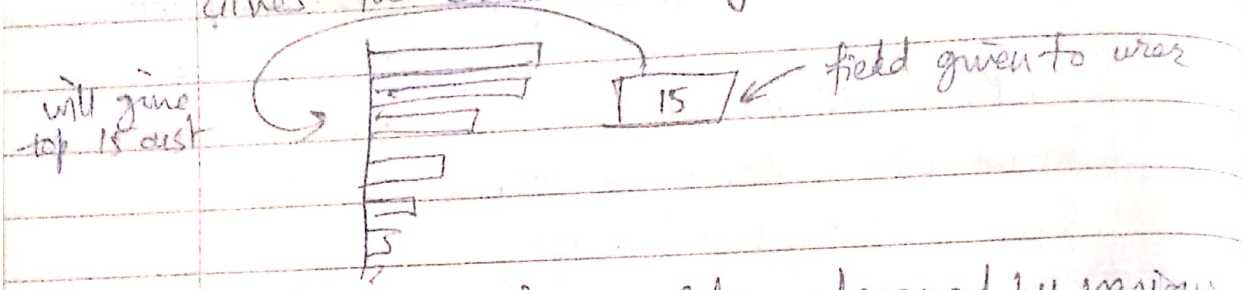
```
Case [state]
    When "Texas" then "South"
    When "Indiana" then "East"
else [Region]
END
```



← put year(date) in color card

Parameters

3 main parameters — used in filters, used in calculated fields, for dynamic labels, what if analysis. Parameters increase interactivity to many folds. Gives the control directly in the hand of user



Suppose we have sales observed by various customers and we make a simple bar chart with sorted values.

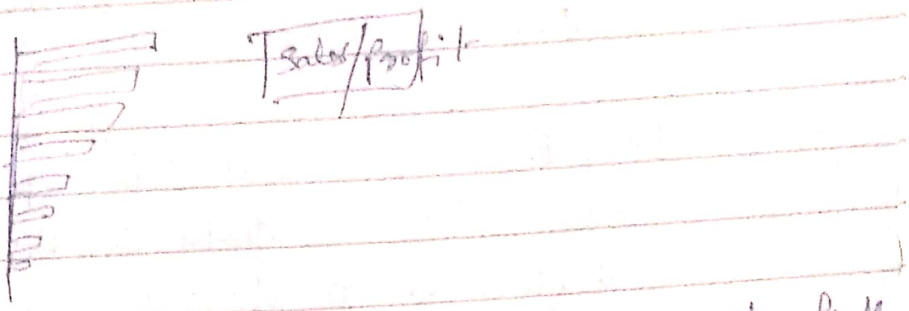
Now we want the user to be able to find top N / Bottom N customers, based on whatever number (N) he enters in the field

→ Create Customer Names vs sum(Sales)
→ Put it in Sorted order (793 customer)
→ I want to give these customers a S. No (index) for 1 to 793 showing above sorted orders
Kind of artificial index / sticky notes)
→ Create Calculated field Index
Index()
→ Move this new field to dimension
→ move it into row next to cust name
change it to discreet.
→ Use Filters, set cust name into filters
set options in Top tab
By field Top 10 by sum(sales)
→ Allow to create new Parameter to set to N instead of 10.

Could be used for bottom 10/(N) also. - least.
performance metric to work upon.

Parameter - Top Customers
Allowable values = All
→ This will allow to choose from 1 to 793,

ex2



Based on ~~region~~ the option chosen in parameter fields (ie sales/profit), Sales or profit for various ~~customers~~ sub categories is displayed
Taking 1 column and replacing it completely with other column. In filter values in one column to display selected rows (ie filters rows)

Rules for parameters -
(1) Create parameter
(2) Show parameter
(3) Create a calculated field using parameter
(4) Apply calculation on worksheet

'Create Parameter' gives options to select values as All, List, Range
Parameter name - Sales/Profit
Data type - String
Allowable values - List ~~(Add from fields)~~
add manually Sales / Profit
Show parameter by clicking parameter field, right click it and select show.

Create Calculated field "2 in 1"

Case [Sales/Profit] ← Purple as parameter
When "Sales" then [Sales]
When "Profit" then [Profit]
End

2 in 1

Now put this calculated field into columns
Same "2 in 1" field can be applied to all these charts where we were using only Sales or Profit

Now how to create "dynamic" labels changing along with changing parameter values;

for ex, labels Category Sales, Sales Trend, State Sales, City Sales etc should be changed to Category Profit, Profit Trend, State Profit and City Profit if "Profit" is selected in the Parameters. (label for various screens/graphs in the dashboard)

Double click on label, instead of <Sheet name> select category [Sales/Profit Parameter]

You cannot customize axis name through (2 in 1)

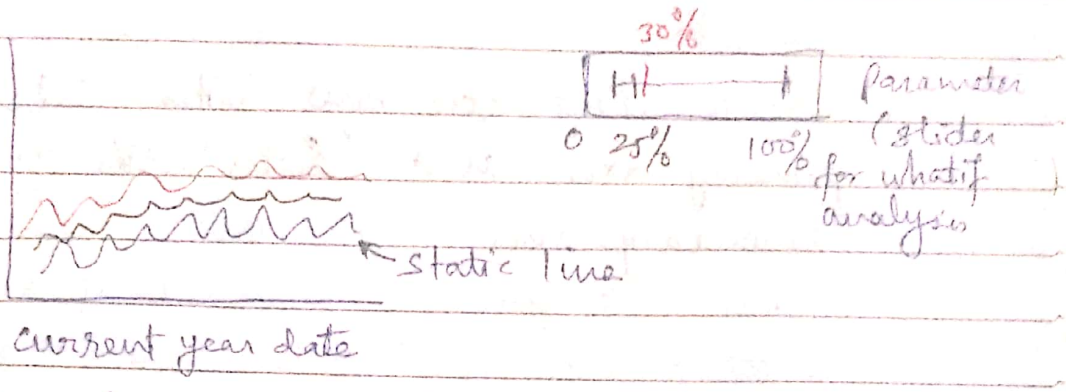
What if analysis

Current profit X

What if market grows by Y% , so next year goal is X+Y

what if analysis to deal with projection

Thresh



Other lines would change based on the parameter selected. Along with what if sales, need a chart/table which shows

	Total
Date	
Sales	
What if Sales	

Columns - Order date

Rows - Sum (Sales)

Filters - Year (Order date) = 2018

Create parameter as "What if Sales"

Choose range of values (Float Parameter) within 0 to 1 or min 0% max 100%

step size can be set (1% increase, 5%/10% etc)

Display format = Percentage instead of automatic

Create Calculated field "Sales with growth factor"

Sum([Sales]) * (1 + [What if Sales])
↑ parameter

Put this aggregated (sales with growth) also above next to the rows (It would give superimposed agg sales)

Create dual axis now with 'Sales' as primary axis and 'Sales with growth' as secondary axis

Now lock primary & secondary axis by selecting secondary axis and say "Synchronize" axis

Now remove "Show header" for secondary axis.

Mimic this "What if Sales" chart as a pivot table by selecting "Duplicate as crosstab"

This Pivot table would also change with slider

Analysis tab - show totals (row totals) will give grand totals

Dynamic Sort

2 type of sort

- Static - Sort using top ribbon, axis, chart header
- Dynamic - based on some conditions, sub-category (dimension) would come on top

Click on sub-category, right click → Sort Based on field "e in one" it would sort the date.

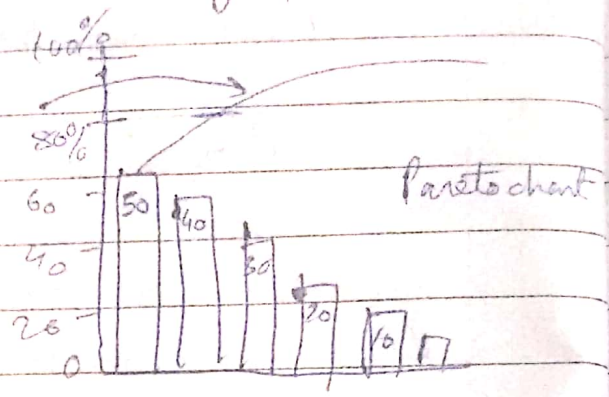
24 default charts in Tableau
More can be created like "Pareto"

Pareto

Columns — Year (order date)
Rows — Month (order date)
Sum(Sales) in text area
Quick table calculation to check growth / degrowth ⇒
Click sum(sales) in text area, "Quick table Calculations" → Percent difference
Sum(Sales) has a Δ icon next to it which tells some table calculation is done there (local only for that table. No other calculated field is created)
Complete using "Previous". Change to "table down" for month on month improvement.
Change Compute using "Table (down then across" to have continuity from dec (last year) to Jan (this year). Profit of Jan 2016 over Dec 2015
Move Sum(sales)Δ into colour card also to show heat map table along with numbers.

Change Fiscal Year as per country

Go to any date field in 'Dimensions' → Default Properties → Fiscal Year Start (say April for India, January for US)



Bars have to be sorted from high to low

$\frac{50}{150}$ = 33% defects

$\frac{40}{150}$ = 26% defects

Rows - Sum(Sales)
Columns - Sub-Category

Sort in descending order (mandatory for pareto)
First Part of Pareto is done

To have the Cumulative line now,
Create another bar chart in similar fashion
Rows - Sum(Sales) Sum(Sales)
Columns - Sub-Category
change it to "Quick Table calculation" → Need Running Total %, So start with running total first (excel has direct function running total % but in tableau, we have to first get running total and then %)
Go back to 2nd Sum(Sales) and Edit Calculation
Use "Add Secondary Calculation"

there as we can't change anything in "Running Total" calculation. We have to do something over and above it.

choose "Percentage of total"

Now we need to make this Cumulative % line.

Go to 2nd sum(sales), select dual axis, it would superimpose 2 charts on one another.

Now go to 1st sum(sales), Mark type → change it to bar

go to 2nd sum(sales), Mark type → change it to line.

Now go to "Measure Names" legend, edit colours, change the colour of % of total running Sum of ~~Sales~~ sales.

To apply Selective labels (Only on line for % and not on bars) → go to label and in Marks select sum(sales) Δ, go to label, Show label.

Word Clouds

Applications of word cloud,

→ Social media
→ Text mining
→ Sentiment analysis
→ Advertisement

Need 1 dimension and 1 measure

for ex, dimensions as actual words, measure as

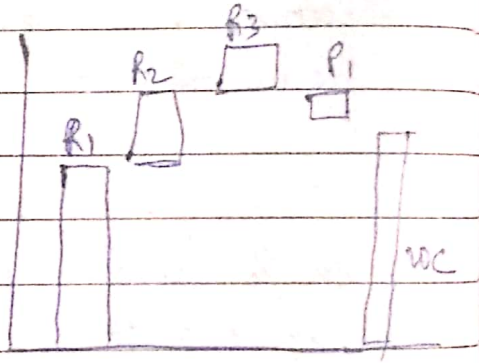
length / number of times word is liked. Whichever word appears more times is in bigger size

Cannot make word cloud for measure with -ve value (say -ve profits)

Text box - Sub category
Size - # of records
marks card - Text
Colour card - Profit
↳ Here -ve values are shown via colour card.

Waterfall charts

Performance KPIs
Revenues analysis
Throughput analysis
Working capital
Stock worked



Project Progress (SDLC) - not a waterfall chart instead, it is a Gantt chart.

Eg - Recievables - Payables = net working cpt
$(R_1 + R_2 + R_3) - P_1 = WC$

Items coming into system — Items going out of the system

→ waterfall charts helps identifying which items are coming into the system and which are

going out of the system.
→ It also tells about the magnitude of items coming in and going out of the system.

Columns – Sub-Category
Rows – sum(Profit)
Quick table calculation, Running total
Mark Type – Gantt bar

Create Calculated field "Waterfall"
– [Profit]
Drop this waterfall measure into size card
→ Tab "Analysis" → Show "Row grand totals
Colour card – Profit
It would show colour coding which items are +ve and which are –ve

Modifying Tool tip

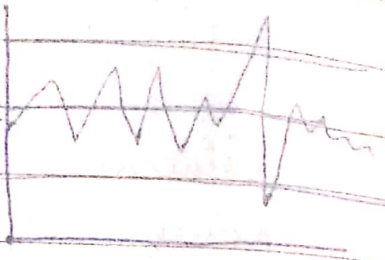
Click on card "Tooltip"
As of <subcategory> the cumulative profits are <running sum of sum(Profit)>
Customize other sentences.

Control chart

Looks an observations during a time period

ex. golf ball with radius of 5 cm

marginal difference (±x) is within threshold (with control limit). Once it goes beyond control limits, product becomes unacceptable.



Generally, in manuf, this kind of quality is done everyday / for every batch

rows - ~~Sum~~ Sum(Sales)

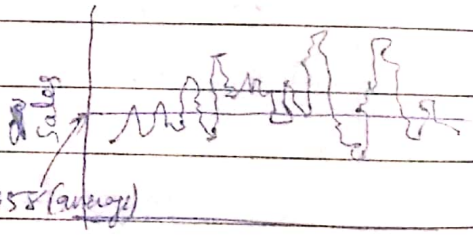
Columns - (Order date) month

for central lines (Reference line), we can take average of all values and create a reference line

Right click Sales axis →

add reference line

change line → average of Sum(Sales)



Change label by customizing it (for central line)

For upper/lower control limit lines, choose "Distribution" ~~line~~ option. Average option given in 60%, 80% of average which can be changed for UCL/LCL

~~choose~~ stddev -1, 1 for UCL/LCL here

To make std-dev levels dynamic, we can create stddev parameters and use it instead

of hard coded -1, 1, 3. It would allow us to do "What if" analysis based on changing σ.

Create calculated field "average"
we need avg (sum ([Sales])) but because of tool limitation, we can't aggregate on an aggregate
Workaround is
window - avg (sum ([Sales]))

2nd calc. UCL
[Average] + window - stddev (sum([Sales]))

3rd calc LCL
[Average] - window - stddev (sum([Sales]))

To apply newly created fields "average", "LCL", "UCL", we need to pull them into chart memory first, instead of keeping them in measures/dimensions.
So pull these 3 fields into "Detail" card
Now add reference line, line, choose average field. Label custom avg <value>
Same for other 2.
Now add parameter for what if analysis
- Create parameter "Control limit"
choose range, min 0.5, max 3, stepsize 0.1 (~~min~~ this is range of sigma)
- Goto LCL field, edit the formula
[Average] + window_stddev (sum([Sales])) * [Control limit]
here if control limit is 1, average + 1σ
else average + σ

Drag another sum(sales) to rows. It would create 2 charts. Go to 2nd sum(sales), change mark type as "Circle"

Apply colour to the 2nd chart, add dual axis and merge 2 charts

Add Calculated field 'KPI'

if sum([sales]) > [UCL] or sum([sales]) < [LCL]

then "Out of control"

else "In control"

End

Add KPI for sum(sales) 0 to colour.

Go to 2nd sum(sales) in rows, select dual axis.

Connecting to google sheets

Coffee dashboard in tableau folders (google link)

→ Connect to a server → more → google sheets

→ Discussed about the need of choosing correct chart

Polygon maps

Instead of predefined states, cities on a map, we can create custom polygon maps for any specific place (say parks) on an image using combination of lat-long values

However to do such mapping you need data with lat-long values of that place to connect the dots.

Park Name	Point ID	Polygon	Latitude	Longitude
Dunkeswar National Park	1	5	50.618	-3.63
Dunkeswar "	2	5	50.616	-3.64

Columns → Avg(Longitude)
Rows → Avg(Latitude)
Colour card → Park names
Change mark type to Polygon
~~Path ID~~ → Point ID to → Path card ✓
(open for polygon maps only)

Adding more data points to the actual data can help many folds with visualisation
for example, no. of trees for each park is given in dataset or pollutants scale/quantity
⇒ So we can colour code polygon maps ~~saying~~ saying darker green showing more # of trees, lighter green means less # of trees in that park.
⇒ Similarly heavily pollutant lakes as black (more # of pollutants)

Route Maps

hub and spoke system — b/w A and B

— To display density, of say flights b/w 2 points, say Delhi & Mumbai, we can show more dense routes with thick lines or darker colour lines

Country Name	City	Path ID	Path order	Flow Amount	Latitude	Longitude
Germany	Berlin	Berlin-Lyon	1		52.52	13.38
France	Lyon	Berlin-Lyon	2	400	45.76	4.84

Here all flight going out of Berlin so Berlin is the Hub and all other cities are Spoke

Copy all data from excel directly and paste into tableau sheet.

Now data is stored in clipboard and you can delete the data from tableau sheet.

Columns — avg(Longitude)
Rows — avg(Latitude)
Details card — Path ID
Mark type — Line
Path card — Path order
Colour card — Sum(Flow amount)
Size card — Sum(Flow amount)