What is data mining?

- nontrivial extraction of implicit, previously unknown and potentially useful information from data.
- exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

What is not data mining?

- (deductive) query processing, we want is statistical algorithm running over the data.

Data mining application

1. Database analysis and decision support
2. Market analysis
   - how market is segmented
   - how the customers fill in the profile you address with the product.
   - determine purchasing patterns.
   - how different products are related.
3. corporate analysis and risk management.
   - what the trends are.
   - monitor competitors and market directions.

Input Data →	Data Preprocessing →	Data mining →	Postprocessing →	Information
• flat files, spreadsheet, relational tables • centralized or distributed	• raw input data to appropriate format for subsequent analysis • data cleaning • feature selection, dimensionality reduction, normalization, data subsetting	• closing the loop. • Integrating data mining results with decision support system.	• only valid and useful results are incorporated • explore from different view • filtering patterns, visualization	

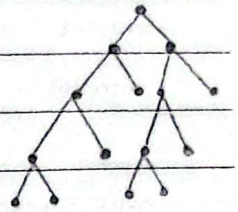
Methods of Data mining

1. Predictive
   - predict values of an attribute based on values of other attributes
   - attribute to be predicted is known as target, while the other one is explanatary or independent.
2. Descriptive
   - derive patterns that summarize relationships in data.

day / date:

1. Predictive Modelling

→ building a model for the target variable as a function of independent variable.



(a) classification

→ discrete target variables

→ predicting whether a web-user will make a purchase online is binary-valued.

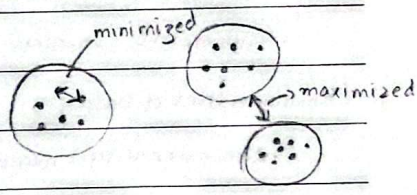
(b) regression

→ continuous target variable.

→ forecasting future price of stock because price is continuous valued.

2. Cluster Analysis

→ groups of closely related observation so that observations that belong to the same cluster are more similar to each other than observations of other cluster.



→ document clustering, similar articles grouped on a topic

→ K-means

3. Association Analysis.

→ discover patterns that describe strongly associated features in the data.

→ represented in form of implication rules or feature subsets.

→ market-basket analysis

4. Anomaly Detection

→ identifying observations whose characteristics are significantly different from the rest of the data.

→ credit card fraud detection.

→ good anomaly detector must have high detection rate and a low false alarm rate.

What is data?

- collection of data objects and their attributes
- an attribute is a property or characteristic of an object
- attribute values are numbers or symbols assigned to an attribute
- height is attribute, feet or meters are values.

Types of Attributes

- nominal : used to categorise/classify objects, groups, individuals. e.g. gender, country, religion.
- ordinal : data is categorized and ranked, magnitude of rank unknown e.g. grades
- interval : categorize, rank and infer equal intervals between neighbouring data points
  e.g. temperature, calendar dates → mean, SD
- ratio : categorize, rank, infer equal intervals and there is a true zero point.
  e.g. weight, income, age, length, mass
- discrete : has only finite set of values, integers, binary attributes e.g: ID number, zipcode.
- continuous has real numbers as attribute value, floating point variables e.g: temperature, height
- symmetric : equal treatment of both values e.g. has a pet → Yes or No
- asymmetric : unequal treatment e.g. presence, non-zero attribute value is important

Characteristics of Data

- dimensionality : number of attributes
- sparsity : those with asymmetric feature, attributes of an object have values of 0, presence of 1 counts
- resolution : low resolution data is historical gathered on hourly, monthly, yearly basis
  high resolution data is collected at intervals of one-minute
- size : type of analysis depends on size

Types of Dataset

1. Record 2. Graph 3. Ordered.

1. Record

→ collection of records, each of which consists a fixed set of attributes.

→ data matrix : each dimension represents a distinct attribute. m objects, n columns for each attribute.

→ document data : each document becomes a 'term' vector, each term is a attribute and value is number of times the corresponding term occurs, non-zero entries are stored.

→ transaction data : each transaction has set of items.

2. Graph

→ relationship among data objects are links, and data objects are nodes.

3. Ordered Data

→ sequential : each item purchased along with time.

→ spatio temporal data.

→

day / date:

Data Quality

- detection and correction of data quality problems
- data cleaning
- problems
  - noise, outliers, fake data, wrong data, missing values

1. Noise
   - refers to modification of original values.
   - techniques used to reduce noise to discover patterns/signal
2. Outliers
   - data objects that are considerably different than most of the other data.
   - anomalies
   - outliers may sometimes be the goal of our analysis. e.g. credit card fraud.

Data Preprocessing

1. Aggregation
   - combining two or more objects in a single object. to reduce number of attributes.
   - cities aggregated into regions, days to weeks or months.
   - quantitative attributes e.g. price aggregated by sum, mean.
   - have less variability.
2. Sampling
   - selecting a subset of data objects to be analyzed.
   - this is done because 'processing' entire set can be expensive.
   - a sample is representative if it has approximately same properties as the original set of data.

(a) simple random
- equal probability of selecting any particular item.
- without replacement: as each item is selected, it is removed from population.
- with replacement: objects are not removed after selecting.

(b) stratified
- split data into several partition and draw randomly.

day / date:

3. Discretization

→ converts a continuous attribute into ordinal

→ infinite number of values are mapped into small number of categories

→ used in supervised and unsupervised.

4. Binarization

→ converts discrete (continuous) attribute into 1 or more binary attribute.

5. Attribute Transformation

→ maps the entire set of values to a new set

→ normalization:

→ standardization.

6. Dimensionality Reduction

→ avoid curse of dimensionality

→ reduce time, memory

→ helps in eliminating irrelevant feature.

→ use of PCA: creating new attributes that are combination of old one.

→ captures largest amount of variation.

7. Feature Subset Selection

→ redundant features duplicate much or all of the information. e.g: price of product, sales tax paid.

→ Irrelevant features: information is not useful for data mining.

8. Feature Creation

→ create new attributes that capture important information more efficiently than originals.

→ feature extraction: new features from raw data.

→ feature construction:

→ mapping data to new space

original value -

day / date:

DM functionalities

- Association

→ one thing dependent on other.

→ age(x, "20..29") , income(x, "20...29K") → buys (x, "PC")

People aged 20-29 with income 20-29K tends to buy PCs.

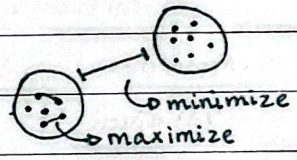
- Classification and Prediction

→ developing models like decision trees or neural networks for future predictions

- Cluster Analysis

→ class label is unknown

→ maximize intra-class similarity and minimize interclass similarity



- Outlier Analysis

→ a data object that is oad one out

→ detecting rare or fraudulent patterns in data.

Association Rule Mining

- finding all co-occurrence relationships among data items.
- market basket analysis → discover how items are purchased.

X → Y (if X is bought, then Y is likely to be bought)

- let $I = \{i_1, i_2, \ldots i_m\}$ be a set of items.
- let $T = \{t_1, t_2, \ldots t_m\}$. be a set of transactions

$X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$

someone who bought item X, also buys item Y e.g. Bread → Butter but not Bread → Bread

- Rules can be weak or strong
- strength of rule is measured by its support and confidence.
- Support → probability that X and Y appear together in transaction.
- Confidence → probability of Y appearing given X is present.
- Lift → dependency between X and Y $Lift(X \Rightarrow Y) = Conf(X \Rightarrow Y) / Supp(Y)$

Lift > 1 , strong association, appear together more than expected.

day / date:

- if support is too low → rules may appear by chance
- if confidence is low → X does not reliably predict Y
- discover all associated rules in T that have support and confidence greater than a minimum threshold (minsup, mincof)

- Approach minsup = 30% minconf = 80%

chicken, clothes → milk = $\frac{\text{chicken} \cap \text{clothes} \cap \text{milk}}{\text{no. of transaction}}$

sup = $3/7$

confidence $\frac{\text{chicken} \cap \text{clothes} \cap \text{milk}}{\text{chicken} \cap \text{clothes} \cap} = \frac{3}{3} = 1$

Transactions

T1 : beef, chicken, milk
T2 : beef, cheese
T3 : cheese, boots
T4 : beef, chicken, cheese
T5 : beef, chicken, clothes, milk, cheese
T6 : clothes, chicken, milk
T7 : chicken, milk, clothes

- Approaches
  - (a) apriori
  - (b) multiple minimum supports
  - (c) mining class association rule

Apriori Algorithm

Step 1 : find all frequent itemsets (support ≥ minsup)

Step 2 : use frequent itemsets to generate candidate rules

STEP 1 :

	TRANSACTIONS
T1	beef, chicken, milk
T2	beef, cheese
T3	cheese, boots
T4	beef, chicken, cheese
T5	beef, chicken, clothes, cheese, milk
T6	clothes, chicken, milk
T7	chicken, milk, clothes

1. find frequent itemsets of size 1 → F_1
2. generalization, K ≫ 2

C_K = those itemsets of size K that could be frequent.

F_K = those itemsets that are actually frequent

- Downward closure property → any subset of a frequent itemset is also a frequent itemset.

example → {chicken, clothes, Milk} = $3/7$ (m) minsup = 30%

$\frac{3}{7}$ = (chicken, clothes) (chicken, Milk) = $\frac{4}{7}$ (clothes, Milk) = $\frac{3}{7}$

chicken, clothes, milk
$\frac{5}{7}$ $\frac{3}{7}$ $\frac{4}{7}$

- items are sorted in lexicographic order

day / date:

· Generalization

step 1 : joining (A_{k-1}, B_{k-1})

example → $F_3 = \{\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{1,3,5\}, \{2,3,4\}\}$

1,2,3	→	1,2,4	→	1,2,3,4	1,2,4	→	1,3,4
		1,3,4					1,3,5
		1,3,5					2,3,4
		2,3,4					

1,3,4	→	1,3,5	→	1,3,4,5	1,3,5	→	2,3,4
		2,3,4					

. only the last element should be different and greater than original set.

. in the above example there ave 2 new candidates which could be frequent

step 2 : pruning → remove those candidates in C_k that do not respect the downward closure property

1,2,3,4		1,2,3 ✓		1,3,4,5		1,3,4 ✓	
		1,2,4 ✓				1,3,5 ✓	
		1,3,4 ✓				1,4,5 ×	
		2,3,4 ✓				3,4,5 ×	

$C_4 = \{(1,2,3,4), (1,3,4,5)\}$

all of these exist in F_{k-1} where $k=4$ so downward closure property exist.

remove from C_4

after pruning $C_4 = [1,2,3,4]$

day / date:

Multiple Minimum Support

- each item has its own minimum support
- some items appear frequently in the data, while others are rare.
- if minsup is too high, no rare items will be found
- to prevent very frequent items and very rare items from appearing in the same itemsets, we introduce a support difference constraint (Φ)

$$\max\{sup(i) - \min\{(sup(i))\}\} \leq \Phi \qquad 0 \leq \Phi \leq 1 \text{ is user specified.}$$

- MIS(i) is the minimum item support
- The minsup of a rule R is the lowest MIS value of all items.

Rule R satisfies its minimum support if;

actual support $\geq \min(MIS(i_1), MIS(i_2), \ldots)$

- downward closure property doesn't exist anymore.

example → MIS(1) = 10% MIS(2) = 20% MIS(3) = 5% MIS(4) = 6%

{1,2} has a support of 9% → $9 \geq \min(10\%, 20\%)$ → not frequent

{1,2,3} has a support of 7% → $7 \geq \min(10, 20, 5)$ → frequent

* applying downward closure would eliminate (1,2) without evaluating (1,2,3) which is frequent

→ How to solve downward closure property?

1. sort all items in I according to their MIS values, starting with smallest

→ Algorithm

Step 1: frequent itemset generation

(a) seeds for generating candidate itemsets.

(b) candidate generation for k = 2

(c) generalization for k > 2

Step 2: rule generation.

day / date:

STEP 1: frequent itemset generation

example → I = [1,2,3,4] MIS(1) = 10·/. MIS(2) = 20·/. MIS(3) = 5·/. MIS(4) = 6·/. n = 100 transaction

(a) · sort according to MIS → [3,4,1,2]

· count → [3]:6, [4]:3, [1]:9, [2]:25

· seeds of list L: { }

→ MIS(3) = 5·/.

item 3 → 6·/. ≥ 5·/. → so L = [3]

item 4 → 3 ≥ 5·/. → so L = [3]

item 1 → 9 ≥ 5·/. → so L = [3,1]

item 2 → 25 ≥ 5·/. → so L = [3,1,2]

· calculate F1 now if any item in L has actual support < MIS(i)

item 3 → 6 < 5 ✓

item 1 → 9 < 10 ×

item 2 → 25 < 20 ✓

Items	1	2	3	4
MIS	10	20	5	6
SUP	9	25	6	3
L	{3,1,2}			

· so frequent itemset = F_1 = [3,2]

(b) candidate generation, k = 2, Φ = 10·/.

★ use L and not F_1 due to downward closure property.

· test chosen item against its MIS: sup(3) ≥ MIS(3)

(a) if true, form level 2 candidates, otherwise go to next element

L = {3,1,2} → [3,1], [3,2]

[3,1] is a candidate → sup[1] ≥ MIS(3) AND |sup(3) − sup(1)| ≤ Φ

9 ≥ 5 AND |6 − 9| ≤ 10, thus C_2 = [3,1]

[3,2] → sup[2] ≥ MIS(3) AND |sup(3) − sup(2)| ≤ Φ

25 ≥ 6 AND |6 − 25| ≤ Φ, thus not true.

· test: sup(1) ≥ MIS(1) = 9 ≥ 10 = not true so cannot proceed further. so C_2 = [3,1]

· calculate support of each item in C_2 → sup(3,1) = 6 ≥ min[5,10] so F_2 = [3,1]

day / date:

Mining Sequence Patterns

- frequently occurring ordered events
- customer retention, targeted marketing
- an event is an itemset, unordered list of items, brought in a single transaction

$(I_1 I_1 I_3)$

- sequence S is an ordered list of events

$\langle e_1 e_2 e_3 \dots \rangle$

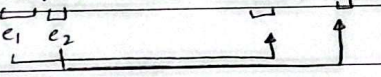
- example;

$e_1 = (abc)$, $e_2 = (ade)$

$s = \langle e_1 e_2 \rangle = \langle (abc)(ade) \rangle$ → length of $s = 6$

- α is a subsequence of β because;

$\alpha = \langle (ab)d \rangle$ and $\beta = \langle (abc)(de) \rangle$ ∴ $\alpha \sqsubseteq \beta$



- support of a sequence
  - ↳ number of tuples in S, containing α
- α is a frequent sequence if $sup_S(\alpha) \geq min\text{-}sup$.
- a sequence pattern is also a frequent sequence.

EXAMPLE

SID	SEQUENCE
1	$\langle a(abc)(ac)d(cf) \rangle$
2	$\langle (ad)c(bc)(ae) \rangle$
3	$\langle (ef)(ab)(df)cb \rangle$
4	$\langle eg(af)cbc \rangle$

$I = [a, b, c, d, e, f, g]$

min-sup = 2

∴ for SID = 1, length = 9, but support of a remains 1 in first transaction even though it has occurred thrice, so total support = 4

∴ $\langle a(bc)af \rangle$ is a subsequence of SID = 1 $\quad \langle a(bc)df \rangle$ / $\langle a(abc)(ac)d(cf) \rangle$

- huge number of sequential patterns are hidden in databases

day / date:

- a priori based methods.
  - → generalized sequential pattern. (based on downward closure property)

Generalized sequential pattern.

step 1 : calculate support of all length 1 sequence and discard those which are less than min sup.

step 2 : generate length 2 candidates from length 1 pattern (ordered, unordered table)
1 event, 2 event

- drawbacks
  - ↳ huge set of candidate sequence generated
  - ↳ multiple scan of database
  - ↳ long patterns grow from short patterns.

- combining sequences of events with repeated measurements, we obtain time series data.

Time - Series Data

- reveal temporal behaviour
- events changing with time.
- data recorded at regular intervals.
- financial, industry, meteorological
- goals
  - → modeling time-series
    - analyze the past data.
  - → forecasting time-series
    - use the model to predict future data.
- methods
  - → trend analysis
  - → similarity search.

Trend Analysis

- statistical techniques. e.g. regression analysis.
- construct a model to explain the behavior of the measurement. e.g. correlation.
- regression analysis
  - ↳ finding trends and outliers in datasets.
  - ↳ dependent and independent variable.
- example
  - → determine appropriate levels of advertising for a particular market segment

day / date:

- linear regression analysis
  - kitna difference on sales.

$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$

Y = predicted score

b_0 = intercept / origin

b_i = regression coefficients

- correlation (R)
  - → denoted by pearson
  - → refers to interdependence.
  - → closeness of linear relationship between X and Y.
  - → between -1 and 1 (positive

error = $y - \hat{y}$

$\hat{Y} = \theta_0 + \theta_1 X_1$

$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$ actual - predicted

$\theta_0 = \bar{y} - \theta_1 \bar{x}$

- regression trend channels
  - → upper and lower trendline from standard deviation of regression line.
  - → when the prices break a well established trend channel.
  - → represented by 3 lines, where middle one is linear regression line and then +/- standard deviation.
- non-linear regression
  - → bayesian methods
- characteristics time-series movement components.
  1. trend → long term progression / upward downward
  2. seasonal → identical patterns that a time series appears to follow. (christmas sales)
  3. cycle → regular fluctuations (business cycle)
  4. irregular → random
- decomposition
  - → additive modal = $T + C + S + I$
  - → multiplicative = $T \times C \times S \times I$

day / date:

Trend Analysis (T), methods

1. freehand method
   - → fit the curve by looking at the graph.
   - → costly
2. least - square method
   - → curve minimizing the sum of the squares.
3. moving average method
   - → eliminates cyclic, seasonal and irregular patterns.
   - → loss of ends as it is a sliding window.
   - → sensitive outliers.
   - → a smoother curve
   - → influence of extreme values can be reduced with weighted moving average (WMA)

Original	WMA(3) → (1 4 1)
3	
7	
2	$(3\times1 + 4\times7 + 1\times2) / 1+4+1$
0	$(7\times1 + 2\times4 + 0\times1) / 1+4+1$
4	
5	
9	
7	
2	

   - → cumulative moving average $CA_i = \frac{x_1 + \dots + x_1}{i} \rightarrow CA_{i+1} = CA_i + \frac{x_{i+1} - CA_i}{i+1}$
   - → exponential weighted moving average.

Estimation of seasonal variation (S)

1. seasonal index
2. deseasonalized
   - → adjusted with seasonal variations.

* happens regularly at fixed points within a year

Estimation of cyclic variation (C)

- → semantic meaning

* happens over irregular time frames, tied to long term economic factor.

Estimation of irregular variation (S)

- → adjusting the data.

- make long or short term predictions (time-series forecasting)

day / date:

- time series forecasting
  - → Auto regressive integrated moving average
  - → non-stationary.
- granularity change
- moving average for trend analysis
  - ↳ identify open position for buying and selling.
- bollinger bands → change of trends.
- resistance lines → market stops at one point.
- momentum analysis.
  - ↳ difference between opening and closing position.

Similarity Search

- exact matches in normal database
- SS finds data sequence that differ only slightly
- financial market, market basket

day / date:

Classification

- collection of records
- each record consists of a set of attributes
  - x : attribute, independent, input
  - y : class, response, dependent, output
- goal is to assign new records to each class without asking
- defining a class attribute as a function of characteristics that you know
- training on a known data and then testing to validate
- financial field, news
- training set → learn model → apply model → deduce → test set → classification.
  - test set: unknown class

Supervised learning

- training data comes with class labels
- new data classified based on training set

Unsupervised learning

- no labels in training set
- clustering

Decision Trees : Base classifiers

internal node : test on an attribute

branch : represents outcome of test

leafnodes: class labels

STEP 1 : INDUCTION

- input : training dataset, attribute list
- output : decision tree

STEP 2 : DEDUCTION

- predict classes of new data
- input : decision tree, new data
- output : classified new data

day / date:

DECISION TREE INDUCTION

1. HUNT'S ALGORITHM

D_t → set of training records (data) that reach a node t

same class label, do not split.

if any data record has different class label, then split.

- split the data in classes on an attribute test
  - nominal (discrete values)
  - ordinal (categories)
  - continuous (numbers) → range, buckets.
  - binary split
  - multi-way split
- If D_t contains records that belong to same class y then t is a leafnode labelled as y_t.
- If D_t contains records that belong to more than one class, choose next best attribute for test condition and recursively apply it.
- If D_t is not an empty set, but no other attributes remain for splitting then t is a leaf nod labelled by label of majority records.

- splitting continuous variable
  - discretization (static, dynamic)
  - binary decision (interval, right or left)
- to determine the best split
  - check ration where it is greatest
  - nodes with homogenous class distribution preferred

example: student → Yes: Yes: 6, no: 1 ✓

income → high, Yes: 2, No: 2 ✗

Methods to measure impurity

1. Information Gain

ratio of classes of some label opposed to classes of other labels

Yes: 6 No: 1 → high

Yes: 2 No: 2 → low

2. Gini Index

- continuous valued, several splits

day / date:

1. Information Gain

- assume there are 2 classes, P and N
- p elements of class P, n elements of class N

$$I(p,n) = \frac{-p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

probability that label is P — probability that label is N

- information gain in decision tree

to partition into sets

$$E(A) = \sum_{i=1} \frac{p_i+n_i}{p+n} I(p_i,n_i)$$ sum of entropy of all subsets

to encoding information

$$Gain(A) = I(p,n) - E(A)$$

- bigger entropy, worst split

EXAMPLE – STEP1

class P : buys computes = yes = 9

class N : buys computer = No = 5

1. $I(p,n) = I(9,5) = \frac{-9}{9+5}\log_2\frac{9}{9+5} - \frac{5}{9+5}\log_2\frac{5}{9+5} = 0.94$

2. compute entropy on basis of age

age	p_i	n_i	$I(p_i,n_i)$
<= 30	2	3	0.971
> 40	4	0	0
31...40	3	2	0.971

$$= \sum \frac{2+3}{14} I(2,3) + \frac{4+0}{14} I(4,0) + \frac{3+2}{14} I(3,2) = 0.694$$

3. calculate gain(age) = $I(9,5) - E(age) = 0.94 - 0.694 = 0.246$

→ higher, important, purer group, less entropy, more information gain.

gain(age) > income, student, credit

4. now remove age and make further splits on income, student, credit.

keep which has higher gain.

day / date:

STOP CONDITION

- when all records belong to same class
- or when similar attribute values, a leaf node labelled as majority class

STEP 2: DEDUCTION

- read from root node and reach leafnode.

- one rule created for each path from root to leaf.
- attribute-value pair forms a conjunction.

Advantages

- inexpensive to construct
- fast in classifying unknown records.
- easy to interpret for small sized trees.

→ avoid overfitting in classification.

↳ prepruning → do not split if IG falling below a threshold

↳ post pruning → remove branches from fully grown trees.

day / date:

Mining Sequence Patterns

- frequently occurring ordered events
- customer retention, targeted marketing
- an event is an itemset, unordered list of items, brought in a single transaction.

  $(I_2 I_1 I_3)$

- sequence S is an ordered list of events

  $\langle e_1 e_2 e_3 \dots \rangle$

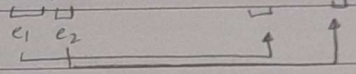
- example;

  $e_1 = (abc)$, $e_2 = (ade)$

  $S = \langle e_1 e_2 \rangle = \langle (abc)(ade) \rangle$ → length of $S = 6$

- α is a subsequence of β because;

  $\alpha = \langle (ab)d \rangle$ and $\beta = \langle (abc)(de) \rangle$ ∴ $\alpha \sqsubseteq \beta$



- support of a sequence

  ↳ number of tuples in S, containing α

- α is a frequent sequence if $sup_S(\alpha) \geq min\_sup$.
- a sequence pattern is also a frequent sequence.

EXAMPLE

SID	SEQUENCE
1	$\langle a(abc)(ac)d(cf) \rangle$
2	$\langle (ad)c(bc)(ae) \rangle$
3	$\langle (ef)(ab)(df)cb \rangle$
4	$\langle eg(af)cbc \rangle$

$I = [a, b, c, d, e, f, g]$

$min\_sup = 2$

∴ for SID = 1, length = 9, but support of a remains 1 in first transaction even though it has occurred thrice, so total support = 4

∴ $\langle a(bc)af \rangle$ is a subsequence of SID = 1 $\quad \langle a'(abc)(ac)d(cf) \rangle$

- huge number of sequential patterns are hidden in databases

day / date:

- a priori based methods
  - → generalized sequential pattern (based on downward closure property)

Generalized sequential pattern

step 1 : calculate support of all length 1 sequence and discard those which are less than min sup.

step 2 : generate length 2 candidates from length 1 pattern (ordered, unordered table)
1 event, 2 event

- drawbacks
  - ↳ huge set of candidate sequence generated
  - ↳ multiple scan of database
  - ↳ long patterns grow from short patterns.

- combining sequences of events with repeated measurements, we obtain time series data.

Time-Series Data

- reveal temporal behaviour
- events changing with time
- data recorded at regular intervals
- financial, industry, meteorological
- goals
  - → modeling time-series
    - analyze the past data.
  - → forecasting time-series
    - use the model to predict future data.
- methods
  - → trend analysis
  - → similarity search.

Trend Analysis

- statistical techniques e.g. regression analysis
- construct a model to explain the behavior of the measurement e.g. correlation.
- regression analysis
  - ↳ finding trends and outliers in datasets
  - ↳ dependent and independent variable.
- example
  - → determine appropriate levels of advertising for a particular market segment

day / date:

- linear regression analysis

$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$

Y = predicted score

b_0 = intercept / origin

b_i = regression coefficients

- correlation

→ denoted by pearson

→ refers to interdependence

→ closeness of linear relationship between X and Y.

→ between -1 and 1
(positive

error = $y - \hat{y}$

$\hat{Y} = \theta_0 + \theta_1 X_1$

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

$\theta_0 = \bar{y} - \theta_1 \bar{x}$

- regression trend channels

→ upper and lower trendline from standard deviation of regression line

→ when the prices break a well established trend channel.

→ represented by 3 lines, where middle one is linear regression line and then +/- standard deviation

- non-linear regression

→ bayesian methods

- characteristics time-series movement components.

1. trend → long term progression
2. seasonal → identical patterns that a time series appears to follow. (christmas sales)
3. cycle → regular fluctuations (business cycle)
4. irregular → random

- decomposition

→ additive modal = $T + C + S + I$

→ multiplicative = $T \times C \times S \times I$

day / date:

Trend Analysis (T), methods

1. freehand method
   - → fit the curve by looking at the graph.
   - → costly
2. least - square method
   - → curve minimizing the sum of the squares.
3. moving average method
   - → eliminates cyclic, seasonal and irregular patterns
   - → loss of ends as it is a sliding window
   - → sensitive outliers
   - → a smoother curve
   - → influence of extreme values can be reduced with weighted moving average (WMA)

Original	WMA(3) → (1 4 1)
3	
7	
2	$(3\times1 + 4\times7 + 1\times2) / 1+4+1$
0	$(7\times1 + 2\times4 + 0\times1) / 1+4+1$
4	
5	
9	
7	
2	

   - → cumulative moving average $CA_i = \frac{x_1 + \dots + x_1}{i}$, $CA_{i+1} = CA_i + \frac{x_{i+1} - CA_i}{i+1}$
   - → exponential weighted moving average.

Estimation of seasonal variation (S)

1. seasonal index
2. deseasonalized
   - → adjusted with seasonal variations.

* happens regularly at fixed points within a year

Estimation of cyclic variation (C)

- → semantic meaning

* happens over irregular time frames, tied to long term economic factor.

Estimation of irregular variation (I)

- → adjusting the data.

- make long or short term predictions (time-series forecasting)

day / date:

- time series forecasting
  - → Auto regressive integrated moving average
  - → non-stationary.
- granularity change
- moving average for trend analysis
  - ↳ identify open position for buying and selling.
- bollinger bands
- resistance lines
- momentum analysis
  - ↳ difference between opening and closing position.

Similarity search

- exact matches in normal database
- SS finds data sequence that differ only slightly
- financial market, market basket

day / date:

Classification

- collection of records
- each record consists of a set of attributes
  - x : attribute, independent, input
  - y : class, response, dependent, output
- goal is to assign new records to each class without asking
- defining a class attribute as a function of characteristics that you know.
- training on a known data and then testing to validate.
- financial field, news
- training set → learn model → apply model → deduce → test set → classification.
  (test set: unknown class)

Supervised learning

- training data comes with class labels
- new data classified based on training set

Unsupervised learning

- no labels in training set
- clustering

Decision Trees : Base classifiers

internal node : teston an attribute

branch : represents outcome of test

leafnodes: class labels.

STEP 1 : INDUCTION

- input : training dataset, attribute list
- output : decision tree.

STEP 2 : DEDUCTION

- predict classes of new data.
- input : decision tree, new data
- output : classified new data.

day / date:

DECISION TREE INDUCTION

1. HUNT'S ALGORITHM

D_t → set of training records (data) that reach a node t

- split the data in classes on an attribute test
  - nominal (discrete values)
  - ordinal (categories)
  - continuous (numbers)
  - binary split
  - multi-way split
- If D_t contains records that belong to same class y then t is a leafnode labelled as y_t.
- If D_t contains records that belong to more than one class, choose next best attribute for test condition and recursively apply it.
- If D_t is not an empty set, but no other attributes remain for splitting then t is a leaf nod labelled by label of majority records.

- splitting continuous variable
  - discretization (static, dynamic)
  - binary decision (interval, right or left)
- to determine the best split
  - check vation where it is greatest
  - nodes with homogenous class distribution preferred

example: student → Yes: Yes: 6, no: 1 ✓

income → high, Yes: 2, No: 2 x

Methods to measure impurity

1. Information Gain

ratio of classes of some label opposed to classes of other labels

Yes: 6 No: 1 → high

Yes: 2 No: 2 → low

2. Gini Index

- continuous valued, several splits

day / date:

1. Information Gain

- assume there are 2 classes, P and N

- p elements of class P , n elements of class N

$$I(p,n) = \frac{-p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

probability that label is p — probability that label is N

- information gain in decision tree

↳ partition into sets

$$E(A) = \sum_{i=1} \frac{p_i+n_i}{p+n} I(p_i,n_i)$$ sum of entropy of all subsets

↳ encoding information

$$Gain(A) = I(p,n) - E(A)$$

- bigger entropy, worst split

EXAMPLE – STEP1

class P : buys computes = yes = 9

class N : buys computer = No = 5

1. $I(p,n) = I(9,5) = \frac{-9}{9+5}\log_2\frac{9}{9+5} - \frac{5}{9+5}\log_2\frac{5}{9+5} = 0.94$

2. compute entropy on basis of age

age	p_i	n_i	$I(p_i,n_i)$
<= 30	2	3	0.971
>40	4	0	0
31...40	3	2	0.971

$$= \sum \frac{2+3}{14} I(2,3) + \frac{4+0}{14} I(4,0) + \frac{3+2}{14} I(3,2) = 0.694$$

3. calculate gain(age) = $I(9,5) - E(age) = 0.94 - 0.694 = 0.246$

→higher important, purer group, less entropy, more information gain.

gain(age) > income, student, credit

4. now remove age and make further splits on income, student, credit.

keep which has higher gain.

day / date:

STOP CONDITION

- when all records belong to same class
- or when similar attribute values, a leaf node labelled as majority class

STEP 2: DEDUCTION

- read from root node and reach leafnode.

- one rule created for each path from root to leaf.
- attribute-value pair forms a conjunction.

Advantages

- inexpensive to construct
- fast in classifying unknown records
- easy to interpret for small sized trees

→ avoid overfitting in classification.

↳ prepruning → do not split if IG falling below a threshold

↳ post pruning → remove branches from fully grown trees

SEQUENCE PATTERN MINING

SID	SEQUENCE
1	<milk (bread, butter) (eggs) diapers>
2	<(milk, eggs) (bread, butter) diapers>
3	<(cereal, milk) juice (bread, butter) eggs>
4	<(eggs, juice) (bread, butter) diapers>
5	<(cereal, milk) eggs (juice, butter)>
6	<juice (milk, bread) (butter, eggs) diapers>
7	<(cereal, juice) (bread, eggs) (butter) milk>

STEP 1 : CREATE 1-LENGTH SEQUENCE

<milk> = a = 6 , <bread> = b = 6 , <butter> = c = 7 , <eggs> = d = 7

<diapers> = e = 4 , <cereal> = f = 3 , <juice> = g = 5

STEP 2 : 1-LENGTH PATTERN

<a> = 6 <b> = 6 <c> = 7 <d> = 7 <e> = 4 <g> = 5 <f> = 3

- discard none as all over > min-sup

STEP 3 : 2-LENGTH SEQUENCE

	<a>	<b>	<c>	<d>	<e>	<f>	<g>
<a>	aa^0	ab^3	ac^5	ad^4	ae^3	af^0	ag^2
<b>	ba^1	bb^0	bc^2	bd^3	be^4	bf^0	bg^0
<c>	ca^1	cb^0	cc^0	cd^2	ce^4	cf^0	cg^0
<d>	da^1	db^2	dc^4	dd^0	de^1	df^0	dg^1
<e>	ea^0	eb^0	ec^0	ed^0	ee^0	ef^0	eg^0
<f>	fa^1	fb^2	fc^3	fd^3	fe^0	ff^0	fg^2
<g>	ga^2	gb^1	gc^4	gd^3	ge^2	gf^0	gg^0

	<a>	<b>	<c>	<d>	<e>	<f>	<g>
<a>		$(ab)^1$	$(ac)^0$	$(ad)^1$	$(ae)^0$	$(af)^2$	$(ag)^0$
<b>			$(bc)^0$	$(bd)^1$	$(be)^0$	$(bf)^0$	$(bg)^0$
<c>				$(cd)^1$	$(ce)^0$	$(cf)^0$	$(cg)^1$
<d>					$(de)^0$	$(df)^0$	$(dg)^1$
<e>						$(ef)^0$	$(eg)^0$
<f>							$(fg)^1$
<g>							

<ab> <ac> <ad> <ae>

<bd> <be>

<ce>

<dc> <de>

<fc> <fd>

<gb> <gc> <gd>

<(bc)>

<a>

<b>

<a>

<bc>

~~1. max speedup~~

Naive Bayes

age	income	student	credit rating	buys comp.
<=30	H	No	fair	No
<=30	H	No	excellent	No
31..40	H	No	fair	Yes
>40	M	No	fair	Yes
>40	L	Yes	fair	Yes
>40	L	Yes	excellent	No
31..40	L	Yes	excellent	Yes
<=30	M	No	fair	No
<=30	L	Yes	fair	Yes
>40	M	Yes	fair	Yes
<=30	M	Yes	excellent	Yes
31..40	M	No	excellent	Yes
31..40	H	Yes	fair	Yes
>40	M	No	excellent	No

1. calculate probability for target column.

positive (p) = buys_computer = Yes = 9/14

negative (n) = buys_computer = No = 4/14

2. calculate probability for "age" column, for each interval corresponding to Yes and No in target column.

age	yes	no
<=30	2/9	3/5
31..40	4/9	0/5
>40	3/9	2/5

3. calculate for remaining columns: income, student, credit rating

income	Yes	no
H	2/9	2/5
M	4/9	2/5
L	3/9	1/5

student	Yes	NO
Yes	6/9	1/5
NO	3/9	4/5

credit rating	Yes	No
fair	6/9	2/5
excellent	3/9	3/5

4. classify according to given instructions

X = Age : youth
Income : low
Student : yes
credit : fair

- seperate for Yes and No

$[P(X|p) * P(p)] \mid P(X)$ → positive

$= \frac{2}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14}$

$= 0.02111$

$[P(X|n) * P(n)] \mid P(X)$ → negative

$= \frac{3}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{4}{14}$

$= 0.0034$

5. conclude and classify

since value of positive > value of negative, so classify it as "buys_computer".

ID	Items Purchased
1	Milk, Bread, Butter, Eggs
2	Milk, Bread, Butter
3	Milk, Eggs, Bread
4	Bread, Butter
5	Milk, Cereal, Bread
6	Jam, Eggs
7	Milk, Butter, Eggs
8	Milk, Bread, Butter
9	Jam, Cereal
10	Milk, Cereal

Item	MinSupport	Sup
(1) Milk	60·/·	7/10
(2) Bread	50·/·	6/10
(3) Eggs	50·/·	4/10
(4) Butter	40·/·	5/10
(5) Jam	25·/·	2/10
(6) cereal	10·/·	3/10

minsupport = 40%
minconf = 50%
ϕ = 10·/·

· sort according to MIS

[6,5,4,3,2,1]

→ seeds of list L : { } → MIS (6) = 10·/·

[1:7], [2:6], [3:4], [4:5],

[5:2], [6:3]

→ check if sup (j) ≥ MIS (6)

item 6 → 3 ≥ 10·/· → so L = [6]

item 5 → 2 ≥ 10·/· → so L = [6,5]

item 4 → 5 ≥ 10·/· → so L = [6,5,4]

item 3 → 4 ≥ 10·/· → so L = [6,5,4,3]

item 2 → 6 ≥ 10·/· → so L = [6,5,4,3,2]

item 1 → 7 ≥ 10·/· → so L = [6,5,4,3,2,1]

→ calculating F_1 from L based on MIS of each item in L

item 6 → 3 ≥ 10 → ✓

item 5 → 2 ≥ 25 → X

item 4 → 5 ≥ 40 → ✓

item 3 → 4 ≥ 50 → X

item 2 → 6 ≥ 50 → ✓

item 1 → 7 ≥ 60 → ✓

F_1 = [6,4,2,1]

→ continuing L1 [6,5,4,3,2,1]

[6,5] is a candidate → sup (5) ≥ MIS (6) AND |sup(6) - sup(5)| ≤ ϕ

$\frac{2}{10} \geq \frac{10}{100}$ AND $|3-2| \leq \frac{10}{100}$, true

[6,4] is a candidate → sup(4) ≥ MIS AND |sup (6) - sup(4)| ≤ ϕ

$\frac{5}{10} \geq \frac{10}{100}$ AND $|3-5| \leq \frac{10}{100}$, false

[6,3] → sup(3) ≥ MIS(6) AND |sup(6) - sup(3)| ≤ ϕ

4 ≥ 10 AND $|3-4| \leq \frac{10}{100}$, true

[6,2] → 6 ≥ 10 AND $|3-6| \leq \frac{10}{100}$, false

[6,1] → 7 ≥ 10 AND $|3-7| \leq \frac{10}{100}$, false

sup(j)

· discard 5

· [4,3] → sup(3) ≥ MIS(4) AND |sup(4) - sup(3)| ≤ ϕ

4 ≥ 40 AND |5 - 4| ≤ 0.1, true

[4,2] → sup(2) ≥ MIS(4) AND |sup(4) - sup(2)| ≤ ϕ

6 ≥ 40 AND |5 - 6| ≤ ϕ, true

[4,1] → sup(1) ≥ MIS(4) AND |sup(4) - sup

7 ≥ 40 AND |5 - 7| ≤ ϕ, fa

· discard 3

· [2,1] → sup(1) ≥ MIS(2) AND |sup(2) - sup(1) ≤ ϕ

7 ≥ 50 AND |6 - 7| ≤ 0.1, true.

C_2 = [6,5], [6,3], [4,3], [4,2], [2,1]

→ calculate support

[6,5]: 1, [6,3]: 0, [4,3]: 2, [4,2]: 4

[2,1]: 5

→ for F_2, check sup(i,j) ≥ min(MIS(i), MIS(j))

· [6,5] = 1 ≥ min(10/25) valid

[6,3] = 0 ≥ min(10,50) invalid

[4,3] = 2 ≥ min(40,50) invalid

[4,2] = 4 ≥ min(40,50) valid

[2,1] = 5 ≥ min(50,60) valid

thus F_2 = [6,5], [4,2], [2,1]

ID	Items Purchased
1	(1)Milk, Bread(2), Butter(3)
2	Bread, Butter, Eggs(4)
3	Milk, Bread, Butter, Eggs
4	Milk, Bread
5	Bread, Butter
6	Milk, Bread, Butter, Eggs
7	Bread, Eggs
8	Milk, Eggs
9	Milk, Bread, Butter
10	Milk, Bread, Eggs

minimum support = 0.3 (30%)

minimum confidence = 0.7 (70%)

1 → [Milk : 7], [Bread : 8], [Butter : 6], [Eggs : 6],

all of them are frequent

all items

→ join, make pairs.

C2 = [Milk, Bread], [Milk, Butter], [Milk, Eggs], [Bread, Butter], [Bread, Eggs], [Butter, Eggs]

→ pruning

all items in C2 belong to F1

2 → C2 = [Milk, Bread : 6], [Milk, Butter : 4], [Milk, Eggs : 4], [Bread, Butter : 6], [Bread, Eggs : 5], [Butter, Eggs : 3]

F2 = all items

→ join, make triples

C3 = [Milk, Bread] → [Milk, Butter] → [Milk, Bread, Butter]
[Milk, Eggs]
[Bread, Butter] → [Milk, Bread, Eggs]
[Bread, Eggs]
[Butter, Eggs]

[Milk, Butter] → [Milk, Eggs] → [Milk, Butter, Eggs]
[Bread, Butter]
[Bread, Eggs]
[Butter, Eggs]

[Milk, Eggs] → [Bread, Butter] → none
[Bread, Eggs]
[Butter, Eggs]

[Bread, Butter] → [Bread, Eggs] → [Bread, Butter, Eggs]
[Butter, Eggs]

[Bread, Eggs] → [Butter, Eggs]

→ pruning

C3 = [Milk, Bread, Butter], [Milk, Bread, Eggs], [Milk, Butter, Eggs], [Bread, Butter, Eggs]

all items in C3 belong to F2

3 → C3 = [Milk, Bread, Butter : 4],
[Milk, Bread, Eggs : 3]
[Milk, Butter, Eggs : 2]
[Bread, Butter, Eggs : 3]

F3 = [Milk, Bread, Butter : 4],
[Milk, Bread, Eggs : 3],
[Bread, Butter, Eggs : 3]

→ join

C4 = [Milk, Bread, Butter] → [Milk, Bread, Eggs]
[Bread, Butter, Eggs]

[Milk, Bread, Eggs] → [Bread, Butter, Eggs] → none

→ pruning

C4 = [Milk, Bread, Butter, Eggs]

all items in C4 belong to F3

4 → C4 = [Milk, Bread, Butter, Eggs : 2]

F4 = none items.

DECISION TREE

OUTLOOK	TEMP	HUMIDITY	WINDY	PLAY GOLF
Rainy	Hot	High	False	No
Rainy	Hot	H	True	No
Overcast	Hot	H	F	Yes
Sunny	Mild	H	F	Y
Sunny	Cool	Normal	F	Y
Sunny	Cool	N	T	N
Overcast	Cool	N	T	Y
Rainy	Mild	H	F	N
Rainy	Cool	N	F	Y
Sunny	Mild	N	F	Y
Rainy	Mild	N	T	Y
Overcast	Mild	H	T	Y
Overcast	Hot	N	F	Y
Sunny.	Mild	H	T	N

1. calculate information gain for target class

classP = play golf = 9

class = doesnot = 5

$$I(9,5) = \frac{-(9)}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.94$$

2. compute entropy for each column and check which column has highest gain, split on that column.

	Pi	ni	I(Pin)
R	2	3	0.97
O	4	0	0
S	3	2	0.97

$$= \sum \frac{2+3}{14} I(2,3) + \frac{4+0}{14} I(4,0) + \frac{3+2}{14} I(3,2)$$

$$= \frac{5}{14}(0.97) + \frac{4}{14}(0) + \frac{5}{14}(0.97) = 0.694$$

gain(outlook) = 0.94 - 0.694 = 0.246

	Pi	ni	I(Pin)
H	2	2	1
M	4	2	0.918
C	3	1	0.81

$$= \sum \frac{4}{14}(1) + \frac{6}{14}(0.918) + \frac{4}{14}(0.81) = 0.918$$

gain(temp) = 0.94 - 0.918 = 0.022

	Pi	ni	I(Pin)
H	3	4	0.985
N	6	1	0.591

$$= \sum \frac{7}{14}(0.985) + \frac{7}{14}(0.591) = 0.788$$

gain(humidity) = 0.94 - 0.788 = 0.152

	Pi	ni	I(Pin)
F	6	2	0.811
T	3	3	1

$$= \sum \frac{8}{14}(0.811) + \frac{6}{14}(1) = 0.891$$

gain(windy) = 0.94 - 0.891 = 0.048

3. outlook has highest gain so split on that.

OUTLOOK

RAINY | SUNNY | OVERCAST

T	H	W	
Hot	High	False	N
Hot	High	True	N
Mild	High	False	N
Cool	Norm	False	Y
Mild	Norm	True	Y

$I(2,3) = 0.97$

for rainy ⇒

	Pi	ni	I(pi,ni)
H	0	2	0
M	1	1	1
C	1	0	0

$= \sum \frac{2}{5}(1) = 0.4$

gain (temp) = 0.97 - 0.4 = 0.57

	Pi	ni	I(pi,ni)
H	0	3	0
N	2	0	0

✓ gain(~~temp~~ humidity) = 0.97 - 0 = 0.97

	Pi	ni	I(pi,ni)
F	1	2	0.918
T	1	1	1

$= \sum \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.9508$

gain(windy) = 0.97 - 0.9508 = 0.0192

T	H	W	
Mild	High	False	Y
Cool	Norm	False	Y
Cool	Norm	True	N
Mild	Norm	False	Y
Mild	High	True	N

$I(3,2) = 0.971$

	Pi	ni	I(pi,ni)
M	2	1	0.918
C	1	1	1

$= \sum \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.95$

gain(temp) = 0.97 - 0.95 = 0.0192

	Pi	ni	I(pi,ni)
H	1	1	1
N	2	1	0.918

= 0.95

gain (humidity) = 0.0192

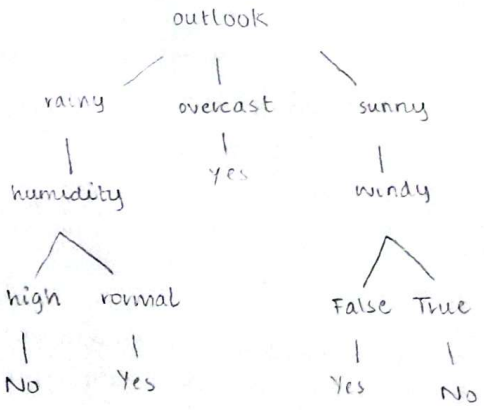
	Pi	ni	I(pi,ni)
F	3	0	0
T	0	2	0

✓ gain (windy) = 0.97

T	H	W	
Hot	High	False	Y
Cool	Norm	True	Y
Mild	High	True	Y
Hot	Norm	False	Y

↓

✓ play-golf = yes.



TID	Items
T100	1, 3, 4
T200	2, 3, 5
T300	1, 2, 3, 5
T400	2, 5

minsup = 0.5

minconf = 50%

STEP 1

1 (a) First Scan (item : count)

$C_1 \rightarrow \{1\}:\frac{2}{4}, \{2\}:\frac{3}{4}, \{3\}:\frac{3}{4}, \{4\}:\frac{1}{4}, \{5\}:\frac{3}{4}$ $\rightarrow \frac{1}{4} < 0.5$ so discard

$F_1 \rightarrow \{1\}:2, \{2\}:3, \{3\}:3, \{5\}:3$

(b) joining C_2 → make pairs of F_1

$C_2 = (1,2), (1,3), (1,5), (2,3), (2,5), (3,5)$

(c) pruning C_2

C_2 = all items in C_2 belong to F_1

2 (a) Second Scan (item : count)

$C_2 = \{1,2\}:\frac{1}{4}, \{1,3\}:\frac{2}{4}, \{1,5\}:\frac{1}{4}, \{2,3\}:\frac{2}{4}, \{2,5\}:\frac{3}{4}, \{3,5\}:\frac{2}{4}$

$F_2 = \{1,3\}:2, \{2,3\}:2, \{2,5\}:3, \{3,5\}:2$

(b) joining C_3 → make triples of F_2

$C_3 =$ [1,3] → [2,3], [2,5], [3,5] [2,3] → [2,5] → [2,3,5] (with [3,5]) [2,5] → [3,5]

(c) pruning $C_3 = [2,3,5]$

$F_2 = [2,3], [2,5], [3,5]$ all belong to $C_3 = [2,3,5]$

3 (a) Third Scan (item : count)

$C_3 = \{2,3,5\}:\frac{2}{4}$

$F_3 = \{2,3,5\}, 2$

(b) joining C_4 → only 1 set of F_3 exist

$C_4 = [2,3,5]$

(c) pruning C_4

$C_4 = [2,3,5]$

STEP2 :

minconf = 50%

non empty subsets = [2,3], [2,5], [3,5], [2], [3], [5]
50% 75% 50% 75% 75% 75%

association rules :

[2,3] → 5 = 2/2 = 100% [2,3,5]/[2,3]

[2,5] → 3 = 2/3 = 67% [2,3,5] / [2,5]

[3,5] → 2 = 2/2 = 100% [2,3,5] / [3,5]

[2] → [3,5] = 2/3 = 67% [2,3,5] / [2]

[3] → [2,5] = 2/3 = 67% [2,3,5] / [3]

[5] → [2,3] = 2/3 = 67% [2,3,5] / [5]

AdaBoost

day / date:

CGPA	Interactiveness	Practical Knowledge	Communication Skills	Job Profile
>=9	Yes	Good	Good	Yes
<9	No	Good	Moderate	Yes
>=9	No	Average	Moderate	No
<9	No	Average	Good	No
>=9	Yes	Good	Moderate	Yes
>=9	Yes	Good	Moderate	Yes.

1. CGPA

C GPA	Predicted values	Actual values	Job Weight
>=9	Yes	Yes	1/6 0.167
<9	No	Yes	1/6
>=9	Yes	No	1/6
<9	No	No	1/6
>=9	Yes	Yes	1/6
>=9	Yes	Yes	1/6

2. calculate error and updated weights.

$$\text{correct} = 4 \times \frac{1}{6}$$
$$\text{incorrect} = 2 \times \frac{1}{6}.$$

$$\alpha = \frac{1}{2} \ln\left(\frac{1 - \frac{2}{6}}{\frac{2}{6}}\right) = 0.347.$$

3. normalize all the weights.

weight x # of correct x $e^{-\alpha}$ + weight x # of incorrect x e^{α}

$$= \frac{4 \times 0.167 \times e^{-0.347} + 0.167 \times 2 \times e^{0.347}}{}$$

$$= 0.472 + 0.945 = 0.9428.$$

4. calculate updated weight for correct and incorrect.

$$\text{new\_weight\_correct} = \frac{\frac{1}{6} \times e^{-0.34}}{0.9428.} = 0.125$$

$$\text{new\_weight\_incorrect} = \frac{\frac{1}{6} \times e^{0.34}}{0.9428} = 0.248.$$

day / date:

2. Interactiveness

Interactiveness	Predicted values	Actual values	Weight
Yes	Yes	Yes	0.125
No	No	Yes	0.248
No	No	No	0.248
No	No	No	0.125
Yes	Yes	Yes	0.125
Yes	Yes	Yes	0.125

3. calculate error and updated weights

correct = 5
incorrect = 1

$$\alpha = \frac{1}{2} \ln\left(\frac{1 - 0.248}{0.248}\right) = 0.555$$

4. normalize all weights

$$= 0.125 \times 4 \times e^{-0.555} + 0.248 \times 1 \times e^{-0.555} + 0.248 \times 1 \times e^{0.555}$$

$$= 0.866$$

5. updated weights

$$\text{new\_weight\_correct} = \frac{0.125 \times e^{-0.555}}{0.866} = 0.0829$$

$$\text{new\_weight\_correct} = \frac{0.248 \times e^{-0.555}}{0.866} = 0.164$$

$$\text{new\_weight\_incorrect} = \frac{0.248 \times e^{0.555}}{0.866} = 0.499$$

day / date:

3. Practical Knowledge

Practical Knowledge	Predicted	Actual	Weights
Good	Yes	Yes	0.0829
Good	Yes	Yes	0.499
Average	No	No	0.164
Average	No	No	0.0829
Good	Yes	Yes	0.0829
Good	Yes	Yes	0.0829

4. calculate error and update weights.

correct : 6 $\quad \alpha = \frac{1}{2} \ln\left(\frac{1-0}{0}\right)$

incorrect : 0

• weights will not be updated.

4. Communication Skill.

Communication Skill	Predicted	Actual	Weights
Good	Yes	Yes	0.0829 →
Moderate	No	Yes	0.499 x
Moderate	No	No	0.164
Good	Yes	No	0.0829 x
Moderate	No	Yes	0.0829 x
Moderate	No	Yes	0.0829 x

5 calculate error and update weights

correct : 2
incorrect : 4
$\alpha = \frac{1}{2} \ln\left(\frac{1-0.7477}{0.7477}\right) = -0.54$

error = $0.499 \times 1 + 0.0829 \times 3 = 0.7477$

day / date:

6. normalize weights

$= 0.0829 \times e^{-(-0.54)} + 0.164 \times e^{-(-0.54)} + 0.0829 \times 3 \times e^{+(0.54)}$

$+ 0.499 \times e^{-0.54}$

$= 1.7067 \quad 0.7627 \quad 0.859$

$\frac{0.0829 \times e^{0.54}}{1.7067 \; 0.7627 \; 0.859} = 0.0833 \quad 0.1865 \quad 0.1663$

$\frac{0.164 \times e^{0.54}}{1.7067 \; 0.7627 \; 0.859} = 0.1648 \quad 0.3689 = 0.3331$

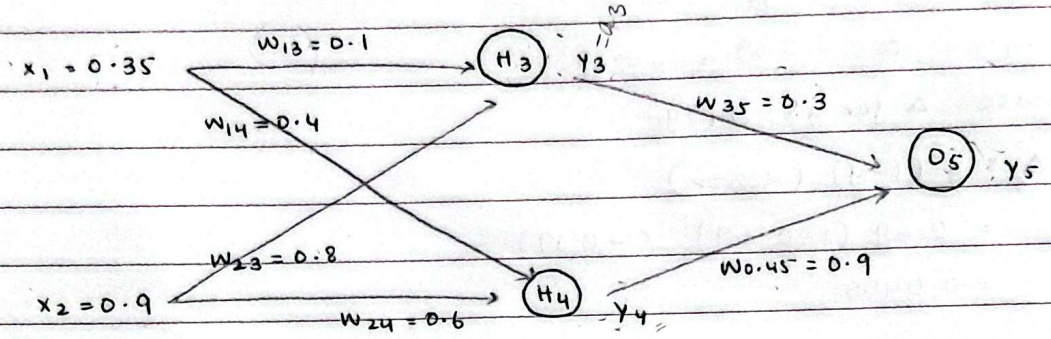
$\frac{0.0829 \times e^{-0.54}}{1.7067 \; 0.7627 \; 0.859} = 0.0633 \quad = -0.8555$

$\frac{0.499 \times e^{-0.54}}{1.7067 \; 0.7627 \; 0.859} = 0.381 \quad = 0.3337$

→ conclusion	$\alpha' = 0.347$	$\alpha = 0.555$	$\alpha = -0.54$	weight	FP
	Y	Y	Y	0.362	Y
	N	N	N	0	N
	Y	N	N	0.347	Y
	N	N	Y	-0.5485	N
	Y	Y	N	0.896	Y
	Y	Y	N	0.896.	Y

day / date:

BACK PROPAGATION



$H_3 \rightarrow H_4 \rightarrow O_5$

$\downarrow$
a_3

- $a_3 = (0.35)(0.1) + (0.8)(0.9)$

sigmoid function $= \dfrac{1}{1+e^{-a_3}}$

$$SF = \frac{1}{1+e^{-0.8755}} = 0.68$$

- $a_4 = (0.35)(0.4) + (0.9)(0.6) = 0.68$

$$SF = \frac{1}{1+e^{-0.68}} = 0.6637$$

- $y_5 a_5 = (0.68)(0.3) + (0.6637)(0.9) = 0.801$

$$SF = \frac{1}{1+e^{-0.801}} = 0.69$$

day / date:

2. calculate error.

error = target − y_5

error = 0.5 − 0.69 = −0.19

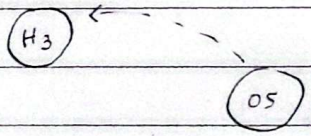
3. calculate Δ for y_3, y_4, y_5

$\Delta_5 = y(1-y)(y_{error})$

$= 0.69(1-0.69)(-0.19)$

$= -0.0406$

$\Delta_3 =$



$\Delta_3 = y(1-y)(w_{35} \times \Delta_5)$

$= 0.68(1-0.68)(0.35 \times -0.0406)$

$= -0.00265$

$\Delta_4 = y(1-y)(w_{45} \times \Delta_5)$

$= 0.6637(1-0.6637)(0.9 \times -0.0406)$

$= -0.0082.$

4. new weight for each line, 45, 35, 13, 14, 23, 24

new weight = $\eta \times \Delta \times y$ $\eta = 1$

$w_{45} = 1 \times -0.0406 \times 0.6637 = -0.0269$

new weight + original = $0.9 + (-0.0269) = 0.8731$

$w_{35} = 1 \times -0.0406 \times 0.68 = -0.0276$

new + original = $0.3 + (-0.0276) = 0.2724$

$w_{13} = 1 \times -0.00265 \times 0.35 = -0.00287$

new + original $= 0.1 + 0.00287 = 0.0991$

$w_{14} = 1 \times -0.0082 \times 0.35 = -0.0287$

new + original $= 0.14 + (-0.0287) = 0.3971$

$w_{23} = 1 \times -0.00265 \times 0.9 = -0.002385$

new + original $= -0.002385 + 0.8 = 0.7976$

w_{24}

new + original = 0.5926