Regression: modeling ~~the~~ and analysis

Histogram: Divide data into buckets and store average (sum) for each bucket.

PCA: a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components

Bining: 19, 71, 48, 63, 35, 85, 69, 81, 72, 88, 99, 85

equal-f: bin1: 19.35.48
bin2: 63, 69, 71

equal-width: width = (99-19)/4 = 20
bin1 = [19,39) 19,35
bin2 = [39,59) 48

Important Characteristics:
- Dimensionality
- Sparsity — KL: the expected number of extra bits required to code q(x) from p(x)
- Resolution
- Distribution

Data warehouse usage:
- information processing - statistical
- Analytical processing - dimensional
- Data Mining -

shell-fragment: 4 | 6 — 2^6 x

4: a = 1
2: b = $\binom{4}{2} \times 2^2 = 24$
3: b = $\binom{4}{3} \times 2 = 8$
4: b = 1

Sup=2, Closed cell $\times 2^6$
$\binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 11$

Mean: $\bar{x} = \frac{1}{n}\sum x_i$, $\mu = \frac{\sum x}{N}$

Median: $L_1 + \left(\frac{n/2 - \sum f}{f}\right)$ width

Mode: unimodal: mean − mode = 3 × (mean − median)

Var: $s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2 = \frac{1}{n-1}\left[\sum x^2 - \frac{1}{n}(\sum x)^2\right]$

$\sigma^2 = \frac{1}{N}\sum(x-\mu)^2 = \frac{1}{N}\sum x^2 - \mu^2$

Data visualization:
- Pixel-oriented
- Geometric: Direct visualization
- Scatterplot & scatterplot matrices (k^2-k)
- Landscape - Prosection - Hyperslice
- Parallel coordinates
- Icon-Based: Chernoff Faces
→ Shape coding [stick Figures
→ color icons → Tile bars
- Hierarchical: Dimensional stacking
- World-within-World - Tree-Map disk — 2D fixed, cannot avoid overlap
- Cone Trees - Info Cube
- Tag cloud: font size/color

$z = \frac{x-\mu}{\sigma}$, robust $z_{if} = \frac{x_{if} - m_f}{s_f}$, $s_f = \frac{1}{n}(|x_{1f} - m_f| + \dots)$

Minkowski distance:
$d(i,j) = \sqrt[p]{|x_{i1}-x_{j1}|^p + |x_{i2}-x_{j2}|^p + \dots + |x_{il}-x_{jl}|^p}$
- $d(i,j) > 0$ if $i \neq j$; $d(i,i) = 0$; $d(i,j) = d(j,i)$
- $d(i,j) \le d(i,k) + d(k,j)$

p=1: Manhattan, # bits diff binary; p=2: Euclidean distance
p=∞: supremum $\max_{f=1}^{l} |x_{if} - x_{jf}|$
maximum diff between 2 vectors

Proximity Measure for Binary Attributes

categorical

		object j 1	0	sum
object i	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p

Symmetric (equal): $d(i,j) = \frac{r+s}{q+r+s+t}$

Asymmetric: $d(i,j) = \frac{r+s}{q+r+s}$

Jaccard: $sim_{Jac}(i,j) = \frac{q}{q+r+s}$ → medical test

Proximity Measure for Categorical:
- Simple Matching: $d(i,j) = \frac{p-m}{p}$
- Use a large number of binary

Ordinal Variables: $r_{if} \in \{1, \dots, M_f\}$, $z_{if} = \frac{r_{if}-1}{M_f-1}$

Mixed type: $d(i,j) = \frac{\sum_{f=1}^{p} w_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} w_{ij}^{(f)}}$
- numeric: ~~no~~ normalized distance
- binary/nominal: 0, if $x_{if} = x_{jf}$, 1
- ordinal: rank $z_{if} = \frac{r_{if}-1}{M-1}$

Cosine Similarity: document term vectors in text file
$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \times \|d_2\|}$

Chapter 3. Data Preprocessing

Data Quality: - accuracy - completeness - Consistency - Timeliness - Believability - interpretability

Incomplete (Missing) Data:
- inference-based: Bayesian formula, decision tree

Noisy Data:
- Binning: sort, partition → smooth: mean, median, boundary
- Regression: regression function
- Clustering: detect & remove outliers
- Semi-supervised: computer & human

Data Integration: Chi-square Correlation Analysis (categorical)
$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$ (Null: independent)
row × column = expected (E_i)
(row−1) × (column−1) = DOF

Covariance for two Variables
$\sigma_{12} = E[(X_1-\mu_1)(X_2-\mu_2)] = E[X_1X_2] - \mu_1\mu_2$
$= E[X_1X_2] - E[X_1]E[X_2]$
if X_1, X_2 are independent, $\sigma_{12} = 0$

Correlation between Numerical — decision tree
$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$
$\rho_{12} > 0$ positive, $X_1 \uparrow$ as $X_2 \uparrow$
$\rho_{12} = 0$ independent (× linear) (normal dist)
$\rho_{12} < 0$ negative — $y = x^2$
[−1, 1]. Normalized covariance

Covariance Matrix: det ≥ 0
$\Sigma[(X-\mu)(X-\mu)^T] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$

Numerosity Reduction:
Parametric: fit some model
Non-Parametric: no model

Regression Analysis: dep (response) → indep (explanatory) → predictors
- Linear ~ - Nonlinear ~
- Multiple ~ - Log-linear ~

Histogram Analysis: binning → distr
Data → buckets → avg.
- equal-width - equal-frequency

Clustering: similarity, representation effective if clustered.

Sampling: Small s represent whole data
Key principle: representative subset
types: - simple random sampling
- Sampling without replacement
- Sampling with replacement
- Stratified sampling

Data Cube Aggregation: The aggregation data for an individual entity of interest.

Data Compression: lossless vs. lossy

Wavelet transform: size ↓
- use hat-shape filters
- effective removal of outliers
- Multi-resolution
- Efficient O(N) - only low D

Data Transformation: to a new set of representatives data
map to new set
- Smoothing - Attribute/feature construction
- Aggregation - Normalization
- Discretization

Normalization:
min-max: $v' = \frac{v-min}{max-min}(new\_max - new\_min) + new\_min$
Z-score: $v' = \frac{v-\mu}{\sigma}$
decimal scaling: $v' = \frac{v}{10^j}$ (Max(|v'|) < 1)

Data Discretization:
- Data size ↓ - supervised vs ×
- split (top-down) vs. merge (bot-up)
- Binning: split. unsup.
  - equal-width: W = (B−A)/N uniform grid
    → outlier dominant. skewed ×
  - equal-depth: N interval, same size
    → good scaling, × categorical
- histogram: split. unsup
- clustering: split, merge, unsup
- Decision tree: split. sup (class info)
- correlation: merge. sup

concept hierarchy generation:
- drilling & rolling / Nominal

Dimensionality Reduction:
- avoid curse - eliminate noise
- reduce space & time - easier visualization

Feature selection: subset
Feature extraction: transform high D to low D

Principal Component Analysis (PCA)
orthogonal transformation, eigenvector
→ Normalize → k ortho vectors → linear combi of k → eliminate weak component
For numeric only → sorted

- Attribute Subset Selection:
  - Redundant attributes
  - Irrelevant attributes
  - Heuristic search: 2^d
    → best single attri → best step-wise select
    → best step-wise eliminate → best combi
    → optimal branch & bound
- Attribute Creation (Feature generation)
  new more effective set
  - Attributes extraction
  - Mapping data to new space
  - Attribute construction

Data integration: - Entity identification - Remove redundant - Detect inconsistency

Data reduction: - Dimensionality - Numerosity - Data Compression

Chapter 4: OLAP & data warehouse

- decision support processing, separately from the organization's operational database
- Support information processing by providing a solid platform of consolidated, historical data for analysis

Data warehouse: - subject-oriented - integrated - time-variant - nonvolatile

subject-oriented: organized around major subjects - modeling & analysis → decision - simple & concise: excluding not useful

Integrated: integrating multiple, heterogeneous data source. - Data cleaning & data integration

Time Variant: longer time horizon
- operational: current time - no key time
- warehouse: historical, key → time element

Nonvolatile: - independent → physical separation store. - Static: no update of data ⇒ 1. initial loading 2. access of data

OLTP: online transactional processing
- Day to Day - application-oriented
- repetitive - small size - more users

OLAP: online analytical processing
- historical - complex - large

Why separate: - High performance
- Different functions & data: missing data → data consolidation - data quality

Architecture: - Top tier: Front End Tools - Middle: OLAP - Bottom: Data Warehouse - Data

Enterprise warehouse: subject, entire org.
Data Mart: specific group {indep, dep (direct fm)}
Virtual warehouse: views, summary

Extraction, transformation, loading (ETL)
- Data extraction - Data cleaning
- Data transformation - load - refresh

Meta data: description of structure
- operational meta-data: → data lineage → currency of data → monitoring info
- algorithm - summary - mapping - business data
- data related to performance

Data Cube: a lattice of cuboids
Base cuboid: n-D base cube
apex cuboid: 0-D highest summary

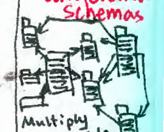
Data warehouse: multidimensional
- Dimension tables: item, time
- Fact table: measures, key

Star Schemas

mid dimension Fact table

Snowflake Schemas

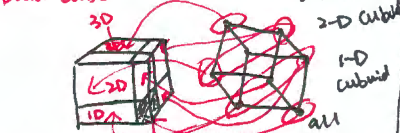
refinement star D-hierarchy normalized

Constellation Schemas

Multiply Fact table share D tables

Data Cube Measure: 3 Categories:

- Distributive: if the result derived by applying the function n aggregate values is the same as that derived by applying the function on all the data without partition
- Algebraic: it can be computed by an algebraic function with M arguments, each of which is obtained by applying a distributive aggregate
  - avg(x) = sum(x)/count(x) • max min
- Holistic: if there is no constant bound on the storage size needed to describe a subaggregate
  - median(), mode(), rank()

Data Cube:



Typical OLAP Operations:

- Roll up (drill-up): summarize data
- climbing hierarchy • D reduction
- Drill down (roll down): reverse
- Slice & dice: project & selection
- pivot (rotate): reorient
- Drill across: involving more than, fact table
- Drill through: bottom level to b-e SQL

Design: Business Analysis Framework:

- Top-down: select of relevant data
- Data Source View: expose info, oper. sys.
- Data warehouse: fact table, D table
- Business query: persepective, end-user

Data Cube Computation: $T = \prod_{i=1}^{n}(L_i + 1)$

Bitmap index: • value → vector

- len(vector) = # of records

Join indices: • $J(R,S) \Rightarrow R(R \cdots) \bowtie S(S \cdots)$

- relates the value of dimensions → row in the fact table

ROLAP: relational OLAP: extended-r. DBMS optimization of DBMS. • Greater Scalability

- MOLAP: multidimensional: • sparse-array based • fast indexing → pre-computed
- Hybrid (HOLAP): flexibility

Chapter 5: Data Cube Technology

Iceberg Cube: having count(*) >= min

- no need to save value below threshold
- avoid computing the un-needed
- avoid explosive growth

close cube: no cell d ∈ descendant of c → d has same measure value as c

cube shell: cuboid, small # D

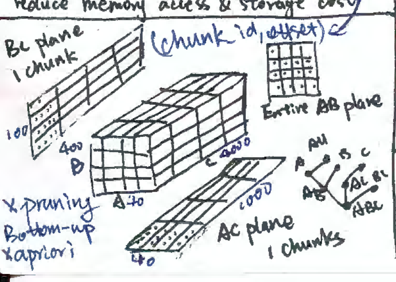
General Heuristic: Sorting, hashing, grouping

- Smallest-child: smallest pre cuboid
- Cache-results: caching result & reduce disk I/O
- Amortize-Scans: as many as possible
- Share-sorts: share sort cross multi cuboid
- Share-partition: hash-based

Multi way array aggregation (bottom up)

- multi-D chunks • simultaneous aggre
- intermedia aggre value reused
- Cannot Apriori pruning: iceberg ×
- Efficient for small
- Partition chunks: small subcube, memory
- compressed sparse array addressing
- aggregation in multiway. Visiting the cube cells → min the # of visit, reduce memory access & storage cost

(chunk id, offset)

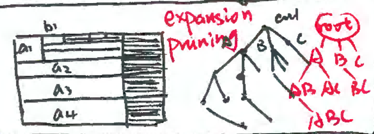


× pruning
Bottom-up
× apriori

BUC: top-down: partition & iceberg pruning

- if partition < min, pruning child
- No simultaneous aggregation

- Entire data cannot fit in memory
- Sort distinct values
- Continue processing • optimizations
- Partitioning - ordering D - x duplicate
- → cannot holistic aggregates ×



High-D OLAP with Shell-fragment

- Trade off between the amount of pre-computation and the speed of online computation
- Reduce high-D ⇒ low-D cubes (shell fragment)
- Online re-construction of h-D space ← pre-compute
- Lossless reduction

tid	A	B	C	D
1	a1	b1	c1	d1
2	a1	b2	c1	d2
3	a1	b2	c1	d1
4	a2	b1	c1	d1
5	a2	b1	c1	d1

Value	TID	Size
a1	123	3
a2	45	2
b1	145	3
b2	23	2
c1	12345	5
d1	1345	4
d2	2	1

cell	∩	TID	size
a1b1	123∩145	1	1
a1b2	123∩23	23	2
a2b1	45∩145	45	2
a2b2	45∩23	∅	0

Shell-fragment

Size & design: T: tuples D: dimension F: shell-fragment size.

- → Cube space: $O\left(T\lceil\frac{D}{F}\rceil(2^F - 1)\right)$
- × disjoint

General Form: $\langle a_1, a_2, \cdots, a_n : M \rangle$

3 values: • instantiated value • Aggregate * function • inquire ? function

Complex Aggregation at Multiple Granularities

- Multi-Feature cubes: compute complex queries involving multiple dependent aggregates at multiple granularities
- Discovery-driven Exploration of Data Cubes
  - huge cube space • pre-compute indicating Exception

Chapter 6: Frequent Pattern

Pattern: itemset $X = \{x_1 \ldots x_k\}$

(absolute) support: frequency of number

- relative support: fraction

Associate Rule: $X \to Y(s,c)$

- Support: $\frac{X \cup Y}{total}$ • Confidence = $\frac{\sup(X \cup Y)}{\sup(X)}$

- closed pattern: no $Y \supset X$ w/ same sup
  $P_1\{a_1 \ldots a_{50}\}: 2$ $P_2\{a_1 \ldots a_{100}\}: 1$
  - lossless compression
- Max Pattern: no $Y \supset X$.
  $P_1\{a_1 \ldots a_{100}\}: 1$ do not care sup subpattern
  - lossy compression

Downward closure (Apriori): Any subset of a frequent itemset must be frequent

Apriori pruning principle: if there is any itemset which is infrequency, no more super-set will be generated.

→ self-join → prune

Partition: reduce passes:

- local frequent → global frequent

DHP: Direct Hashing & Pruning

- reduce number of candidates
- whose corresponding hash below threshold

ECLAT: Exploring Vertical Data Format:

- t(x) = t(Y), together
- t(x) ⊂ t(Y): x, always Y.

10	a,c,d,e
20	a,b,e
30	b,c,e

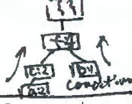
→

a	10,20
b	20,30
c	10,30

diffset
diffset → tid

FP Growth: Pattern Growth.

100 {f,a,c,d,g,i,m,p} {f,c,a,m,p}
200 {a,b,c,f,l,m,o} {f,c,a,b,m}
300 {b,f,h,j,o,w} {f,b}



a f: 2
b f: 1
conditional

$r_1 = \{\}$ $r_2 = \{\}$

Parallel projection: each frequency item
Partition project: in order

closed: if Y appears in every occurrence of X, then Y is merged with X

Interestingness: subjective vs objective

- Lift: $lift(B,C) = \frac{c(B \to C)}{s(C)} = \frac{s(B \cup C)}{s(B) \times s(C)}$
  =1: independent >1 + <1 −
- X^2: $X^2 = \sum \frac{(O-E)^2}{E}$ {=0 indep; >0 dep

- Null-invariant: value x change → # Null
- IR (Imbalance Ratio): $IR(A,B) = \frac{|S(A) - S(B)|}{S(A) + S(B) - S(A \cup B)}$
- Kulczynski: $Kul(A,B) = \frac{1}{2}\left(\frac{S(A \cup B)}{S(A)} + \frac{S(A \cup B)}{S(B)}\right)$

Chapter 7: Advanced FP

Multiple-Level: uniform / reduced min

- Redundant Rule • group based min

Multiple-D: • 1-D. • Multi-D

- inter-D (no repeat) • Hybrid-D (repeat)

Quantitative: numerical • Static: concept

- Dynamic: distribution • Clustering: distance
- Deviation: G = Fe ⇒ Wage: mean = ? (average)

- Extraordinary: • Rule: confirm • Sub: highly
- Rare Pattern: group-based min_sup
- Negative: • sup-based ⇒ $\sup(A \cup B) \ll \sup(A) \cdot \sup(B)$ → × null
  - kulcy: $(P(A|B) + P(B|A))/2 < \varepsilon$
- Compressed: $D(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|}$

Constrains: • knowledge type • Data • D / Level • Rule (Pattern) • interesting

all autonomously × interactive ✓

Pattern Space Pruning (data mining query)

- Anti-monotonic: Violate → terminate
  i.e. Apriori is anti-monotonic
- Monotone: Satisfies
  e.g. • sum < v → anti > v mono
  - range < v → anti > v mono
  - min < v mono • sup > σ anti

Data Space Pruning:

- Anti-monotone: if entry cannot satisfy pattern p. under c.
  e.g. sum ≥ v. min ≤ v. range > v.
  - check recursively

Succinctness: if the constrain c can be enforced by directly manipulating the data

- no i → remove i • with i → i-projected
- min < v ✓ • sum > v × → in any

Convertible Constraints:

tough → (anti) monotone by ordering

Core Pattern: Subpatterns that cluster around α by sharing a similar τ sup

Robustness: $\left|\frac{D_\alpha}{D_\beta}\right| \geq \tau$ $0 < \tau \leq 1$

(d,τ) robust if d is max number of items that can removed from α

- $\Omega(2^d)$ core Pattern. • colossal → core

- Dense ball: $D(\alpha,\beta) = 1 - \frac{|D_\alpha \cap D_\beta|}{|D_\alpha \cup D_\beta|}$

Sequential Pattern: set of sequences ⇒ complete set of frequency subsequences

⟨(ab)c⟩. an element may contains a set of items

GSP: Apriori-Based SPM:

- initial → scan → generate

SPADE: Vertical Data Format

- map to ⟨SID, EID⟩
- Grow the subsequences one item at a time by priori candidate generation

Pattern Growth: PrefixSpan

- Given ⟨a(abc)(ac)d(cf)⟩
- Prefix ⟨a⟩, ⟨aa⟩, ⟨a(ab)⟩, ⟨a(abc)⟩
- Suffix: Prefixes-based projection

CloSpan: There exists no super-P s' ⊃ s, s' & s has same sup.

→ reduce # of Pattern → same expressive power

- S ⊃ S₁, S₁ is closed iff same size

Constraint-Based SPM:

- Anti-monotonic • Monotonic
- Data Anti-monotonic • Succinct
- Convertible

Timing-Based Constraints:

- order constraints • min-gap / max-gap Constrs
- Max span: time diff 1st & last
- Window size: events in an ele do not occur at same time

Episodes, Episode SPM:

- Regular Expression: serial A→B
  Parallel: A|B ⇐ partial order relationship
  Regular Exp: (A|B)C*(D→E)
- Method: Variation of GSP

Frequent Graph Patterns:

$D = \{G_1, G_2, \cdots G_n\}$
$D_g = \{G_i | g \subseteq G_i, G_i \in D\}$
$Support(g) = |D_g| / |D|$ ≥ min_sup

- Apriori-Based: candidate generation → pruning → sup counting → cand eliminate
- Vertex Growing vs. Edge Growing: new graph with new vertex/edge
- Pattern-Growth: (duplicate)

$(a_1, a_2, a_3, a_4, \cdots a_{10})$
$(b_1, a_2, a_3, a_4, \cdots a_{10})$ 3/6
$(c_1, c_2, a_3, a_4, \cdots a_{10})$ 11
$(d_1, d_2, d_3, a_4, \cdots a_{10})$
$(D_1,$ —— $\cancel{a_4 - a_{10}}$ $)$ 4×2^7
$(*, D_2,$ — $\cancel{a_4 - a_{10}}$ $)$ 3×2^8
$(*, *, D_3$ $\cancel{a_4 - a_{10}}$ $)$ 2×2^7
$(*, *, *,$ —— $)$ 2^7

$X^2(A,B)$	$\sum_{i,j=1} \frac{(e-o)^2}{e(a_i b_j)}$	[0, ∞]	No
Lift (A,B)	$\frac{S(A \cup B)}{S(A) \times S(B)}$	[0, ∞]	No
AllConf (A,B)	$\frac{S(A \cup B)}{\max\{S(A), S(B)\}}$	[0, 1]	Yes
Jaccard (A,B)	$\frac{S(A \cup B)}{S(A) + S(B) - S(A \cup B)}$	[0, 1]	Yes
Cosine (A,B)	$\frac{S(A \cup B)}{\sqrt{S(A) \times S(B)}}$	[0, 1]	Yes
Kulczynski (A,B)	$\frac{1}{2}\left(\frac{S(A \cup B)}{S(A)} + \frac{S(A \cup B)}{S(B)}\right)$	[0, 1]	Yes
max conf (A,B)	$\max\left\{\frac{S(A)}{S(A \cup B)}, \frac{S(B)}{S(A \cup B)}\right\}$	[0, 1]	Yes

$Sup(A) = A / total$

Bitmap: