**Data Exploration**

1. Types of Raw Data Resources:

Record, Graph and Network, Ordered, Spatial, image and multimedia

2. Types of Attributes:

Nominal, Ordinal, Interval, Ratio

3. Proximity:

Similarity: [0,1]

Dissimilarity: [0,inf)

4. Proximity for Binary

A white paper with black text and numbers

Description automatically generated

5. Proximity for Nominal

Method1: Simple Matching SMC(i,j)=#attributes match/#all attributes

Method2: Convert to binary

6. Distance on numerical:

Minkowski Distance: A math equations and formulas

Description automatically generated with medium confidence

H=1: Manhattan distance

H=2: Euclidean distance

H=inf: Supremum distance

7. Similarity on numerical:

Cosine similarity:

cos(x, y) = (x • y) /(||x|| ||y||)

8. Correlation:

Limited to linear relationship

Corr = covariance(x,y)/sd(x)sd(y)

A math equations and formulas

Description automatically generated with medium confidence

-1: perfect negative relationship

+1: perfect positive relationship

0: no linear relationship

A black and white box with black text

Description automatically generated

9. Proximity for ordinal -> numerical:

A math equation with black text

Description automatically generatedA black number with dots

Description automatically generated

10. Proximity on mixed typeA math equations on a white background

Description automatically generated

11. Skewed Data

Right(positive) skew: mode<median<mean

Left(negative) skew: mean<median<mode

A black text on a white background

Description automatically generated

**Data Preprocessing**

1. Measure for data quality:

Accuracy, Completeness, Consistency, Timeliness, Believability, Interpretability

2. Major tasks:

Data cleaning, integration, reduction, transformation and discretization

**3. Data cleaning:**

Incomplete, noisy, inconsistent, intentional

4. Handle missing data:

Ignore tuples, Fill manually, Fill with a global constant, the attribute mean, the attribute mean for all samples belonging to the same class, inference-based such as Bayesian formula or decision tree

**5. Data integration:**

Entity identification problem: Identify real world entities from multiple data sources

6. Handle redundant data:

Due to object identification or derivable data,

detected by correlation analysis and covariance analysis

7. Correlation Analysis (Nominal Data)

A black and white text

Description automatically generated

8. Correlation Analysis (Numerical Data)

A math equations and formulas

Description automatically generated with medium confidence

Correlation Calculation:

A group of math equations

Description automatically generated

9. Covariance (Numerical Data)A math equations with black text

Description automatically generated with medium confidenceA black text on a white background

Description automatically generated

Simplified:

**11. Data Reduction**

Dimensionality Reduction:

When dim increase, data is increasingly sparse, density and distance less meaningful, possible combinations of subspaces grow exponentially

12. Principal Component Analysis

A diagram of a mathematical equation

Description automatically generated

A black text on a white background

Description automatically generated

A math equations and formulas

Description automatically generated with medium confidence

In general, 

the kth principal component corresponds to the eigenvector of the kth largest eigenvalue of S

13. Attribute Creation

Attribute extraction, Mapping data to new space, Attribute construction

14. Numerosity Reduction

Parametric method: Assume model

Linear regression, multiple regression, log-linear model

Non-parametric: Do not assume model

histograms, clustering, sampling

15. Regression Analysis:

Dependent (response, measurement) vs independent (explanatory, predictors)

Least Squares Method

16. Histogram Analysis:

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

Partitioning rules: Equal-width: equal bucket

range / Equal-frequency (or equal-depth)

17. Clustering:

Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only

18. Sampling:

obtaining a small sample s to represent the whole data set N

A black and white text

Description automatically generated

19. Data Compression

Original Data ->lossless<- Compressed data ->lossy->original data approximated

**20. Data Transformation**

Normalization

**A math equations and formulas

Description automatically generated with medium confidence**

21. Data Discretization Methods

A close up of words

Description automatically generated

22. Binning Methods for Data Smoothing

Equal-width: max-min/#bin (When you need consistent interval sizes.)

Equal frequency: # inside bin the same (When balanced bin sizes are important)

Clustering-based:Use cluster algo to form bin (when the data has natural groupings)

**Frequent Pattern Mining**

1. Relative support of itemset: fraction of transaction containing the itemset

2. Relative support of rule: 

Confidence of rule:

Strong rule: with relative support and confidence higher than their threshold

3. Association Rule Mining Task: All strong rules

Step1: Frequent Itemset Generation

Step2: Rule generation

4. Closed itemset: none of its immediate supersets has the same support as the itemset

Maximal itemset: none of its immediate supersets is frequent

5. Apriori Principle:



A close up of a text

Description automatically generated

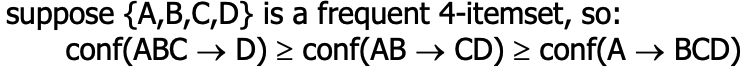
Limitation: BFS, generate huge candidates, repetitive scan of supports

6. Rule generation

A black text with black letters

Description automatically generated

Relation:



7. Correlations (Lift)

A close up of black text

Description automatically generated

8. FP growth

DFS, avoid generate huge candidate, avoid repetitive scan

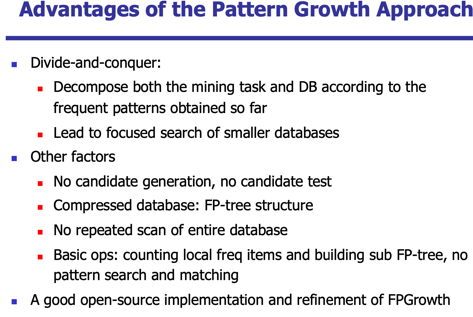
Construct FP tree:

A black text on a white background

Description automatically generated

A diagram of a system

Description automatically generated

A list of information on a computer

Description automatically generated

9. Handling binary attributes:

Asymmetric can be converted to item occurrence

Symmetric can be convert to asymmetric then occurrence

10. Handling categorical attributes:

Some attributes have many values->aggregate over low-support values

Distribution highly skewed->discard the highly frequent items

11. Discretization-based: bin continuous variable in rule left part

Interval too wide:

May merge several disparate patterns

May lose some interesting patters

Interval too narrow:

Pattern is broken up into smaller patterns

Some windows my not meet supportA screenshot of a computer

Description automatically generated

12. Statistics-based: rule right part use statistics

A white background with black text

Description automatically generated

13. Multi-level association rules

A white text with black text

Description automatically generated

14. Constraint-Based Frequent Pattern Mining

-Prune the pattern space:  
-Anti-monotonic: If constraint c is violated, its further

-Monotonic: If c is satisfied, no need to check c again for its further mining.

-Succinct: we can explicitly and precisely determine if any itemset satisfies the constraint by examining if it contains some specific items.

-Convertible: c is not monotonic nor anti-monotonic nor succinct, but it can be converted into it if items in the transaction can be properly ordered

-Strongly convertible constraint: if both the following are satisfied:  
The constraint is convertible anti-monotone w.r.t. item value descending order;  
The constraint is convertible monotone w.r.t. item value ascending order

**Classification**

1. supervised learning: Regression vs Classification

2. Decision tree

A white text with black text

Description automatically generated

A black text on a white background

Description automatically generatedA black text on a white background

Description automatically generated

3. Measure of node impurity

-Gini IndexA black text on a white background

Description automatically generated

A mathematical equation with black text

Description automatically generated

-EntropyA math equations and formulas

Description automatically generated with medium confidence

-Classfication ErrorA black and white text

Description automatically generated

4. Finding the best to split

A white background with black text

Description automatically generated

5. Gain Ratio

A white text with black text

Description automatically generated

6. Bayes Classifier

A mathematical equation with black text

Description automatically generatedA black text with a white background

Description automatically generated with medium confidence

Issue with Naïve. Bayes:

If one of the conditional probabilities is zero, then the entire expression becomes zero

9. Estimate probabilities for continuous attribute

**A white paper with black text

Description automatically generated**

A math equation with black text

Description automatically generated

10. Methods for estimating a classifier’s accuracy:

Holdout, random subsampling, cross-validation, bootstrap

11.

A black and white box with black text

Description automatically generated

12. Evaluation metric:

**Accuracy = (TP + TN)/All Error rate = (FP + FN)/All**

(Class imbalance problem)

**Sensitivity = TP/P** **Specificity = TN/N**

A screenshot of a math problem

Description automatically generated

ROC:A graph of different colored lines

Description automatically generated

A white and blue text on a white background

Description automatically generated

13. Ensemble method:

Ensemble Methods work better than a single base classifier if:

All base classifiers are independent of each other

All base classifiers perform better than random guessing (error rate < 0.5 for binary classification)

14. Bagging:

Bootstrap sampling: sampling with replacement

Build classifier on each bootstrap sample

Add final classification/ majority vote

15. Boosting

An iterative procedure to adaptively change distribution of training data by focusing more on previously misclassified records

Initially, all N records are assigned equal weights (for being selected for training)

Unlike bagging, weights may change at the end of each boosting round, wrong classification increase weight, vice versa

16. Random Forest

Construct an ensemble of decision trees by manipulating training set as well as features

Use bootstrap sample to train every decision tree (similar to Bagging)

Use the following tree induction algorithm:

At each node use a random selection of attributes as candidates and split by the best attribute among them

Repeat this procedure until all leaves are pure (unpruned tree)

17. Ensemble methods try to reduce the variance of complex models (with low bias) by aggregating responses of multiple base classifiers

A white text with black text

Description automatically generated