University of Colorado
Boulder

CSCI 4502/5502
Data Mining

Fall 2020
Lecture 16 (Oct 22)

Homework Assignments

- Homework 1 to 4: graded, check scores in Canvas
- Homework 5: due at 9:30am, Thursday, Oct 22
  - will be graded soon
- Scores statistics and common mistakes
  - will be posted in Canvas soon

Midterm Exam (1)

- Availability survey: Tuesday, Oct 20
  - check for exam schedule confirmation from TA
- Midterm review: Thursday, Oct 22
- Practice exam review: Tuesday, Oct 27
  - practice exam posted in Canvas (Proctorio & PDF)
  - in-person (wait for confirmation) & remote/online

Midterm Exam (2)

- Closed-book exam
- Formula will be provided
- No coding, no calculator
  - write out calculation steps
- Confirmation from TA
  - exam time & mode
- Exam time
  - 9:35am-10:50am (Oct 29)
  - 7:00pm-8:15pm (Oct 29)
  - extra slot (TBD)
- Exam mode: Proctorio; manual proctoring; in-person (class time only)

Office Hours

- Instructor
  - Th Oct 22: 7pm-8pm
  - Tu Oct 27: 11am-12pm
  - Tu Oct 27: 7pm-8pm (moved from Oct 29)
- TA:
  - M Oct 26: 8am-9am
  - W Oct 28: 4pm-5pm
- GSS
  - W Oct 28: 10am-11am

Midterm Review

- Chap 1: Introduction
- Chap 2: Getting to know your data
- Chap 3: Data Preprocessing
- Chap 4 & 5: Data Warehouse, Data Cube
- Chap 6 & 7: Frequent Pattern Mining
- Chap 8 & 9: Classification
- Chap 10 & 11: Cluster Analysis

Chap 1: Introduction

- Why data mining?
  - data explosion (generation & sharing)
  - data rich but information poor
- Data mining (knowledge discovery from data)
  - automated analysis of massive data
  - quality vs. efficiency
  - interesting patterns: valid, novel, potentially useful, ultimately understandable by human

Data Mining: Various Views

- Data view
  - kinds of data to be mined
- Knowledge view
  - kinds of knowledge to be discovered
- Method view
  - kinds of techniques utilized
- Application view
  - kinds of applications adapted

Chap 2: Getting to Know Your Data

- Data objects and attribute types
  - nominal, binary, ordinal, numeric
  - interval-scaled, ratio-scaled
  - discrete vs. continuous
- Measuring data similarity and dissimilarity
  - data matrix vs. dissimilarity matrix
  - proximity measures for nominal, binary, ordinal, numeric attributes

Descriptive Summarization

- Basics: N, min, max
- Central tendency
  - mean, median, mode, midrange
- Dispersion
  - quartiles, IQR, five number summary
  - variance, standard deviation
- Graphic displays: box plot, histogram, quantile plot, quantile-quantile plot, scatter plot

Chap 3: Data Preprocessing

- Why preprocessing data?
- Data cleaning
- Data integration
- Data reduction
- Data transformation and discretization

Why Preprocessing Data?

- Imperfect data
  - incomplete: missing attributes, values
  - noisy: errors, outliers
  - inconsistent: discrepancies
- Measure of data quality
  - accuracy, completeness, consistency, timeliness, believability, value added, interpretability, accessibility

Data Cleaning

- Missing data
  - ignore, manual
  - automatic: constant, attribute mean, class attribute mean, most probable
- Noisy data
  - binning: bin mean, median, boundary
  - regression
  - clustering

Correlation Analysis

- Correlation coefficient (numerical data)

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

- X^2 (chi-square) test (categorical data)

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

- Correlation does not imply causality!

Data Reduction

- Data reduction: Why? Goal?
- Dimensionality reduction
  - attribute subset selection; Wavelet transform; principle component analysis (PCA)
- Numerosity reduction
  - regression, log-linear models
  - data cube aggregation
  - histograms, clustering, sampling

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization
  - e.g., daily sales => monthly, annual sales
- Generalization: concept hierarchy climbing
  - e.g., street => city => state
- Normalization: scale to fall within a range
  - min-max, z-score, decimal scaling

Chap 4: Data Warehouse

- Separate data store for information processing
- Characteristics
  - subject-oriented
  - integrated
  - time-variant
  - nonvolatile
- OLTP vs. OLAP
  - day to day operation
  - vs. decision support

Chap 4 & 5: Data Cube

- Multi-dimensional data model
- Dimensions, facts
- Schema: star, snowflake, fact constellation
- Typical operations
- roll-up, drill-down, slice and dice, pivot, drill-across, drill-through
- Cuboid cells, materialization of data cube
- full, partial, or no materialization

Chap 6 & 7: Frequent Pattern Mining

- Basic concepts, roadmap
- Mining frequent itemsets
- Mining association rules
- Correlation analysis
- Constraint-based association mining

Mining Frequent Itemsets

- Apriori algorithm
  - k-itemsets, candidate (k+1)-itemsets
- Challenges
  - #scans, #candidates, support counting
- Partition, sampling, dynamic itemset counting, hash-tree
- FP-growth: FP-tree, conditional pattern base, conditional FP-tree
- Vertical data format

Associations & Correlations

- Association: Support & confidence
- Correlation rule
  - A ⇒ B [support, confidence, correlation]
  - lift, x^2, null-invariant

$$lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)}$$

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \qquad e_{ij} = \frac{count(A=a_i) \times count(B=b_j)}{N}$$

Rule (Pattern) Constraints

- Metarule
- Anti-monotonicity
- Monotonicity
- Succinctness (skip)
- Convertible constraints
  - with proper ordering of items
- Strongly convertible (skip)
- Inconvertible (skip)

Chap 8: Classification

- Basic concepts
- Decision tree induction
- Bayesian classification
- Rule-based classification (skip)
- Model evaluation and selection
- Improve classification accuracy
- Summary

Classification vs. Prediction

- Classification: categorical class labels
- Prediction: continuous-valued functions
- Training vs. testing
- Supervised vs. unsupervised
- Evaluation criteria
  - accuracy, speed, robustness, scalability, interpretability, goodness of rules

Decision Tree Induction

- Top-down, recursive, divide-and-conquer
- Attribute selection
  - information gain
  - gain ratio
  - gini index
- Attribute split
  - discrete, continuous, discrete-binary

Bayesian Classification

- Bayes' Theorem

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

- Naive Bayesian classifier

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i)$$

Classifier Accuracy Measures

- Accuracy, error rate
- Confusion matrix
- Costs and benefits of TP, TN, FP, FN
- Sensitivity, specificity, precision
- Predictor error measures
  - error, square error
  - absolute, mean, relative

Classifier/Predictor Evaluation

- Holdout, random sampling
- Cross-validation
  - k-fold stratified cross-validation
- Bootstrapping
  - sample with replacement
  - .632 bootstrap
- Model selection
  - t-test, ROC curves

Ensemble

- Ensemble
  - combination of multiple models
  - bagging: majority vote (equal weight)
  - boosting (e.g., Adaboost): weighted

Chap 9: Advanced Classification

- Bayesian belief networks
- Backpropagation
- Support vector machines
- Lazy learning (skip)
- Other classification methods (skip)
- Additional topics regarding classification (skip)
- Summary

Chapter 10: Cluster Analysis

- Basic concepts
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods (skip)
- Evaluation of clustering (skip)
- Summary

Cluster Analysis

- Unsupervised learning
- Intra/inter-cluster similarity
- Requirements
  - scalability, various data types, arbitrary shape, minimal domain knowledge, noisy data, incremental, high dimensionality, constraint-based, interpretability

Clustering Methods

- Partitioning methods
  - k-means, k-medoids
- Hierarchical methods
  - agglomerative, decisive
  - BIRCH, CHAMELEON
- Density-based methods
  - DBSCAN, DENCLUE
- Grid-based methods
  - STING, CLIQUE (skip)

- Chapter 11: Advanced Cluster Analysis
  - probabilistic model-based clustering
  - clustering high-dimensional data (skip)
  - clustering graph and network data (skip)
  - clustering with constraints (skip)

Reminder

- Practice exam review: Tuesday, Oct 27
- Midterm exam: Thursday, Oct 29
- Exam schedule confirmation from TA