CSE 5243 INTRO. TO DATA MINING

Review Session for Midterm

Yu Su, CSE@The Ohio State University

Notes

- Time: 02/26/2020 (Wed), 9:35 – 10:55 AM
- Location: Caldwell Lab 171
- One-page cheat sheet: both sides allowed
- Calculator: allowed
- I will update the slides over the weekend to make them cleaner. No major changes.

Agenda

- Summary of key concepts and equations
- HW1 Discussion (TA)
- HW2 Discussion (TA)

Probability and Statistics

- Bayes rule: prior, likelihood, marginal probability, posterior

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- Chain rule

$$p(x_1, \ldots, x_n) = p(x_n|x_1, \ldots, x_{n-1})p(x_1, \ldots, x_{n-1})$$

$$= p(x_n|x_1, \ldots, x_{n-1})p(x_{n-1}|x_1, \ldots, x_{n-2})p(x_1, \ldots, x_{n-2})$$

$$= p(x_1)\prod_{i=2}^{n} p(x_i|x_1, \ldots, x_{i-1})$$

Probability and Statistics

- Bayes rule: prior, likelihood, marginal probability, posterior

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- Chain rule
- Maximum Likelihood Estimation (MLE)
  - Obtain parameter estimates that maximize the probability that the sample data occurs for the specific model.

$$L(\Theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \Theta)$$

Data Preprocessing

- Major tasks: cleaning, integration, reduction, and transformation
- Cleaning: Smoothing noisy data by binning

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Data Preprocessing

- Major tasks: cleaning, integration, reduction, and transformation
- Cleaning: Smoothing noisy data by binning
- Integration: Detecting redundant attributes by correlation analysis
  - χ^2 test for discrete random variables

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

Data Preprocessing

- Major tasks: cleaning, integration, reduction, and transformation
- Cleaning: Smoothing noisy data by binning
- Integration: Detecting redundant attributes by correlation analysis
  - χ^2 test for discrete random variables
  - Correlation/covariance for continuous random variables

Data Preprocessing

- Major tasks: cleaning, integration, reduction, and transformation
- Cleaning: Smoothing noisy data by binning
- Integration: Detecting redundant attributes by correlation analysis
- Reduction: Types of data reduction methods
  - Regression, sampling, histogram, dimensionality reduction, clustering

Data Preprocessing

- Major tasks: cleaning, integration, reduction, and transformation
- Cleaning: Smoothing noisy data by binning
- Integration: Detecting redundant attributes by correlation analysis
- Reduction: Types of data reduction methods
- Transformation
  - Normalization: min-max, z-score, L2 norm
  - Discretization: general concept

Classification

- Decision tree
  - How to construct a decision tree given a dataset
  - Attribute selection measures: information gain, gain ratio, Gini index
  - Categorical attribute vs. continuous attribute
  - What is pruning and why?

Classification

- Decision tree
- Classifier evaluation
  - Metrics: confusion matrix/accuracy/error rate/precision/recall/F-measure/ROC curve
  - Methods: Holdout/cross validation

Classification

- Decision tree
- Classifier evaluation
- Practical issues: overfitting/underfitting
  - Concepts
  - What could cause that? How to detect? How to fix?

Classification

- Decision tree
- Classifier evaluation
- Practical issues: overfitting/underfitting
- Naïve Bayes classifier (zero-probability problem)
- Ensemble methods: general concepts. Why ensemble often improves performance?
- K-nearest neighbor classifier
- Neural network and SVM
  - general concepts, e.g., what are support vectors? What is maximum marginal hyperplane? What is back propagation?

Clustering

- Distance and similarity measures (in the Statistics review lecture)
- Partitioning-based
  - K-means: algorithm, objective, complexity
  - K-medoids
- Hierarchical clustering
  - Dendrogram
  - MIN (single linkage), MAX (complete linkage)
- Density-based
  - DBSCAN: general concepts like core/border/noise, density-reachable/connected
- Cluster evaluation
  - Similarity matrix, silhouette coefficient

Good luck!