Data Mining
Topics, 2024

1. Definition of data mining. Motivations, examples. Origins of data mining. Prediction and description, data mining tasks. Challenges of data mining. Exploring data: Summary statistics and visualization.
   (Slides: Introduction + Exploring Data)

2. Data analysis pipeline, the 5-step process. Data, types of data and types of attribute. Types of data sets. Quality of data. Preprocessing: aggregation, sampling, dimension reduction (PCA), feature subset selection, feature creation, discretization and binarization, attribute transformation.
   (Slides: Data preprocessing)

3. Classification: Basic concepts. Decision trees. Tree induction algorithms: Hunt, CART, CHAID, C4.5. Model evaluation: confusion matrix, metrics, ROC curve.
   (Slides: Classification: Basic Concepts, Decision Trees, and Model Evaluation)

4. Classification techniques I: Rule-based classifiers, Nearest neighbor classifiers (k-NN), Bayesian classifiers.
   (Slides: Classification: Alternative Techniques)

5. Classification techniques II: Artificial neural networks (ANN), Support vector machines (SVM), Logistic regression, Ensemble methods (bagging and boosting).
   (Slides: Classification: Alternative Techniques)

6. Similarity and dissimilarity (distance). Clustering: definition, K-means algorithm and its variants. Cluster validity: similarity matrix, correlation, SSE, silhouette coefficient.
   (Slides: Data preprocessing + Cluster Analysis: Basic Concepts and Algorithms)

7. Clustering: Hierarchical and density based methods. Agglomerative clustering algorithm: single, complete and average link. Dendrogram. DBSCAN.
   (Slides: Cluster Analysis: Basic Concepts and Algorithms)

8. Market-basket data. Frequent itemset, support. Apriori principle and Apriori algorithm. Candidate generation. Other algorithms. Confidence. Association rule mining. Pattern evaluation: statistical based measures, lift value.
   (Slides: Association Analysis: Basic Concepts and Algorithms)

9. Anomaly detection: definition, problems, applications. Anomaly detection schemes: graphical and statistical-based, distance-based, model-based. Base rate fallacy.
   (Slides: Anomaly Detection)

Debrecen, 20 of May, 2024

Dr. Márton Ispány
Professor