7CCSMDM1 Data Mining
Exam Revision Note
Dimitrios Letsios

During the term, we cover a range of (1) Data Mining tasks, (2) approaches for implementing these tasks, notably statistical concepts, models, and algorithms, and (3) approaches for evaluating the various Data Mining outcomes. The following table summarizes the key elements that we cover during the term to facilitate your preparation for the final exam. A key learning outcome of the module is to develop understanding and applied skills that will enable you to use these approaches in real contexts. You may want to study the various concepts, how the various data mining approaches (statistical components, models, algorithms) operate, and ways that they can be used in practice.

Task	**Data Mining Approach (Statistical Concepts, Models, Algorithms)**	**Evaluation Approach**
Regression & Classification: (Non-) Linear Models	- Linear Regression - Gradient Descent algorithm - Linear Classification, Logistics Regression - Multilayer Perceptron Neural Networks	Mean Squared Error Mean Absolute Error Cross-validation Regularization**
Regression & Classification: Rule-/Tree-Based Models	- Enumeration algorithm for Classification Rules - Decision Trees - Entropy-based Divide-and-Conquer algorithm - Gradient-Boosted Trees and Random Forests	Bootstrapping Confusion Matrix ROC Curve
Cluster Analysis	- Centroid-based clustering via Kmeans - Density-based clustering via DBSCAN - Hierarchical clustering and dendrograms - Outlier detection	Within-cluster distances Between-cluster distances Silhouette coefficient
(Density) Estimation	- Probabilistic models, e.g. Binomial, Gaussian. - Maximum Likelihood Estimation. - Gaussian Mixture Models and EM algorithm. - Naïve Bayes Classification	Likelihoods Posterior probabilities
Instance-Based Learning	- k-Nearest Neighbors (kNN) and kD-trees - Histogram and Kernel Density Estimation - Locally Weighted Linear and Kernel Regression - Support Vector Machines	Bias-Variance Trade-off
Text Classification	- Text preprocessing operations - Bag-of-words - Multinomial Naïve Bayes - Language models	Perplexity
Information Retrieval	- TF-IDF - PageRank	Precision-Recall
Association Rule Mining	- Frequent Item-Set Enumeration - Apriori Algorithm - FP-trees & FP-Growth Algorithm	Support Confidence Lift
Time Series Analysis	- Temporal Feature Engineering - Classic Decompositions and Visualizations - ARIMA Modelling**	Stationarity**
Feature Selection	- Basic Feature Engineering Operations - Filtering with Statistical Dependency Measures, e.g. Correlation, Average Mutual Information - Forward and Backward Sequential Selection	Variable Dependencies & Redundancies
Dimensionality Reduction	- Variable transformations and projections. - Principal Component Analysis (PCA) - Singular Value Decomposition (SVD)**	Reconstruction Error