Dr. Christoph F. Eick

Review List Midterm Data Mining COSC 6335

Tuesday, October 17, 2023, 2:30p

*Last updated: October 11, noon*

*Th*e exam will be “open books and notes” in **SEC 105**; the use of computers is not allowed, but calculators are okay; there will be no R-programming or other programming tasks in this exam; topics covered are:

1. \*\*\*\*\*\*\* Exploratory Data Analysis (class transparencies discussion of Chapter3 in the first edition of the textbook; capability to apply EDA to a problem at hand (questions are similar to Task1 centering on histograms, box plots, scatter plots, fitting a linear function to a dataset and statistical summaries))
2. \*\* Basics of correlation, linear regression, Normal distribution; additional reading material for this topics includes: <http://en.wikipedia.org/wiki/Correlation_and_dependence> . <http://en.wikipedia.org/wiki/Normal_distribution> , <http://en.wikipedia.org/wiki/Standard_score> ,

<https://en.wikipedia.org/wiki/68–95–99.7_rule>

1. \*\*\*\* Introduction to Density Estimation (lecture transparencies, <https://en.wikipedia.org/wiki/Multivariate_normal_distribution>, [Kernel density estimation - Wikipedia](https://en.wikipedia.org/wiki/Kernel_density_estimation) )
2. \*\* Similarity Assessment (Class Transparencies, capability to design a distance function for a dataset)
3. \*\*\*\*\*\*\*\*\*\*\*\*\* Clustering centering on objectives, cluster validity (only overview, correlation-based methods and silhouette are relevant for the exam), K-means, PAM, DBSCAN, DENCLUE (lecture transparencies, DENCLUE 2.0 paper, textbook pages 525-531, 534-542, 549-553, 565-574, 579-584)
4. \*\*\*Outlier Detection and Anomaly Detection
   1. Covered transparencies
   2. Book pages 703-711, 719-724
5. \*\*\* Decision tree induction algorithm

You should have detailed knowledge concerning the following algorithms, concepts and procedures: Apply density estimation techniques to an example; K-Means, PAM, DBSCAN and DENCLUE; decision tree induction algorithm.

The exam will not cover data storytelling and the introduction to data mining; these topics will be covered in Dec. 7 final exam.

The midterm exam counts 21% towards the overall course grade and should take about 75-80 minutes.