Dr. Christoph F. Eick

Review List Midterm1 Exam Data Mining COSC 4335

Monday, October 1, 2018, 2:30-3:50p in our classroom

*Last updated: September 24, 9a*

The exam will be “open books and notes” (but use of computers & internet is **not** allowed) and will center on the following topics (at least 85% of the questions will focus on material that was covered in the lecture); there will be no R-programming in this exam:

1. \*\*\*\*\*\*\* Exploratory Data Analysis (class transparencies including “interpreting displays” and discussion of Chapter3 in the first edition of the textbook; capability to apply EDA to a problem at hand (similar to Assignment 1 centering on histograms, box plots, scatter plots and statistical summaries))
2. \*\*\*\*\*\*\* Clustering and Similarity Assessment(covered class transparencies and textbook[[1]](#footnote-1) pages 76-80, 83-84, 97-98, 525-538, 565-570 of the Tan/Steinback/Kumar book.) . Only the discussions of clustering in general and k-means and k-medoids/PAM and DBSCAN[[2]](#footnote-2) are relevant for the exam. For k-medoids/PAM study the class transparencies, but also read

<https://en.wikipedia.org/wiki/K-medoids> as k-medoids is not discussed in the textbook.

1. \*\*\* Basics of correlation, linear regression, Normal distribution, density estimation, and decision trees; 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>

<http://en.wikipedia.org/wiki/Density_estimation> .

1. \*\* Introduction to Data Mining

a. Transparencies covered in the first 2 lectures

b. Textbook pages 4-13

You should have detailed knowledge concerning the following algorithms and measures: PAM/K-medoids (not covered in the textbook), K-means, DBSCAN, SSE; be able to define a distance function for a specific example.

Relevant Slides: I Introduction to Data Mining [Part1](http://www2.cs.uh.edu/~ceick/UDM/DM-Part1a.pptx), [Part2](http://www2.cs.uh.edu/~ceick/UDM/DM-Part1b.pptx), [Differences between Clustering and Classification](http://www2.cs.uh.edu/~ceick/UDM/claclu.pptx)).  
II [Exploratory Data Analysis](http://www2.cs.uh.edu/~ceick/UDM/DM-Part2.pptx) (updated on August 29, 2018; covers [chapter 3 from the the First Edition of the Tan Book](http://www2.cs.uh.edu/~ceick/UDM/DA_Tan.pdf) (download as this material is not in the second edition); see also [Interpreting Displays](http://www2.cs.uh.edu/~ceick/UDM/dm_interpreting.pptx);   
IV Clustering and Similarity Assessment ([Introduction](http://www2.cs.uh.edu/~ceick/UDM/dm_clustering1.pptx) and [~~Hierachical Clustering~~ and DBSCAN](http://www2.cs.uh.edu/~ceick/UDM/dm_clustering2b.pptx);

Midterm1 will count about 16% towards the overall course grade.

1. All pages numbers refer to the second edition of the textbook~ [↑](#footnote-ref-1)
2. Hieararchical Clustering is not relevant for the midterm1 exam! [↑](#footnote-ref-2)