Dr. Christoph F. Eick

Review List Midterm1 Exam Data Mining COSC 4335

Thursday, March 3, 2015, 11:30-1p in T 120K

*Last updated: Fr. , Feb. 26, 4p*

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):

1. \*\*\*\*\*\* Exploratory Data Analysis (class transparencies including “interpreting displays” and and text book pages 97-131; 76bottom-78; 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 pages 65-78, 81-83, 487-506, 515-522, 525 of the Tan/Steinback/Kumar book.) Also look at: <https://en.wikipedia.org/wiki/Hierarchical_clustering> and see below
3. \*\*\* 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> , <http://en.wikipedia.org/wiki/Density_estimation> .

4. \*\*\* Introduction to Data Mining

a. Transparencies covered in the first 3 lectures

b. Textbook pages 19-36, 47-48

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

Relevant Slides:I Introduction to Data Mining ([COSC 4355 Knowledge Sources](http://www2.cs.uh.edu/~ceick/UDM/Cosc4335-Knowledge_Sources.docx), [Part1](http://www2.cs.uh.edu/~ceick/UDM/DM-Part1a.pptx), [Part2](http://www2.cs.uh.edu/~ceick/UDM/DM-Part1b.pptx), [Part3: Data](http://www2.cs.uh.edu/~ceick/UDM/DM-Part1c.pptx), [Differences between Clustering and Classification](http://www2.cs.uh.edu/~ceick/UDM/claclu.pptx) --- covers chapter 1 and Section 2.1)
II [Exploratory Data Analysis](http://www2.cs.uh.edu/~ceick/UDM/DM-Part2.pptx) (covers chapter 3 in part; see also [Interpreting Displays](http://www2.cs.uh.edu/~ceick/UDM/dm_interpreting.ppt); [Some R Data Analysis Functions I](http://www2.cs.uh.edu/~ceick/UDM/UDM-R1.pptx); [Some R Data Analysis Functions II](http://www2.cs.uh.edu/~ceick/UDM/UDM-R2.pptx))
III R ([Can Cao's R Tutorial](http://www2.cs.uh.edu/~ceick/UDM/Cao_R_Tutorial.pdf) (used in lab on Feb. 2, 2016), [Arko's Short Intro Into R](http://www2.cs.uh.edu/~ceick/UDM/Arko_R_Sh_Intro.pptx), [Scatter Plot Code](http://stats.stackexchange.com/questions/30788/whats-a-good-way-to-use-r-to-make-a-scatterplot-that-separates-the-data-by-trea), [Decision Trees in R](http://www2.cs.uh.edu/~ceick/UDM/DecisionTreesinR.pptx), [Some useful code for Project1](http://www2.cs.uh.edu/~ceick/UDM/bna.r), [Functions and Loops in R](http://www2.cs.uh.edu/~ceick/UDM/Functions%20and%20Loops%20in%20R.pptx), [Directory containing R-code for Project2](http://www2.cs.uh.edu/~ceick/UDM/RExamplesP2) (lecture on Feb. 26)
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); R-scripts demonstrating: [K-means/medoids](http://www2.cs.uh.edu/~ceick/UDM/r-clustering.r), [DBSCAN](http://www2.cs.uh.edu/~ceick/UDM/dbscan.r); [Clustering Exercises K-Means, HC, and DBSCAN](http://www2.cs.uh.edu/~ceick/UDM/Clustering_Exercises1.pdf))

The Midterm1 will **not** cover R-programming—you will not be asked to write any R-code; however, you should have a clear understanding about the R-functions you used to solve the tasks of Assignment1.

Midterm1 will count about 16-19% towards the overall course grade.