Dr. Christoph F. Eick

Review List Midterm1 Exam DS 1 COSC 3337

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

*Last updated: September 25, 11a*

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, density plots and statistical summaries))
2. \*\*\* Classification and decision trees (mostly centering on basics, decision tree induction algorithm, and discussion of “greedy” algorithms) Slides and Tan second edition text book pages 113-134
3. \*\*\* Basics of correlation, attribute normalization, 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](https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule)

1. \*\*\*\* Non-Parametric Density Estimation slides and <http://en.wikipedia.org/wiki/Density_estimation> .
2. \*\*\*\* Data Visualization slides and <https://eagereyes.org/basics/banking-45-degrees> and

<https://thedoublethink.com/tuftes-principles-for-visualizing-quantitative-information/>

 and <https://vwo.com/blog/crap-design-principles/>

1. \*\* Introduction to Data Mining

a. Transparencies covered in the first 2 lectures

b. Textbook pages 4-13

You should have detailed knowledge how non-parametric density estimation and the decision tree induction algorithm works.

Relevant Slides:

I Introduction to Data Mining/Data Science ([Part1: Introduction to Data Mining](http://www2.cs.uh.edu/~ceick/UDM/DS1-Part1a.pptx), [Part2: Course Information](http://www2.cs.uh.edu/~ceick/UDM/DS1-Part1b.pptx)).
II [Exploratory Data Analysis](http://www2.cs.uh.edu/~ceick/UDM/DS1-Part2.pptx) (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 and other material); [Interpreting Displays](http://www2.cs.uh.edu/~ceick/UDM/dm_interpreting.pptx); [Introduction to Non-Parametric Density Estimation](http://www2.cs.uh.edu/~ceick/UDM/Non-ParametricDE.pptx); [KDE Density Functions](http://www2.cs.uh.edu/~ceick/UDM/Non-Parametric%20Density%20Functions.pptx))

III R ([Decision Trees in R](http://www2.cs.uh.edu/~ceick/UDM/DecisionTreesinR.pptx))

IV Introduction to Data Visualization ([Part1](http://www2.cs.uh.edu/~ceick/UDM/COSC3337-DV1.pdf), [Part2](http://www2.cs.uh.edu/~ceick/UDM/COSC3337-DV2.pdf))

V Classification ([Introduction to Classification: Basic Concepts and Decision Trees](http://www2.cs.uh.edu/~ceick/UDM/dm_classification1.pptx) (only those slides that were covered in the lecture!))

Midterm1 will count about 15% towards the overall course grade.