Dr. Christoph F. Eick

Review List Midterm1 Exam Data Mining COSC 6335

Tuesday, October 13, 2:30p in Blackboard

*Last updated: October 7, 2020, 1p*

The exam will be “open books and notes”; there will be no R-programming in this exam:

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 (similar to Task1/2 ProblemSet1 centering on histograms, box plots, scatter plots 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. \*\*\*\*\*\*\*\*Decision Trees, and General Topics for Classification including overfitting (covered class transparencies and textbook[[1]](#footnote-1) pages 117-156 (skip 3.3.5) and 162(starting with 3.5.4)-170
2. \*\*\*\*\*\*SVM (class transparencies, <http://en.wikipedia.org/wiki/Kernel_method> , and pages 276-296
3. \*\*\*Nearest Neighbor Classifiers (class transparencies, textbook pages 208-212)
4. \*\*\*Association Rule Mining centering on basics and the APRIORI algorithm (class transparencies, textbook pages 357-375, 377-384

You should have detailed knowledge concerning the following algorithms and measures: Decision Tree Induction Algorithm, gain computations for decision trees, APRIORI Algorithm, how kNN and SVM classify examples.

Relevant Slides and other reading material:

Slides of Groups C, D, E, F (can be found in the DM2000 Teams respective channels)

II 2020 [Exploratory Data Analysis](http://www2.cs.uh.edu/~ceick/DM/DM-Part2.pptx) (covers [chapter 3 of the first edition of the Tan book](http://www2.cs.uh.edu/~ceick/DM/Tan_Chapter3.pdf))  
III R ( [Decision Trees in R](http://www2.cs.uh.edu/~ceick/DM/DecisionTreesinR.pptx),  [Computing Statistical Summaries In the Presence of Missing Value (NA)](http://www2.cs.uh.edu/~ceick/DM/NA.r))

V 2020 Classification ([Introduction to Classification: Basic Concepts and Decision Trees](http://www2.cs.uh.edu/~ceick/DM/dm_classification1.pptx), [Overfitting](http://www2.cs.uh.edu/~ceick/DM/Overfitting.pptx), , [kNN-Classifiers and Support Vector Machines](http://www2.cs.uh.edu/~ceick/DM/KNN+SVM.pptx))  
VI Association Analysis: Rule, Sequence, Graph and Collocation Mining ([Part1](http://www2.cs.uh.edu/~ceick/DM/AA1.pptx)) centering only on APRIORI algorithms and Association rule mining basics.

Midterm1 will counts 18-22% towards the overall course grade.

1. All page numbers refer to Second Edition of the Textbook [↑](#footnote-ref-1)