Dr. Christoph F. Eick

Review List Midterm Exam DS 1 COSC 3337

Thursday, October 14, 11:30a-12:45p in F 160

*Last updated: October 6, 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 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. \*\*\* 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>

1. \*\*\*\*\*\*Decision Trees, and General Topics for Classification, particularly decision tree induction algorithm, overfitting, classification model performance evaluation (covered class transparencies and textbook[[1]](#footnote-1) pages 117-157 (skip 3.3.5) and 162(starting with 3.5.4)-169
2. \*\*\*\*SVM (class transparencies, <http://en.wikipedia.org/wiki/Kernel_method> , and pages text book 90-94 and 276-296
3. \*\*\*\*\* Neural networks (class transparencies, textbook pages 249-262); content of the  two introductory videos by 3blueonebrown about neural networks:  
   [Introduction to Neural Networks](https://www.bing.com/videos/search?q=neural+network+video&view=detail&mid=54402D363ABB8903202F54402D363ABB8903202F&FORM=VIRE) (watch the whole video)  
   [Weight Learning in Neural Networks](https://www.youtube.com/watch?v=IHZwWFHWa-w&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi&index=3&t=0s) (just watch the first 15 minutes of the second video)
4. \*\*Nearest Neighbor Classifiers (class transparencies, textbook pages 208-212)

Other relevant material: Group Homework Credit Slides for Groups A-F.

You should have detailed knowledge concerning the following algorithms and approaches: Decision Tree induction algorithm, information and GINI gain computations, SVM hyperplane approach, kernels (only basic ideas), architecture of neural networks and some basis understanding how neural networks learn models and how back propagation works; how the nearest neighbor classifiers work.

Relevant Lecture Slides:

II [Exploratory Data Analysis](http://www2.cs.uh.edu/~ceick/UDM/DS1-EDA.pptx) (covers [chapter 3 from the the First Edition of the Tan Book](http://www2.cs.uh.edu/~ceick/UDM/DA_Tan.pdf) )

IV Classification ([Introduction to Classification: Basic Concepts and Decision Trees](http://www2.cs.uh.edu/~ceick/UDM/dm_classification1.pptx), [Overfitting](http://www2.cs.uh.edu/~ceick/UDM/Overfitting.pptx), [kNN-Classifiers and Support Vector Machines](http://www2.cs.uh.edu/~ceick/UDM/oc1.pptx), [Neural Networks](http://www2.cs.uh.edu/~ceick/UDM/NN2021.pptx), )

The introduction to Data Mining/Data Science, covered in the first week of the semester will be relevant for the course final exam, but not for the midterm exam. The Fall 2021 Midterm will count about 20% towards the overall course grade.

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