Dr. Christoph F. Eick

Review List 2023 Midterm1 Exam DS 1 COSC 3337

Tuesday, October 3, 11:30a-12:45p in **SW 101**

*Last updated: Sept. 27, 8a*

*Missing COSC 3337 Exams:* If you miss a course exam for reasons that are not covered by the Undergraduate Excused Absence Policy or you missed the exam and did not follow the procedure outlined in this policy, you will get a grade of ‘F’ for the missed exam (for more details, see COSC 3337 Syllabus).

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 Task1 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. \*\*Nearest Neighbor Classifiers (class transparencies, textbook pages 208-212)
4. \*\*\*\* Naïve, Parametric and Non-parametric Density Estimation; additional reading material:

[Kernel density estimation – Wikipedia](https://en.wikipedia.org/wiki/Kernel_density_estimation) (KDE for short); only Definition, Example, Bandwidth Selection

[03Gaussiankernel.nb (wisc.edu)](https://pages.stat.wisc.edu/~mchung/teaching/MIA/reading/diffusion.gaussian.kernel.pdf.pdf) only 3.1 and 3.2

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), how nearest neighbor classifiers work; how non-parametric density estimation work (e.g. capability to apply it to an examples)

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) (download as this material is not in the second edition);

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),)

V Density Estimation ([Naive and Parametric Density Estimation](https://www2.cs.uh.edu/~ceick/UDM/PDE.pptx), [Non-parametric Density Estimation](https://www2.cs.uh.edu/~ceick/UDM/NPDE.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 midterm1 exam. There will be no programming tasks in the Midterm1 Exam!

Midterm1 counts approx. 14% towards the COSC 3337 course grade!

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