Dr. Christoph F. Eick

Review List Midterm Data Mining COSC 6335

Thursday, October 10, 2024, 2:30p

*Last updated: October 7, noon!*

*Th*e exam will be “open books and notes” in **SEC 203**; the use of computers is not allowed, but calculators are okay; there will be no R-programming or other programming tasks in this exam. There will be 3+versions, and your neighbors will likely have different versions of the Midterm Exam. Topics covered are:

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 (questions are similar to Task1 centering on histograms, box plots, scatter plots, fitting a linear function to a dataset 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. \*\*\*\* Introduction to Density Estimation (lecture transparencies, <https://en.wikipedia.org/wiki/Multivariate_normal_distribution>, [Kernel density estimation - Wikipedia](https://en.wikipedia.org/wiki/Kernel_density_estimation) )
2. \*\* Similarity Assessment (Class Transparencies)
3. \*\*\*\*\*\*\*\*\*\*\*\*\* Clustering centering on objectives, cluster validity and evaluation, K-means, PAM, DBSCAN, EM textbook pages 525-531, 534-542, 549-553, 565-574, 579-584)

Background material about EM and GMMs: [Gaussian Mixture Model Explained | Built In](https://builtin.com/articles/gaussian-mixture-model)

1. \*\*\*\*Outlier Detection and Anomaly Detection
   1. Covered transparencies
   2. Book pages 703-711, 719-724
2. \*\*\*\* Data Storytelling  [Data Storytelling](https://www2.cs.uh.edu/~ceick/DM/DataStorytelling.pptx) slides

Read: [Data Storytelling: The Essential Data Science Skill Everyone Needs (forbes.com)](https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/?sh=146220b252ad)

You should have detailed knowledge concerning the following algorithms, concepts and procedures: Apply density estimation techniques to an example; K-Means, PAM, DBSCAN, EM; compute silhouette of a clustering; correlation-base cluster evaluation.

The exam will not cover the introduction to data mining; these topics will be covered in in the second course exam

The midterm exam counts 23% towards the overall course grade and takes 75 minutes.