Lectures

Preliminary schedule of lectures
L1 Mon 4.9. Course logistics, Introduction, preprocessing
L2 Tue 5.9. Distance and similarity
L3 Mon 11.9. Dimension reduction (PCA, SVD), Clustering I
L4 Tue 12.9. Clustering II (K-representatives, hierarchical)
L5 Tue 19.9. Clustering III (spectral, clustering validation)
L6 Tue 26.9. Ass. mining I (frequency-based)
L7 Tue 3.10. Ass. mining II (statistical)
L8 Tue 10.10. Ass mining III (special tricks), Episode mining (guest)
week 42 no teaching
L9 Tue 24.10. Web mining, recommender systems
L10 Tue 31.10. Graph mining, social networks
L11 Tue 7.11. Data randomization (guest)
L12 Tue 14.11. Text mining
L13 Tue 28.11. Recap

During lectures, we will use presemo (for anonymous questions and sometimes polls). You can find the session here:
https://presemo.aalto.fi/mdm2023

Lecture 1 (4.9. 2023)
course logistics slides
introduction to DM slides
data preprocessing slides
book Ch 1 and Ch 2 except 2.4.3–2.4.4

Lecture 2 (5.9. 2023)
distance and similarity slides
A note on metrics (includes a proof showing that fractional Lp norms are not metrics) slides
book Ch 3

Lecture 3 (11.9. 2023)
Dimension reduction with PCA and SVD slides
Clustering I (clustering tendency) slides
book Sec. 2.4.3, 6.1-6.2

Lecture 4 (12.9. 2023)
Clustering II (K-representatives, hierarchical clustering) slides1 slides2
book Sec. 6.3-6.4, 7.2.1
External material (these are optional, if you understand the main idea from the slides):
Tim Löhr: K-Means Clustering and the Gap-Statistics
For other methods to decide the number of clusters see wikipedia article
During lectures, we'll watch two videos:
"K-means clustering: how it works" (7.5 min) by Victor Lavrenko https://www.youtube.com/watch?v=_aWzGGNrcic
"Hierarchical Clustering - Fun and Easy Machine Learning" (10min) by Augmented Startups https://www.youtube.com/watch?v=EUQY3hL

Lecture 5 (19.9. 2023)
1) Spectral clustering slides video
book 2.4.4.3, 6.7, 6.10, 19.3.4. Recommended external material: von Luxburg (2007): A Tutorial on Spectral Clustering. Statistics and Computing, vol. 17, pp. 395–416, 2007. Reading guide.
2) Clustering validation slides video
book Ch 6.9. External material: Halkidi et al. (2002): Cluster Validity Methods: Part I. ACM SIGMOD Record 31(2): 40–45.
https://doi.org/10.1145/565117.565124 (you can access the paper in Aalto network or using ACM e-library through Aalto, but the pdf is also in Section Extra material, with restricted access to course participants)

Lecture 6 (26.9. 2023)
Association discovery 1 slides video1 video2 (slides not visible)
Reading: Hämäläinen and Webb: A tutorial on statistically sound pattern discovery, 2019, section 2.2 (Definitions of statistical dependence and association patterns, measuring strength of association)
Book: 4.1-4.4.2, 4.4.3 overview and 4.4.3.1, 4.6.3, 4.7, 5.2.1-5.2.2 (you can skip 4.4.2.1 on hash trees and the rest of 4.4.3 will be covered later)

Lecture 7 (3.10. 2023)
Association discovery 2 slides video1 video2 (The funny p-value video available in Extra material)
Reading: Hämäläinen and Webb: A tutorial on statistically sound pattern discovery, 2019, Sections 3.1, 4.1, 4.4.
Properties of statistical association rules explained with the Mega Party example (detailed explanation of the examples covered in the lecture)
It suffices to know the algorithm/pruning ideas in the scope presented in lectures, but if you want to deepen your knowledge, you can optionally read from Hämäläinen: Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. Knowledge and Information Systems: An International Journal 32(2):383-414, 2012. Sections 1-3 (overview, basic concepts), 4.1-4.2 (pruning the search space), and the main idea from 5.1 (general branch-and-bound strategy for statistical association rule mining). The simulation in the lecture is explained in Sec 5.2

Lecture 8 (10.10. 2023)
Association discovery 3 slides video1
Guest lecture: Juho Rinta-Paavola: Episode mining slides video2
Reading: Book 4.4.3.2-4.4.3.3 (it suffices to get the main idea of database projections and vertical counting). The FP-tree is extra-course material but if you are interested, you read more in Sec 4.4.4.3. On specious associations and Simpson's paradox, it suffices to know the idea at the level presented in the lectures, but if you are interested you can find further reading in section Extra material.
Material on episode mining?

Lecture 9 (24.10. 2023)
Web mining and recommender systems slides1 slides2 example (presented in the lecture) video1 video2
Reading: Book 17.1 (overview of graph mining tasks), 18.1, 18.3-18.5, 18.7.

Lecture 10 (31.10. 2023)
Graph mining and overview of Social network analysis slides1 slides2 video1 (about 10min missing from the end) video2
Reading for Graphs: Book 17.2 (from 17.2.3.2 only p. 567 before edit dist. alg.) 17.3 (from 17.3.2 only the main idea and Wiener index, from 17.3.3 only preview),17.4, 17.5, 17.7.
Reading for Social networks: Book 19.1, 19.2 preview-19.2.3, 19.2.5 (only measures for undirected graphs), 19.3 preview,19.5 preview-19.5.3, 19.7.
Note: There is no reading next week, so you can divide the reading on two weeks.

Lecture 11 (7.11. 2023)
Guest lecture: Heikki Mannila: Data randomization for assessing the results of data mining slides video1 video2

Lecture 12 (14.11. 2023)
Text mining slides video (note: this is from MDM2021, so just ignore the two announcements on the first slide)
Reading: Book 31.1-13.2, 13.3 preview-13.3.1, 13.3.3, 13.5.1.1-13.5.1.2, 13.6 (only the main idea), 13.7
External material: collocations (wikipedia article, only the general idea)
Note: The lecture will be arranged using a video recording but there will be a zoom session for online questions during lecture 16:15-18:00. You can follow the video and zoom in the lecture hall or anywhere you like. Zoom https://aalto.zoom.us/j/65280378922.

Lecture 13 (28.11. 2023)
Recap lecture. Program: 1) How to prepare for the exam, 2) Main sources of errors and misleading results in each step of the DM process, 3) recapping wish topics (please, fill the questionnaire soon).

Wishes for the recap lecture

Next section ►
Prerequisite questionnaire