Syllabus “Data Mining and Machine Learning” (DM868-DM870-DS804)

Arthur Zimek

January 2023

This syllabus gives an overview of topics in “Data Mining and Machine Learning” (DM868-DM870-DS804) along with pointers to related chapters in textbooks and some original papers. Note that these pointers are just given as alternatives, and that no textbook or paper reading is *required* for the course. The hints are only given to point to *possible* readings to go deeper after the lecture or to prepare before the lecture, if desired by the individual student.

Knowledge Discovery in Databases

- overview
- KDD process model [Fayyad et al., 1996]

Frequent Pattern Mining

- Frequent itemset problem and association rules
- APRIORI algorithm [Srikant and Agrawal, 1996]

some related textbook chapters:

- Zaki and Meira Jr. [2020], Ch. 8+9
- Tan et al. [2020], Ch. 4

Feature Spaces, Distance Measures

- features
- LP norms and related distance measures
- feature spaces
- example feature spaces
  - color space for images
  - term frequency vectors for documents

some related textbook chapters:

- Zaki and Meira Jr. [2020], Ch. 2+3
- Tan et al. [2020], Ch. 2

Clustering basics

- families of clustering approaches
- partitional clustering, k-means and basic variants
- some basic evaluation measures

some related textbook chapters:

- Zaki and Meira Jr. [2020], Ch. 13.1
- Tan et al. [2020], Ch. 5.1+5.2+5.5

Classification basics

- classification problem, inductive learning assumption
- hypothesis space and bias
- evaluation
- k-nearest neighbor classification

some related textbook chapters:

- Mitchell [1997], Ch. 1+2+8
- Tan et al. [2020], Ch. 3.6+6.3

Bayesian Learning

- basics of probability theory
- Bayes' rule
- probabilistic learning (Bayesian learning theory, minimum description length, Bayes optimal classification, Naïve Bayes classifier)

some related textbook chapters:

- Mitzenmacher and Upfal [2017], Ch. 1
- Mitchell [1997], Ch. 6
- Zaki and Meira Jr. [2020], Ch. 18

Learning with distributions

- Expectation, Variance, Deviations, Continuous Distributions
- Bayesian learning: EM-clustering [Dempster et al., 1977]
- non-parametric learning: density estimation, density-based clustering, hierarchical clustering
- outlier detection

some related textbook chapters:

- Mitzenmacher and Upfal [2017], Ch. 2+3+8+9
- Zaki and Meira Jr. [2020], Ch. 13+14+15
- Tan et al. [2020], Ch. 5.3+5.4+8.2.2+8.3+8.4.6-8.4.9+9

Entropy, Purity, Separation

- Entropy, randomness, information
- decision tree learning
- neural networks
- support vector machines
- regression

some related textbook chapters:

- Mitzenmacher and Upfal [2017], Ch. 10
- Mitchell [1997], Ch. 3+4
- Zaki and Meira Jr. [2020], Ch. 19+21
- Tan et al. [2020], Ch. 3.3+6.7+6.9

Ensemble Learning

- diversity and combination
- bias, variance, noise
- example methods (e.g., bagging, boosting, error-correcting output codes, random forests)

some related textbook chapters:

- Zaki and Meira Jr. [2020], Ch. 22
- Tan et al. [2020], Ch. 6.10

Resources Online resources are available for some of the recommended books:

- Tan et al. [2020] (sample chapters):
  `https://www-users.cs.umn.edu/~kumar001/dmbook/index.php`
- Zaki and Meira Jr. [2020] (complete book):
  `https://dataminingbook.info/`

References

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–31, 1977.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR*, pages 82–88, 1996.

T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

M. Mitzenmacher and E. Upfal. *Probability and Computing. Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2nd edition, 2017.

R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Montreal, QC, Canada*, pages 1–12, 1996. doi: 10.1145/233269.233311.

P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2nd edition, 2020.

M. J. Zaki and W. Meira Jr. *Data Mining and Analysis. Fundamental Concepts and Algorithms*. Cambridge University Press, 2nd edition, 2020.