DSE-3(i): Mathematical Data Science

Weeks 1 and 2: Types of Data: nominal, ordinal, interval, and ratio; Steps involved in data science case-study: question, procurement, exploration, modeling, and presentation; Structured and unstructured data: streams, frames, series, survey results, scale and source of data – fixed, variable, high velocity, exact and implied/inferred; Overview of problems with data – dirty and missing data in tabular formats – CSV, data frames in R/Pandas.
[2]: Chapter 2, Chapter 3, and relevant material for different presentation styles from Chapter 9.
[1]: Chapter 1 (up to page 28).

Weeks 3 and 4: Anomaly detection, assessing data quality, rectification and creation methods, data hygiene, meta-data for inline data-description-markups such as XML and JSON; Overview of other data-source formats – SQL, pdf, Yaml, HDF5, and Vaex.
[1]: Relevant material from Chapters 4, 5, and 6.
[1]: Chapter 1 (pages 29- 44, and 58-60).

Week 5: Model driven data in R^n, Log-likelihoods and MLE, Chebyshev, and Chernoff-Hoeffding inequalities with examples, Importance sampling.
[3]: Chapter 1 (pages 12-13), and Chapter 2 (Section 2.2, 2.3 [2.3.1 to 2.3.3], and 2.4).

Weeks 6 and 7: Norms in Vector Spaces– Euclidean, and metric choices; Types of distances: Manhattan, Hamming, Mahalanobis, Cosine and angular distances, KL divergence; Distances applied to sets– Jaccard, and edit distances; Modeling text with distances.
[3]: Chapter 3 (Section 3.3), and Chapter 4 (Sections 4.1 to 4.4).

Weeks 8 and 9: Linear Regression: Simple, multiple explanatory variables, polynomial, cross-validation, regularized, Lasso, and matching pursuit; Gradient descent.
[3]: Chapter 5, and Chapter 6 (Sections 6.1 to 6.3).

Weeks 10 and 11: Problem of dimensionality, Principal component analysis, Singular value decomposition (SVD), Best *k*-rank approximation of a matrix, Eigenvector and eigenvalues relation to SVD, Multidimensional scaling, Linear discriminant analysis.
[3]: Chapter 7 (Sections 7.1 to 7.7).

Weeks 12 and 13: Clustering: Voronoi diagrams, Delaunay triangulation, Gonzalez's algorithm for *k*-center clustering, Lloyd's algorithm for *k*-means clustering, Mixture of Gaussians, Hierarchical clustering, Density-based clustering and outliers, Mean shift clustering.
[3]: Chapter 8.

Weeks 14 and 15: Classification: Linear classifiers, Perceptron algorithm, Kernels, Support vector machines, and *k*-nearest neighbors (*k*-NN) classifiers.
[3]: Chapter 9 (Sections 9.1 to 9.5).

Essential Readings

1. Mertz, David. (2021). Cleaning Data for Effective Data Science, Packt Publishing.
2. Ozdemir, Sinan. (2016). Principles of Data Science, Packt Publishing.
3. Phillips, Jeff M. (2021). Mathematical Foundations for Data Analysis, Springer. (https://mathfordata.github.io/).