Data Science
Introduction

Stéphane Marchand-Maillet

Department of Computer Science

Master en Sciences Informatiques - Autumn semester

What is the Data Science course about?

- ⋆ Study the modelling of phenomenons into digital (numerical, quantitative) data
- ⋆ Understand the geometrical and statistical properties of this data
- ⋆ Understand the geometrical and statistical properties of the spaces this data lives into
- ⋆ Analyse the data and develop tools for this analysis
- ⋆ Understand the assumptions made in the design of these tools
- ⋆ Work out the theory (in depth)

$\Rightarrow$ involved in linear algebra, probability and statistics

Course content

Introduction

Part I: Data Analysis (SMM)

- ⋆ High-dimensional representation spaces
- ⋆ Component analysis: PCA
- ⋆ Component analysis: FCA
- ⋆ Component analysis: LDA
- ⋆ Density estimation: k-Means
- ⋆ Density estimation: Gaussian Mixture Models and the EM algorithm
- ⋆ Temporal Data Analysis: Autoregressive models
- ⋆ Temporal Data Analysis: Markov models

Part II: Information Processing (Prof. S. Voloshynovskiy)

Relationship to Machine Learning

- Data Science and Machine Learning are synonyms
- Data Science is the study of representation spaces within which Machine Learning acts
- ...

Required Background (BSc)

⚠ Mathematical formalism

Linear Algebra

- Vector space, inner product, matrix computation
- Projection, eigensystems, SVD, properties
- Optimisation, Gradient Descent, Lagragian multipliers
- Hyperplane representation, homogenisation of coordinates
- ...

Statistics and probabilities

- Random variables, expectation, variance
- Probability density function, CDF, entropy
- Joint and conditional probabilities, Bayes theorem
- ...

Notation

These notations should be consistent throughout the slides:

- ⚠ : critical, φ : redo this computation/proof
- "$:=$" definition, $[\![N]\!] = \{1, \cdots, N\}$
- x scalar, $\boldsymbol{x}$ vector, $\boldsymbol{X}$ matrix, $\mathcal{X}$ set, X random variable
- $\boldsymbol{x}(i)$ component i of vector $\boldsymbol{x}$
- $\mathbb{F}$ family, $\mathfrak{F}$ subspace
- 1_N $\mathbf{1}$ vector/matrix full of 1's, $\mathbf{Id}_N$ identity
- $\boldsymbol{x}^\top$ $\boldsymbol{X}^\top$ transpose (dual), $\boldsymbol{A}^+$ Moore-Penrose inverse
- $\mathbb{P}(X = \boldsymbol{x})$ probability, $\mathbb{E}X$ expectation, $\mathrm{Var}(X)$ variance
- $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ inner product
- $\|\cdot\|$ norm (default $= L_2$), $d(\cdot, \cdot)$ distance function (default=Euclidean)

References I

[1] Sheldon Axler. *Linear Algebra Done Right*. Springer, 2015.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. (available online).

[3] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020. (available online).

[4] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2 edition, 2001.

[5] Jonathan S. Golan. *The Linear Algebra a Beginning Graduate Student Ought to Know*. Springer, 2004.

[6] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press, 2003. (available online).

[7] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass., 2013.

[8] Gilbert Strang. *Linear Algebra and Its Applications*. Brooks/Cole, 4th edition edition, 2005.

[9] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.