Foundations of Data Science

Data-driven computation is now a ubiquitous feature of variety of programs designed for estimation, modeling and inference. In all cases, the starting point is an assumption that there is now so much data about a system being studied, that it is possible to answer practically any kind of question about the system by modelling, estimation and inference (thus giving a modern twist to Liebniz' exhortation "Let us calculate!"). Two developments have been driving this pervasive growth of data. First, it has become increasingly easier to generate data in an automated manner, due to algorithmic and technological advances. Secondly, from 1990, mass storage (disk) capacity has been roughly doubling every year, keeping up with the demands of machine-generated data. Combined with significant increases in computational power, there are now a range of new techniques for the computational modelling of data using a combination of techniques from algebra, geometry, applied mathematics and computation. This course will cover the basics of this form of computation by introducing some of the basic mathematical tools and computational techniques that are widely used in data-driven modelling of systems.

Lecturer(s): Ashwin Srinivasan and Sujith Thomas

Pre-requisites

- None

Post Condition (on student capability after successfully completing the course):

- At the end of the course, the student will be able to: understand some basic techniques from linear algebra; optimisation, probability and statistics that form the foundations of current-day data-science

Brief Description

Module	Lecture	Topics Covered
Introduction	1	Data and Science
Algebra	2	Vector spaces and Linear independence
	3	Subspaces, Column space, Null space
	4	Basis and dimension of vector spaces
	5	Solving Ax = b when solution exists
	6	Four fundamental subspaces of A
	7	Rank of A and existence of inverse of A
	8	Linear transformations
	9	Orthogonal subspaces
	10	Eigenvalues and Eigenvectors
	11	Diagonalization of a matrix
	12	Singular value decomposition (SVD)
	13	Applications of SVD
Geometry	14	Curse of high-dimensional data
	15	Geometry of high-dimensional data
	16	Distances in high-dimensional space
	17	Principal Components Analysis (PCA)
	18	Application to Dimensionality Reduction
Calculus	19	Introduction to Optimisation
	20	Constrained Convex Optimisation
	21	Numerical Optimisation
	22	Numerical Optimisation
	23	Application to Data Fitting

Module	Lecture	Topics Covered
Probability and Statistics	24	Data Reduction
	25	Probability Distributions 1
	26	Probability Distributions 2
	27	Probability Bounds
	28	Distance between Distributions
	29	Generative and Discriminative models
	30	Estimation, Bias and Variance
	31	Risks and Decisions
	32	Model Selection
	33	Simple Hypothesis Testing
	34	Application to Data Visualisation
	35	Application to Experimentation
Conclusion	36	Correlation and Causation
	37	Concluding Remarks

Evaluation

Assessment will consist of the following: quizzes (30%), mid-semester exam (30%), final examination (40%).

Texts/Other Resources

- **A.S.C. Ehrenberg**, A Primer in Data Reduction, Wiley, 1982
  - Still one of the best introductory statistics textbooks around.
- **J.K. Kruschke**, *Doing Bayesian Data Analysis*, Academic Press, 2010.
  - A good text introducing Bayesian data analysis
- **G. Strang,** *Introduction to Linear Algebra. 5th Edition*. Wellesley-Cambridge Press, 2016.
  - The unmatched book for linear algebra
- **David Rosenberg,** Extreme Abridgement of Boyd and Vandenberghe's Convex Optimization
- **Sheldon M Ross,** Introduction to Probability and Statistics for Engineers and Scientists. *5th Edition*. Elsevier Academic Press, 2014.
  - A good introductory textbook on probability