DS 294 Jan. 3:0

Data Analysis and Visualization

Instructor

Anirban Chakraborty
Email: anirban@iisc.ac.in

Teaching Assistant

Email:

Department: Department of Computational and Data Sciences (CDS)

Course Time: Mon., Wed., Fri., 3:00 PM - 4:00 PM

Lecture venue: CDS 202

Detailed Course Page:

Announcements

The first meeting of the 2018 session would be held at 3:00 PM on January 8, 2018 (Monday) in CDS room 202.

Brief description of the course

DAV (DS 294) is designed as an introductory first course (core) in the general areas of data analytics, machine learning and visualization tools and techniques for better understanding, analyzing and presenting data and drawing inferences based on it. The course is targeted towards the first year M.Tech./Ph.D. students with zero to beginner's knowledge in the aforementioned areas. The course will start by providing a brief review of probability, statistics and random processes and introduce basics of estimation theory. Various machine learning techniques would subsequently be introduced to the students including basic theory and practical implementation details to solve real world problems. It this course, students will be taught how to mine and tell visual stories from data. It would cover fundamentals of visual presentation, open source software tools for the same and when to use a specific tool for visualization. Visualization toolkit for 3D computer graphics and visualization for deep learning would conclude the course.

Prerequisites

- Basic knowledge in linear algebra, probability and statistics, calculus, algorithms and data structures are preferred.

- Consent of the instructor.

Syllabus

-Importance of analytics and visualization in the era of data abundance.

-Review of probability, statistics and random processes.

- Brief introduction to estimation theory.

- Introduction to machine learning, supervised and unsupervised learning, gradient descent, overfitting, regularization etc.

- Clustering techniques: K-means, Gaussian mixture models and expectation-maximization, agglomerative clustering, evaluation of clustering - Rand index, mutual information based scores, Fowlkes-Mallows index etc.

- Regression: Linear models, ordinary least squares, ridge regression, LASSO, Gaussian Processes regression.

- Supervised classification methods: K-nearest neighbor, naive Bayes, logistic regression, decision tree, support vector machine.

- Sparse coding and dictionary learning, orthogonal matching pursuit.

- Introduction to artificial neural networks (ANNs), deep NNs, convolutional neural network (CNN), and other recent topics.

- Data visualization: Basic principles, categorical and continuous variables.

- Exploratory graphical analysis.

- Creating static graphs, animated visualizations - loops, GIFs and Videos.

- Data visualization in Python and R, examples from Bokeh, Altair, ggPlot, ggplot2, gganimate, ImageMagick etc.

- Introduction to Visualization Toolkit (VTK) for 3D computer graphics, image processing and visualization.

- Visualization for deep learning.

Course outcomes

At the end of the course, the students should be able to parse a real-world data analysis problem into one or more computational components learned in this course, apply suitable machine learning and/or visualization techniques and analyze the results obtained to enable optimal decision making. This would also act as a first course in data science which would provide necessary pre-requisites and knowledge to explore more specialized and involved topics in machine learning, analytics, statistics etc.

Grading policy

Home work: 40% (about 5-6 in the semester)

Project: 20% (to be submitted and evaluated at the end of the course)

Mid Term Exam: 20%

End Term Exam: 20%

Assignments

About 5-6 take home assignments would be provided during the course.

Resources

There is no prescribed course textbook. Slides and materials from course lectures would be provided. The references and recommended readings besides the contemporary literature are -

1. Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie, The Elements of Statistical Learning, Springer, 2001.
2. Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
3. David G. Stork, Peter E. Hart, and Richard O. Duda, Pattern Classification (2nd edition), Wiley, 2000.
4. Edward Tufte, The Visual Display of Quantitative Information (2nd edition), Graphics Press, 2001.
5. Colin Ware, Information Visualization: Perception for Design (2nd edition), Morgan Kaufmann, 2004.
6. Alberto Cairo, The Functional Art: An Introduction to Information Graphics and Visualization, New Riders, Pearson Education, 2013.
7. Nathan Yau, Data Points: Visualization That Means Something, Wiley, 2013.
8. Charles D. Hansen and Chris R. Johnson, Visualization Handbook, Academic Press, 2004.
9. Will Schroeder, Ken Martin, and Bill Lorensen, The Visualization Toolkit: An Object-Oriented Approach to 3D Graphics, Kitware Inc. Publishers, 2004.