Course Introduction and Administrivia

- Imdadullah Khan

Course Goals

Presented with data, the student should be able to

- Describe the behavior of data
- Select the appropriate statistical tool and algorithm to solve a real world problem based on the data
- Appreciate the strengths and weaknesses of different solutions
- Understand and convey the result generated by the algorithm, as well as the assumptions and limitations of the methods

Textbook

No textbook will be strictly followed, but for some topics we will use material from the following books

- **Foundations of Data Science**
  A. Blum, J. Hopcroft, R. Kannan
- **Mining Massive Datasets**
  J. Leskovec, A. Rajaraman, J. Ullman
- **Data Mining: Concepts and Techniques**
  J. Han, M. Kamber, J. Pei
- **Data Streams: Algorithms and Applications**
  S. Muthukrishnan
- Other notes, slides or chapters

Topics

- Introduction to Big Data and Applications
- Descriptive and Exploratory Data Analysis - Data Visualization
- Preprocessing & Data transformation
- Proximity Measures (Distance and Similarity)
- Common Analytics Tasks and Techniques
  - Classification & Regression
  - Cluster Analysis
  - Recommendation System
  - Text Analytics and Information Retrieval
- High Dimensional Geometry and Curse of Dimensionality
- Locality Sensitive Hashing
- Dimensionality Reduction
- Streaming and Sampling
- Link Analysis & Web Search
- Spectral Clustering and Social Network Analysis

Topics

Introduction to Big Data Analytics

- Applications
- Aspects of Big
- Sources & Types
- Data analytics process

Topics

Getting to know your data

- Statistical Data Description
- Exploratory Data Analysis
- Graphical Data Description
- EDA on Text Data

Information Visualization

- Motivation and value of visualization
- Principle of Visualization
- Visual Encoding
- Visual Perception

Topics

Data Preprocessing & Data transformation

- Data Cleaning
- Data Integration
- Data Reduction
- Standardization & Normalization
- Data Transformation

Topics

Proximity Measures (Distance and Similarity)

- Need for measures
- Similarity measures for Vector Data
- Non-Vector Data (Sets and Bags of words)
- TF-IDF

Topics

Common Analytics Tasks, Techniques and Evaluations

- Classification
  - Naive Bayes Classifier
  - Nearest Neighbor Classifier
  - Decision Tree
- Regression
  - Linear Regression
  - Logistic Regression
- Clustering
  - Point Assignment Clustering - K-Means and K-Medoids
  - Hierarchical Clustering - Agglomerative and Partition based clustering
- Recommendation Systems
  - Content Based Recommender
  - Collaborative Filter based recommender
- Text Analytics and Information Retrieval
  - Vector Space Model
  - Sentiment Analysis and other NLP Tasks, Ranked Retrieval

Topics

High Dimensional Geometry and Curse of Dimensionality

- Computational Complexity
- Data Sparsity
- Combinatorial Complexity of nearest neighbors search
- Diminishing volume of n-ball
- Distance Concentration
- Angle Concentration
- Generating random angles

Topics

Locality Sensitive Hashing

- LSH for Dimensionality Reduction
- The S-Curve and theory of LSH
- LSH for Hamming Distance
- LSH for Cosine Distance
- LSH for Jaccard distance (min-wise hashing)
- LSH for Euclidean distance (Random Projection)

Topics

Dimensionality Reduction

- Random Projection, the Johnson-Lindenstrauss Lemma
- Linear Algebra Review
- Singular Value Decomposition
- Principal Component Analysis

Topics

Recommendation Systems

- Content Based Filtering
- Collaborative Filtering
- Latent Factor Analysis (UV-decomposition)

Topics

Streaming and Sampling

- Models of data streams
- Frequency Estimation
- The Count-min Sketch
- The Count Sketch
- AMS Sketch & dimensionality reduction
- Lower bound
- Sampling, Weighted Sampling, Dynamic weights
- Stream Sampling

Topics

Link analysis & Web Search

- Basics Information Retrieval
- Web Search
- Pagerank algorithm
- Pagerank algorithm: Algebraic Formulation
- Pagerank algorithm: Markov Chain Formulation
- Pagerank algorithm: Matrix Formulation
- Spam Farm and TrustRank
- Topic Sensitive Pagerank
- The HITS algorithm: Hubs and Authorities view
- The HITS algorithm: Matrix formulation

Topics

Spectral Clustering and Spectral Graph Theory

- Proximity Graphs
- Graph Laplacian
- Spectral Partitioning of Graphs

Topics

Social Network Analysis

- Communities and Direct Community Detection
- Important Players and Centrality Measures
- Overlapping Communities The Affiliation-Graph Model

Grading (Tentative)

- **Quizzes, Attendance & Class Participation** 15%
  - 6-10 10 minutes (online) quizzes
- **Homework Assignments and Labs** 25%
  - Data Analysis Assignments, datasets and tasks will be assigned
  - May conduct them jointly as Labs some could be (data) assignment
- **Research Project** 60%
  - Review of recent research papers (according to the provided template)
  - Every student will review some assigned papers, to write reports and present in class
  - You will choose 1/2 papers from a list (in search of project proposal)
  - All reports must be typed in LaTeX(.tex & .pdf to be submitted)
  - Project done in groups and will have **separately graded phases**
  - Initial Proposal, Final Proposal, Data Report, Literature Review, Intermediate Report, Final Report, Presentation
  - Templates for each deliverable will be provided and will be graded
  - All this grading will be done with a viva

No Makeup for anything.

Other Information

- Course Website: LMS tab
- important announcements
- homework assignments
- readings and reading assignments
- templates
- Calendar
- Zoom links
- check regularly!

Course Staff

- Instructor: Imdadullah Khan
  - Office Hours:
  - Email: imdad.khan@lums.edu.pk
  - Grading of each instrument will be done during meetings
  - Will conduct a thorough discussion (almost a viva)
  - Necessary feedback will be provided
  - Project related meetings will also be done during office hours
  - You can meet us any other time by appointment

All this and other contact and office hours information along with a project calendar will be posted on LMS as an announcement