Second Semester Area B (data science)

CSE 382M, Foundational Techniques of Machine Learning and Data Sciences. Spring 2021.

References:

Foundations of Data Science, by Blum, Hopcroft, and Kannan.

Non-convex optimization for Machine Learning, by Jain and Kar.

Mathematics of Data Science, by Bandeira, Singer, Strohmer (in progress).

Prerequesites: Undergraduate course in probability and (proof-based) linear algebra. Basic computer programming knowledge (Matlab, Python, or C).

Tentative list of topics:

1. Introduction to Data Science. Historical perspective. Applications.
2. High-dimensional Probability: Multivariate Gaussians, CLT, concentration of measure
3. Linear algebra: SVD, condition number, overdetermined/underdetermined systems.
4. Regression/Interpolation: Generalization/overfitting. Regularization. Cross-validation.
5. Optimization: Gradient Descent, convexity, smoothness.
6. Stochastic Optimization: Stochastic Gradient descent. Importance sampling. Condition number. Implicit regularization.
7. Classification: Support Vector Machines (SVM), Kernel SVM Classification
8. Principal Component Analysis (PCA): Best rank-k approximation, Power method.
9. Random Projections: JL lemma, approximate K-nearest neighbor
10. Matrix sampling and sketching.
11. Compressed Sensing: sparsity, L1 regularization.
12. Learning models from data: Convex vs non convex optimization, alternating minimization / expectation maximization.
13. Unsupervised Clustering: k-means. Lloyd's method. Spectral clustering

14. Statistical estimation: Bayes Rule , Maximum Likelihood, Maximum a Posteriori (MAP) Probability
15. Graph analysis: Random walks. Markov chains. Community detection.
16. Control theory: Hidden Markov models, Kalman Filter/ Linear quadratic estimation
17. Neural networks: Resnet for supervised learning, Generative Adversarial Network for unsupervised learning. Implicit regularization. Overparameterization.