Data Science - Syllabus

Statistics

- Understand the fundamentals of statistics
- Learn how to work with different types of data
- How to plot different types of data
- Calculate the measures of central tendency, asymmetry, and variability
- Calculate correlation and covariance
- Distinguish and work with different types of distributions
- Estimate confidence intervals
- Perform hypothesis testing
- Make data driven decisions
- Understand the mechanics of regression analysis
- Carry out regression analysis
- Use and understand dummy variables
- Understand the concepts needed for data science even with Python and R!

Python & R Language

- Introduction
- Basic constructs
- Various Libraries in Python and R Language
- OOPs in Python
- NumPy for mathematical computing
- SciPy for scientific computing
- Data manipulation
- Data visualization with Matplotlib
- Machine Learning using Python
- Supervised learning
- Unsupervised Learning
- Python integration with Spark

Machine Learning

1. **Introduction**
   Definition of learning systems. Goals and applications of machine learning. Aspects of developing a learning system: training data, concept representation, function approximation.
2. **Regression**
   Studying various regression techniques used for classification and Prediction in Supervised Learning.
3. **Decision Tree Learning**
   Representing concepts as decision trees. Recursive induction of decision trees. Picking the best splitting attribute: entropy and information gain. Searching for simple trees and computational complexity. Overfitting, noisy data, and pruning.
4. **Ensemble Learning**
   (read this paper) Using committees of multiple hypotheses. Bagging, boosting, and DECORATE. Active learning with ensembles.
5. **Experimental Evaluation of Learning Algorithms**
   Measuring the accuracy of learned hypotheses. Comparing learning algorithms: cross-validation, learning curves, and statistical hypothesis testing.
6. **Artificial Neural Networks**
   Neurons and biological motivation. Linear threshold units. Perceptrons: representational limitation and gradient descent training. Multilayer networks and backpropagation. Hidden layers and constructing intermediate, distributed representations. Overfitting, learning network structure, recurrent networks.
7. **Support Vector Machines**
   Maximum margin linear separators. Quadractic programming solution to finding maximum margin separators. Kernels for learning non-linear functions.
8. **Bayesian Learning**
   Probability theory and Bayes rule. Naive Bayes learning algorithm. Parameter smoothing. Generative vs. discriminative training. Logisitic regression. Bayes nets and Markov nets for representing dependencies.
9. **Instance-Based Learning**
   Constructing explicit generalizations versus comparing to past specific examples. k-Nearest-neighbor algorithm. Case-based learning.
10. **Text Classification**
    Bag of words representation. Vector space model and cosine similarity. Relevance feedback and Rocchio algorithm. Versions of nearest neighbor and Naive Bayes for text.
11. **Clustering and Unsupervised Learning**
    Learning from unclassified data. Clustering. Hierarchical Aglomerative Clustering. k-means partitional clustering. Expectation maximization (EM) for soft clustering. Semi-supervised learning with EM using labeled and unlabled data.