Machine Learning with Spark

Course Outline:

Section 1: Introductions and overviews

- Machine learning: goals, results, supervised/unsupervised
- Spark as a tool for Big Data
- Scala as the language of Spark (together with Python, Java and R)

  If the students do not have the Spark/Scala prerequisites, a thorough introduction of these is taught in the section

Section 2: SVM (Supervised Vector Machines)

- Theory
- Lab
- Use case: anomaly detection

Section 3: Logistic Regression

- Theory
- Lab
- Use case: healthcare prediction

Section 4: Linear regression

- Theory
- Lab

- Use case: financial modelling

Section 5: Naive Bayes

- Theory
- Lab
- Use case: spam filtering

Section 6: Decision Trees

- Theory
- Lab
- Use case: vessel shipment planning

Section 7: Clustering (K-Means)

- Theory
- Lab
- Use case: topic grouping

Section 8: LDA (Latent Dirichlet Allocation)

- Theory
- Lab
- Use case: unsupervised topic discovery

Section 9: Principal Component Analysis (PCA)

- Theory
- Lab

- Use case: stock analysis

Section 10: Recommendation (Collaborative filtering)

- Theory
- Lab
- Use case: dating

Section 11: Graphs – graph operations

- Theory
- Lab
- Use case: finding followers

Section 12: Graphs – optimizations with Pregel

- Theory
- Lab
- Use case: shortest routes, PageRank