# Applied Machine Learning: syllabus

Dmitry Adamskiy David Barber Silvia Chiappa

March 26, 2019

Here is the list of topics covered in the Applied Machine Learning course which you might find helpful while preparing for the exam.

1. Competitive Data Science aspects
   - Preprocessing issues
   - Competition Metrics. Various metrics for classification and regression and their properties.
   - Train/test validation split. Random, time-based, id-based splits.
   - Data leaks. The concept of a data leak, examples.
2. Optimisation
   - Gradient descent. Toy example: GD for quadratic form. Learning rates and convergence, optimal learning rate and eigenvalues.
   - Line search. The idea, updates for the quadratic form.
   - Conjugate gradients. What are conjugate directions, intuition behind the method. Why CJ is an efficient method for sparse linear systems.
3. Tree Ensembles
   - Bagging and Boosting
   - Random forests as bagging of decision trees. Out of Bag estimate. Feature importance.
   - Gradient Boosting.
4. Matrix Factorisation
   - Low-rank matrix factorisation: problem setting and motivation
   - SVD as a solution.
   - Connection between PCA and SVD.
   - Non-negative Matrix Factorisation. Why SVD is not enough for topic modeling.
   - Standard algorithms for NNMF: Alternating non-negative least squares.
   - Separable case. Anchor words assumption and the efficient algorithm based on it.

5. Clustering
    - Clustering problem.
    - K-means. Algorithm description, issues with K-means.
    - Spectral clustering: algorithm, the intuition behind it, similarity graphs, connection between graph Laplacian eigenvalues and connected components. RatioCut and Random walk viewpoints (intuitions).
6. Visualisation
    - SNE. The goal of SNE, the resulting objective function.
    - t-SNE. The idea behind using heavy-tail distribution.
    - t-SNE gradient and its interpretation in terms of attractive and repulsive forces.
    - Landmark datapoints and random walk trick for dealing with large datasets.
    - Speeding up: approximating input similarities with vantage-point trees. Barnes-Hut approximation.
7. Fast-NN
    - Using triangle inequality to speed-up NN queries.
    - Orchard's algorithm and AESA algorithm.
    - KD-trees and Vantage point trees.
8. Fairness
    - General knowledge of Bayesian networks .
    - How to assess independence between random variables in Bayesian networks.
    - Difference in the basic fairness definitions demographic parity, equal false positive/false negative rates, calibration, and fairness through unawareness.
    - Discussed example of least-squares predictor using and not using the sensitive attribute.