Course Title (in English)	Introduction to Data Science
Course Title (in Russian)	Введение в анализ данных
Lead Instructor(s)	Belyayev, Mikhail Panov, Maxim

Is this syllabus complete, or do you plan to edit it again before sending it to the Education Office?	The syllabus is a work in progress (draft)
Contact Person	Maxim Panov
Contact Person's E-mail	m.panov@skoltech.ru

1. Annotation

Course Description

The course gives an introduction to the main topics of modern data analysis such as classification, regression, clustering, dimensionality reduction, reinforcement and sequence learning, scalable algorithms. Each topic is accompanied by a survey of key machine learning algorithms solving the problem and is illustrated with a set of real-world examples. The primary objective of the course is giving a broad overview of major machine learning techniques. Particular attention is paid to the modern data analysis libraries which allow solving efficiently the problems mentioned above.

Course Prerequisites / Recommendations	Linear algebra, mathematical analysis, algorithms. At least intermediate programming skills are necessary! During the course you'll write simple Python programs like this http://scikit-learn.org/stable/auto_examples/plot_cv_predict.html Second year Data Science track students aren't eligible for credits, but are allowed to attend the course.

2. Structure and Content

Course Academic Level	Master-level course suitable for PhD students
Number of ECTS credits	3

Topic	Summary of Topic	Lectures (# of hours)	Seminars (# of hours)	Labs (# of hours)
General introduction	A definition of data science, real-world examples of data science applications, an overview of main topics in machine learning	4		
Solving machine learning problems in Python	Why Python, overview of Python libraries: scikit-learn, pandas, seaborn, visual exploration. Practical example: exploring the Titanic dataset	1	1	1
Elements of Multivariate Statistics	Multivariate Normal, Conditional Normal, Wishart Distributions; Hotellings T2 test; Analysis of Variance; Multivariate Analysis of Variance; Multiple testing correction; Histograms; Kernel Density Estimation. Practical Example: the dead salmon study	1	1	1
Regression, cross-validation	Supervised learning, k nearest neighbours, linear regression, L1&L2 regularization, overfitting & underfitting concepts (the Bias-Variance Tradeoff). Practical example: the bike sharing demand dataset	1	1	1
Classification, quality metrics	Classification problems, logistic regression, SVM, loss functions, precision & recall, ROC curve. Practical example: the Titanic dataset (continued)	1	1	1
Decision trees	Overview, handling missing values, calculating features importance, algorithms complexity, visualisation. Practical example: the Iris dataset	1	1	1
Ensembling	Bagging, Boosting, Random Forest, Gradient Boosting, XGboost library. Practical example: Forest Cover Type Prediction	1	1	1
Features engineering & selection	Feature selection approaches: wrappers, filters, embedded methods; categorical features, text features, time-series features. Practical example: Amazon Employee Access	1	1	1
Dimensionality Reduction	Principal Component Analysis, overview of nonlinear methods (Isomap, LTSA, tSNE). Practical examples: DR for airfoils & generation of new airfoils, genetic signature of Jewish ancestry	1	1	1
Clustering	K-means, Gaussian Mixture Model, Hierarchical clustering Spectral clustering. Practical example: text documents clusterization	1	1	1
Basics of Neural Networks	Stochastic Gradient Descend, Multilayer perceptron, activation functions (ReLu, tanh), Dropout, training and validation. Early Stopping, Convolutional networks; Keras library. Practical example: toy problems, Facial keypoints recognition	1	1	1
Scalable algorithms	Overview, MapReduce paradigm; collaborative filtering. Practical example: Netflix	1	1	1

3. Assignments

Assignment Type	Assignment Summary
Project	Solve a real-life data science problem. An example: build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year

4. Grading

Type of Assessment	Graded

Grade Structure

Activity Type	Activity weight, %
Homework Assignments	50
Projects	50

Grading Scale

A:	85
B:	70
C:	55
D:	40
E:	25
F:	0
Attendance Requirements	Mandatory

5. Basic Information

Maximum Number of Students

	Maximum Number of Students
Overall:	80
Per Group (for seminars and labs):	

Course Stream	Science, Technology and Engineering (STE)
Course Term (in context of Academic Year)	Term 1B (last four weeks)
Course Delivery Frequency	Every year

Students of Which Programs do You Recommend to Consider this Course as an Elective?

Masters Programs	PhD Programs
Advanced Manufacturing and Materials Computational Science and Engeneering Data Science Petroleum Engineering	Computational and Data Science and Engineering Engineering Systems Petroleum Engineering

Course Tags	Programming Engineering

6. Textbooks and Internet Resources

Required Textbooks	ISBN-13 (or ISBN-10)
The Elements of Statistical Learning, 2nd edition by Hastie, Tibshirani and Friedman, Springer-Verlag, 2008	9780387848570
Pattern Recognition and Machine Learning by Bishop, Springer, 2006	9780387310732

Recommended Textbooks	ISBN-13 (or ISBN-10)
Machine Learning: A Probabilistic Perspective by Kevin P. Murphy, MIT Press, 2012.	9780262018029
Bayesian Reasoning and Machine Learning by David Barber, Cambridge University Press, 2012.	9780521518147
Deep Learning by Yoshua Bengio, Ian Goodfellow, and Aaron Courville.	9780262035613

Web-resources (links)	Description
http://scipy-lectures.org	Tutorials on the scientific Python ecosystem.

7. Facilities

Equipment
Laptop with pre-installed python

Software
Python 3.4+

8. Learning Outcomes

Knowledge
Statements of all major machine learning problems.
Mathematical details of the most important data analysis methods and algorithms.

Skill
Select an appropriate method for solving particular data analysis problems.
Perform basic data processing and visual analysis, generate features for subsequent machine learning.
Apply machine learning libraries, select algorithm's hyperparameters.
Critically evaluate the obtained results and redesign data-processing pipelines.
Solve real-world data science problems using modern machine learning techniques.

9. Assessment Criteria

Input or Upload Example(s) of Assigment 1:

Select Assignment 1 Type	Project
Input Example(s) of Assignment 1 (preferable)	Students should perform full analysis of the chosen real-life data science problem and prepare a Jupyter notebook as a report.
Assessment Criteria for Assignment 1	1) The general literacy and style of the report — 10%; 2) Data science methods and approaches — 20%; 3) Depth of the subject understanding— 45%; 4) The presentation and answers to questions — 25%.

Input or Upload Example(s) of Assigment 2:

Select Assignment 2 Type	Problem Set

Input Example(s) of Assignment 2 (preferable)	Homework 4: 1. Implement k-nearest neighbors method in Python. 2. Estimate bias and variance as a function of neighborhood size. 3. Estimate quality of kNN prediction in two scenarios: a) the data is used as is, b) the data is normalized in advance.
Assessment Criteria for Assignment 2	1) The general literacy and style of the report — 10%; 2) Data science methods and approaches — 20%; 3) Depth of the subject understanding— 45%; 4) The presentation and answers to questions — 25%.

Input or Upload Example(s) of Assigment 3:

Select Assignment 3 Type	Problem Set
Input Example(s) of Assignment 3 (preferable)	Homework 5: Deep analysis of a real-life data science problem: Classification of shopping trips based on market basket analysis. Perform the following analysis: 1. Parse data from file. 2. Perform visual analysis of the data. 3. Build cross validation procedure. 4. Propose and evaluate several feature generation methods based on special characteristics of the dataset. 5. Compare classification algorithms (including different sets of hyperparameters). 6. Evaluate performance of the best model from the business point of view.
Assessment Criteria for Assignment 3	1) The general literacy and style of the report — 10%; 2) Data science methods and approaches — 20%; 3) Depth of the subject understanding— 45%; 4) The presentation and answers to questions — 25%.

Input or Upload Example(s) of Assigment 4:

Input or Upload Example(s) of Assigment 5:

10. Additional Notes