

MAKE THE SWITCH TO DATA SCIENCE

Get ready to enter Data Science using your prior experience and skills, within 14 weeks



Pre course: OOP in Python + Computing fundamentals	Classes and Objects	id() Function, using __init__, encapsulation, how are Objects and Classes stored in memory
	Attributes and Methods	Class Attributes and Instance Attribute, 'self' parameter, Static Methods and Instance Methods
	Inheritance	Basic Inheritance, Method Overriding, Constructor Inheritance
	Decorators	Decorators overview, @property decorator
	"Magic Methods"	Comparison, equality and inequality Magic Methods (__add__, __sub__, etc)
	Testing	Using the pytest module to write test for your code
	Git	Git structure, operations, use-cases, shell commands, interface with PyCharm
	Linux	Basic structure, important directories, permissions, bash, using the shell
Pre course: Statistics Fundamentals	Probability basic concepts	Expectation, variance, covariance (univariate and multivariate)
	Important distributions	Bernoulli, Poisson, Binomial, Uniform, Exponential, Normal
	LLN and CLT	LLN and CLT
	Random variables	Random variables- definition, pmfs and pdfs, operations
	Types of convergence	In distribution, in probability, almost surely
	Bayesian statistics	Bayes formula, prior and posterior distributions
	Statistics Basic concepts	Population, sampling, inference, estimators, statistics
Advanced Data science tools in Python	Jupyter	Jupyter notebooks: general operations, tips and tricks.
	Scientific Python	Pandas - dataframes, series, timeseries, merges, groupby Numpy - arrays and matrices, broadcasting, linalg tools
	Matplotlib & Seaborn	Matplotlib & Seaborn packages, various plots (line, scatter, histograms, distplots.) and their parameters
	Data analysis pipeline and processes	Exploratory data analysis. Structuring ML and data science projects. DVC. Presenting results. Best practices. ML and DS landscape and tools.
Statistical modeling	Probability & decision making	Probability in data science
		Decision making with probabilistic models
		Descriptive statistics and visualization
		Evaluating uncertainty of sample estimates
	Distribution and parameters estimation	Important distributions and their characteristics
		Methods of parameter estimation (Maximum Likelihood Estimation)
		Unbiased estimators
		Bayesian inference
	Hypotheses testing	Single parameter methods, conjugate distributions
		Tests for comparing means and goodness of fit
		Sequential tests
		Multiple hypothesis (Bonferroni, FDR)
Predictive Statistics	T-test and T-distribution, Chi squared distribution	
	Confidence intervals, Type 1 and type 2 errors	
	Kolmogorov-Smirnov	
	Basic non-parametric tests (sign, Mann-Whitney, Wilcoxon-rank)	
	Linear regression and inference from a regression model	
	Multinomial regression	
	Logistic regression, Poisson regression, GLMs	
	Mathematics of joint distributions	
Definition of latent variables		
Machine learning	ML Intro	Introduction, definition of ML, types of ML
	Regression	Definition, Linear regression: closed form solution. GLM(Poisson, Exponential etc.)
	Classifiers	Linear separators, SVM, Logistic regression
		Naive Bayes
		Classification metrics
		KNN, Bayesian classifiers, GLM
	Model Selection	Bias-variance tradeoff, validation set, cross-validation
		Overfitting and underfitting
		Regularization and hyperparameter tuning
Features selection		
Decision Trees	Explainability	
	The decision tree algorithm, usage in regression and classification, random forests, examples of non-linear classification Ensemble methods, XGboost, regularization and hp-tuning	
Unsupervised learning	Intro to unsupervised learning (k-means, clustering, EM etc)	
	K-Means algorithm	
	Dimensionality reduction (PCA)	
SQL Basics	Filtering, Sorting, and Calculating Data with SQL	WHERE, BETWEEN, IN, OR, NOT, LIKE, ORDER BY, and GROUP BY, wildcard function
	Aggregate functions	AVERAGE, COUNT, MAX, MIN
	Subqueries and Joins	key field, cartesian join, inner join, left and right joins, full outer joins and a self join, aliases and pre-qualifiers
	Modifying and Analyzing Data with SQL	modify strings by concatenating, trimming, changing the case, using the substring function, case statements, views
Big data handling	Parallel computing	Parallel computing
	PySpark on python	PySpark on python
	Parallelising and Concurrency in Python	Multiprocessing and Threading
	Handling Big Data	Working with partitioned dataframes, parallelizing, finding outliers
	Bias and Drop Analysis	Analysing large dataframes, dealing with outliers, the cost of dropping nans
Neural networks	Back propagation	Back propagation
	CNNs and RNNs	CNNs and RNNs
	Well-known architectures	Well-known architectures
	Training an NN model	Hyperparameters, exploding/vanishing gradients, Batch normalization, Overparameterization
	Use cases	Use cases of Neural Networks
Coding in a research community	Best practices in big data	Best practices in big data
	Debugging	Debugging in big data
	PRing and CRing	How to review code and how to write effective pool requests on git
	Time management in data work	Estimating running time of procedures
	AWS	Knowledge of the AWS cloud infrastructure (S3, VMs, containers)