Data Scientist Training Syllabus (May 2017)

Week	Content	Date
Week One	**Python data analytics eco-system** 1. What is data scientist 2. Key data structures in Python & Numpy 3. Pandas for data analytics 4. Fast data visualization in Pandas	Saturday (5-7PM)
	Introduction to Data Application 1. Data science project lifecycle 2. Cluster and distributed computing 3. Hadoop eco-system 4. HDFS	Sunday (5-7PM)
	Monday TA Office Hour	Monday (6-7:30PM)
	Advanced Python - 1 Basic CS Algorithm -1	Tuesday (6-8PM)
	Wednesday TA Office Hour	Wednesday (6-7:30PM)
	Thursday TA Office Hour	Thursday (6-7 PM)
Week Two	**大作业讲解** function syntax basic data structure pandas read files basic data exploration & data slicing	Saturday (3:30-4:30PM)
	Python machine learning eco-system 1. Machine learning introduction 2. Full machine learning flow in Python 3. Scikit-learn package 4. Basic regression model	Saturday (5-7PM)
	Probability and distribution with R	Sunday (5-7PM)
	Monday TA Office Hour	Monday (6-7:30PM)
	Advanced Python - 2 Basic CS Algorithm -2	Tuesday (6-8PM)
	Wednesday TA Office Hour	Wednesday (6-7:30PM)
	Thursday TA Office Hour	Thursday

		(6-7 PM)
Week Three	大作业讲解 matplotlib--histogram, scatter plot simple linear regression and cross validation	Saturday (3:30-4:30PM)
	Supervised learning: classification 1. Classification measurement 2. Basic classification models (Logistic, Tree, SVM) 3. Ensemble models	Saturday (5-7PM)
	Data Analysis using Hadoop Hive 1	Sunday (5-7PM)
	Monday TA Office Hour	Monday (6-7:30PM)
	Machine Learning Algorithm -1 Brief Introduction to Machine Learning Algorithm	Tuesday (6-7PM)
	Wednesday TA Office Hour	Wednesday (6-7:30PM)
	Thursday TA Office Hour	Thursday (6-7 PM)
Week Four	大作业讲解 data exploration methods comparison on real data data slicing on textual data simple data insights find out unusual data draw bar chart with multiple components	Saturday (3:30-4:30PM)
	Supervised learning: regression 1. bias variance trade off 2. underfitting vs. overfitting 3. regularization (Lasso, Ridge, Elastic-Net) 4. Advanced techniques in regression	Saturday (5-7PM)
	Data Analysis using Hadoop Hive 2	Sunday (5-7PM)
	Monday TA Office Hour	Monday (6-7:30PM)
	Machine Learning Algorithm -2 SVM Classifiers	Tuesday (6-7PM)
	Wednesday TA Office Hour	Wednesday (6-7:30PM)
	Thursday TA Office Hour	Thursday (6-7 PM)

Week Five	**大作业讲解** draw bar chart with multiple components draw correlation heatmap & matrix sklearn package intro RandomForestClassifier intro GridSearchCV intro model ensemble intro	Saturday (3:30-4:30PM)
	Advanced visualization & A/B Testing 1. Basic & interactive visualization in Python 2. Exploratory analysis 3. A/B test and experimentation	Saturday (5-7PM)
	Data Visualization with Tableau	Sunday (5-7PM)
	Monday TA Office Hour	Monday (6-7:30PM)
	Machine Learning Algorithm -3 ANN	Tuesday (6-7PM)
	Wednesday TA Office Hour	Wednesday (6-7:30PM)
	Thursday TA Office Hour	Thursday (6-7 PM)
Week Six	**大作业讲解** Kaggle competition-Airbnb run through Classifiers code realization common feature engineering methods data insights exploration label encoding & one hot encoding use GridSearchCV to do model parameter tuning	Saturday (3:30-4:30PM)
	Unsupervised learning: dimension reduction 1. Dimension reduction overview 2. Linear projection methods 3. Manifold learning	Saturday (5-7PM)
	Data Processing using Spark SQL and DataFrame 1. Spark introduction 2. Spark SQL & data frame 3. Spark for data analytics	Sunday (5-7PM)
	Monday TA Office Hour	Monday (6-7:30PM)
	Machine Learning Algorithm -4	Tuesday

	RNN&CNN	(6-7PM)
	Wednesday TA Office Hour	Wednesday (6-7:30PM)
	Thursday TA Office Hour	Thursday (6-7 PM)
Week Seven	**大作业讲解** kaggle competition-AllState run through common feature engineering methods Regression code realization log-transform; box cox; standard scaling	Saturday (3:30-4:30PM)
	Unsupervised learning: clustering and outlier detection 1. Unsupervised learning introduction 2. Clustering methods & techniques 3. Outlier and anomaly detection	Saturday (5-7PM)
	Machine Learning using Spark MLLib	Sunday (5-7PM)
	Monday TA Office Hour	Monday (6-7:30PM)
	Machine Learning Algorithm -5 Decision Tree	Tuesday (6-7PM)
	Wednesday TA Office Hour	Wednesday (6-7:30PM)
	Thursday TA Office Hour	Thursday (6-7 PM)
Project 1	**Week 7 to Week 9**	
Project 2	**Week 10 to Week 12**	
Kaggle Project	**Week 7 to Week 12 (Three classes on week 7 to 9)**	