Data Application Lab

Data Scientist 1711 Training Syllabus – Student

Technical Class:

- Saturday 5-7 pm PST
- Sunday 5-7 pm PST
- Tuesday 6-7 pm PST

Office Hour:

- Wednesday & Friday 5-7 pm PST in first 8 weeks

Mini Project Problem Solving Session:

- Saturday 3:30 – 4:30 pm PST from week 2 ~ week 8

Week	Content
Week 1	**Introduction to Data Application** 1. Data science project lifecycle 2. Cluster and distributed computing 3. Hadoop Eco-system 4. HDFS 5. Basic Linux Operation
	Python Data Analytics Eco-system 1. What is data scientist 2. Key data structures in Python & Numpy 3. Pandas for data analytics • Importing data into Python • Exploring dataset • Renaming the columns of a DataFrame • Filtering a Data Frame • Basic operations with a Data Frame • … 4. Fast data visualization in Pandas
	Statistical Foundations 1. Probability Distribution: Normal, Binomial, χ^2… 2. Central Limit Theorem 3. Bayes’ Theorem 4. Conditional Probability 5. Hypothesis Testing: Confidence interval, T-test… 6. Sampling: proportion sampling, t-distribition… 7. Statistical modeling …

Week 2	Mini Project 1 Session: Sberbank Data Manipulation with Pandas
	Best Practice in Data Processing 1. The importance of data quality 2. The data formats and types 3. The use of RE and BeautifulSoup to collect data from webpage 4. Regular Data Cleaning skills on missing data and outliers
	Python Machine Learning Eco-system 1. Machine learning introduction • The basic concept of machine learning • Differences between Supervised and Unsupervised machine learning • What can supervised and unsupervised learning do 2. Full machine learning flow in Python 3. Scikit-learn package 4. Basic use of sklearn to build simple regression model 5. What is Cross Validation
	Machine Learning Algorithm -1 **Brief Introduction to Machine Learning Algorithm** 1. Supervised Machine Learning vs. Unsupervised Machine Learning 2. Regression vs. Classification 3. Evaluation Methods for regression 4. Evaluation Methods for classification: • how to generate confusion matrix; • how to generate ROC and calculate AUC 5. Basic principles of linear regression and logistic regression

Week 3	Mini Project 2 Session : Data Cleansing Practice on Zillow Data
	Data Analysis using Hadoop Hive 1 1. The basic hive concept: • What is hive • How hive works • Hive architecture 2. Basic operation of hiveQL
	Supervised Learning: Classification 1. Evaluation Methods of classification 2. Basic classification model: logistic regression, decision tree 3. Classification Types (how binary and multi-class works) 4. Ensemble model method: • Bagging • Boosting • Stacking
	Machine Learning Algorithm -2 **SVM Classifiers** 1. Basic principles of SVM 2. Know the procedures to derive SVM 3. What are Kernels and kernel tricks 4. Some basic Kernels such Gaussian Kernel 5. Some important parameters such as the slack variable 6. What kind of problems can be solved by SVM

Week 4	**Mini Project 3 Session: Bank Fraud Detection (binary classification)**
	Analysis using Hadoop Hive 2 1. What is partition table 2. The differences between external and internal table 3. Advanced use of HiveQL 4. The basic use of HiveQL in Spark
	Supervised Learning: Regression 1. Basic concept of Regression 2. Bias-Variance trade off 3. Underfitting vs. Overfitting 4. Linear regression analytical solution 5. Regularization: • Lasso • Ridge • Elastic-Net • Pros and cons of L1 and L2 regularization 6. Advanced techniques in regression • Gradient Descendent • Coordinated Descendent • Stochastic Gradient Descendent • Random sample consensus (RANSAC)
	Machine Learning Algorithm -3 **ANN** 1. Basic structure of ANN • Neuron • Perceptron 2. Activation function and the common activation functions 3. Procedures of forward propagation and backward propagation 4. Derivation of ANN

Week 5	**Mini Project 4 Session: History Kaggle Demo: Allstate Claims Severity**
	Data Visualization with Tableau 1. Hands-on data visualization & Analysis on Tableau **2.** Business Insights Extraction
	Advanced visualization & A/B Testing 1. Basic & interactive visualization in Python 2. Levels of visualization 3. Matplotlib • Basic elements • Visualization for distribution (histogram, pie chart…) • Visualization for bi-variable relationship on continuous and categorical features 4. Seaborn 5. Exploratory analysis 6. A/B test and Experimentation • Business need • Design Experiment • Power Analysis • Analyzed Result
	Machine Learning Algorithm -4 **CNN & RNN** 1. What can CNN and RNN work for 2. Basic CNN architecture: Convolutional, RELU, Pooling and Fully Connected Layer 3. Basic RNN architecture: Forward Propagation and Backward Propagation in RNN 4. RNN Example in add operation

Week 6	Mini Project 5 Session: Data Visuliazation with Duoligo User Datasets
	Data Processing using Spark SQL and DataFrame 1. Spark introduction • What is Spark • Why Spark is better 2. Spark SQL & data frame 3. Spark for Data Analytics 4. Demos to fully practice
	Unsupervised Learning: Dimension Reduction 1. Dimension reduction overview 2. Dimension reduction methods • Randomized Projection • Principal Component Analysis o PCA Calculation o Randomized PCA o Sparse PCA 3. Manifold learning 4. Multidimensional Scaling • MDS • Isomap
	Machine Learning Algorithm -5 **Decision Tree & Ensemble Methods** 1. Details in Decision Tree • How Decision Tree works • Measures to select the best split 2. Details in Ensemble Methods: • Why do we need ensemble model • Committees, Weighted, Predictor of Predictors • Bagging and Boosting 3. Random Forest Tree 4. Gradient Boosting 5. Adaboost

Week 7	**Mini Project 6 Session: History Kaggle Demo: Airbnb New User Bookings**
	Machine Learning using Spark MLLib 1. Relational Database and No-SQL Database 2. Graph Analytics • What is graph database and its applications • Spark GraphX/GraphFrame 3. Machine Learning in Spark 4. Demo Practice by PySpark
	Unsupervised Learning: Clustering and Outlier Detection 1. Unsupervised learning introduction 2. Clustering methods & techniques • K-mean Algorithm • Hierarchical Clustering Algorithm • DBSCAN algorithm 3. Outlier and anomaly detection
	Advanced Python **Basic CS Algorithm -1** 1. Basic data structure 2. What is Algorithm: • Algorithm Analysis (Time and Space Efficiency) • Theoretical Analysis and Asymptotic Notation (Big-Θ, Big-Ω and Big-O) • Master Theorem 3. Search Algorithm • Sequential search • Binary search 4. Sort Algorithm: Bubble Sort, Selection Sort, Insertion Sort, Shell Sort, Count Sort, Merge Sort 5. Divide and Conquer: Quick Sort

Week 8	Mini Project 7 Session: PySpark Machine Learning
	Real Case Data Processing & Machine Learning in R Use the skills of what we have learned in R
	Deep Learning 1. Neutal Network Anatomy 2. Uniform approximator 3. CNN & RNN 4. LSTM (Long-Short-Term-Memory) 5. Use Keras to build the neural network
	Advanced Python **Basic CS Algorithm -2** 1. Dictionary 2. Hashing: • Hash Code Conventions • Hash Code Design • Collision • Bucketing • Separate Chaining 3. Linear Probing 4. Quadratic Probing 5. Double Hashing

Kaggle

Week 9	**Kaggle Introduction** 1. What is Kaggle 2. Why we need to attend Kaggle 3. Tools we will use in Kaggle 4. The basic precedures in Kaggle (Feature Engineering, Parameter Tuning, Model Ensemble …)
	Kaggle 1 (We will cover the following topic with the data of Kaggle Topic we choose) 1. Exploratatory Data Analysis (Data Types, Distributions, Missing Values, Correlations …) 2. Validation Method 3. Feature Engineering on: • Numetic Features (Log Transformation, Standardization …) • Categortical Features (One Hot Encoding, Label Encoding, Mean Response Encoding…) • Missing Value • Interactions 4. Feature Selection Methods: • Low-variance & High-correlation filters • Recursive-feature-elimination
Week 10	**Kaggle 2 (We will cover the following topic with the data of Kaggle Topic we choose)** 1. XGBoost highlights and its parameters 2. Tuning Process of XGBoost 3. Grid Searching and bayesian optimization
Week 11	**Kaggle 3 (We will cover the following topic with the data of Kaggle Topic we choose)** 1. LightGBM highlights and its parameters 2. Model Ensemble Details 3. Blending with demo code

NLP

Week 9	**NLP 1** 1. Basic NLP Introducton 2. The Naïve Bayian Algorithm in NLP
Week 10	**NLP 2** 1. Detailed Coding
Week 11	**NLP 3** 1. Web Application Development by Flask 2. Enhanced NB 3. Negation Handling 4. Advanced algorithm such as RNN

FinTech

Week 12	**FinTech 1** 3. FinTech Domain Knowledge 4. Introduction of Lending Club 5. How to request data by Lending Club API 6. Features at a first look 7. Data Preparation
Week 13	**FinTech 2** 2. Feature Engineering 3. Baseline Model by Logisic Regression 4. Gradient Boost Example 5. Insights for this project
Week 14	**FinTech 3** 5. Web Application Development by Flask • Pickling • Routing • Rendering • Js Basics

Recommendation System

Week 15	**Recommendation System 1** 1. Introduction for this project and its insights 2. Understand the API and the use of Github 3. Learn how to use API to crawl data from Steam 4. Basic Function of Requests and BeautifulSoup
Week 16	**Recommendation System 2** 1. How to use the Github to do the version control 2. The HTTP Basics 3. The API Basics and request structure 4. Process the raw data and set up our own database 5. The knowledge of collaborative filtering 6. The knowledge of content-based filtering 7. The knowledge of popularity based recommendation
Week 17	**Recommendation System 3** 1. Build a recommender engine in Spark 2. Build your own demo with Python Flask