DataSci 420: Supplemental (Optional) Reading BETA for Spring 2020

Spring 2020

About This Document

- The intent is that if a topic isn't quite clear to you, you can consult one of the sources/links below.
- That is, if you read the lesson materials & attended the lesson, and a topic still isn't quite clear to you, you can consult one of the links below.
- The reading is organized by lesson, starting on page 5 below.
- For each lesson, there are generally 3-8 links to supplemental reading.
- For clarity, this is supplemental/optional reading - it is not required reading.
- Most of the links below go to O'Reilly books, which are freely available to you online via your UW netID.

Main Textbooks

All of the books are available for free to UW students via https://learning.oreilly.com/home/

Title	Python Data Science Handbook
Author	Jake VanderPlas
Publisher	O'Reilly Media, Inc.
ISBN	9781491912058
Abbreviation	PSDH

Lesson 1: Introduction to Machine Learning & the Data	
Lesson 2: Pitfalls of Machine Learning (and how to avoid them)	
Lesson 3: Feature Engineering	
Lesson 4: Feature Selection	
Milestone 1: Capstone - Project Definition	
Lesson 5: Decision Trees	• Chapter 5 - In-Depth - Decision Trees and Random Forests
Lesson 6: Ensemble Models	
Lesson 7: Support Vector Machines	• Chapter 5 - In-Depth: Support Vector Machines
Milestone 2: Capstone - Initial Models	
Lesson 8: Neural Networks	
Lesson 9: Deep Learning I: Deep Neural Networks	
Lesson 10: Deep Learning II: Recurrent Neural Networks	
Milestone 3: Capstone - Final Project	

Title	Data Science from Scratch, 2nd Edition
Author	Joel Grus
Publisher	O'Reilly Media, Inc.
ISBN	9781492041139
Abbreviation	DSS

Lesson 1: Introduction to Machine Learning & the Data	
Lesson 2: Pitfalls of Machine Learning (and how to avoid them)	• Chapter 11 - Overfitting and Underfitting • Chapter 11 - The Bias-Variance Tradeoff • Chapter 15 - Regularization
Lesson 3: Feature Engineering	
Lesson 4: Feature Selection	
Milestone 1: Capstone - Project Definition	
Lesson 5: Decision Trees	• Chapter 17 - Decision Trees
Lesson 6: Ensemble Models	
Lesson 7: Support Vector Machines	
Milestone 2: Capstone - Initial Models	
Lesson 8: Neural Networks	• Chapter 18 - Neural Networks • Chapter 19 - Deep Learning
Lesson 9: Deep Learning I: Deep Neural Networks	
Lesson 10: Deep Learning II: Recurrent Neural Networks	• Chapter 21 - Recurrent Neural Networks
Milestone 3: Capstone - Final Project	

Additional Textbooks

All of the books below are available for free to UW students via https://learning.oreilly.com/home/

Title	Principles of Data Science - Second Edition
Author	Sunil Kakade, Sinan Ozdemir
Publisher	Packt Publishing
ISBN	9781789804546

Title	Python Data Science Essentials - Third Edition
Author	Luca Massaron, Alberto Boschetti
Publisher	Packt Publishing
ISBN	9781789537864

Title	Python Feature Engineering Cookbook
Author	Soledad Galli
Publisher	Packt Publishing
ISBN	9781789806311

Title	Practical Statistics for Data Scientists, 2nd Edition
Author	Peter Gedeck, Andrew Bruce, Peter Bruce
Publisher	O'Reilly Media, Inc.
ISBN	9781492072942

Title	Python Machine Learning - Third Edition
Author	Vahid Mirjalili, Sebastian Raschka
Publisher	Packt Publishing
ISBN	9781789955750

Title	Visual Models for Software Requirements
Author	Joy Beatty and Anthony Chen
Publisher	Microsoft Press
ISBN	9780735667730

Title	Hands-On Machine Learning with Scikit-Learn and TensorFlow
Author	Aurélien Géron
Publisher	O'Reilly Media, Inc.
ISBN	9781491962299

Lesson 1: Introduction to Machine Learning & the Data Science Process

Supplemental (Optional) Materials

- Learning objectives
    - Explain why a standardized data science process should be followed: Wikipedia - CRISP-DM, Why using CRISP-DM will make you a better Data Scientist
    - Demonstrate the steps of a standard data science process: Wikipedia - CRISP-DM, Why using CRISP-DM will make you a better Data Scientist
- Assignment
    - Optimizing a Manufacturing Process: Overview of the five steps, Why using CRISP-DM will make you a better Data Scientist

Lesson 2: Pitfalls of Machine Learning (and how to avoid them)

Supplemental (Optional) Materials

- Learning Objectives
  - Avoid overfitting and underfitting: DSS Chapter 11 - Overfitting and Underfitting, DSS Chapter 11 - The Bias-Variance Tradeoff, DSS Chapter 15 - Regularization
  - Solve class imbalance problems: Chapter 5 - Strategies for Imbalanced Data, Dealing with class imbalance (scroll down)
- Assignment
  - Good and Bad Connections: Chapter 5 - Evaluating Classification Models, DSS Chapter 11 - Overfitting and Underfitting, Chapter 5 - Strategies for Imbalanced Data, Dealing with class imbalance (scroll down)

Lesson 3: Feature Engineering

Supplemental (Optional) Materials

- Learning Objectives
    - Perform one-hot encoding of the categorical variables: Creating binary variables through one-hot encoding, Guide to Encoding Categorical Values in Python
    - Calculate the risk values of categorical variables, and use risk values to replace categorical variables: *no supplemental reference for risk values, but here is a description of similar concept - Weight of Evidence
    - Apply RFM feature engineering framework to generate features on some applicable scenarios: Recency, Frequency, Monetary Model with Python, RFM Analysis Tutorial
- Assignment
    - Customer Segmentation: Recency, Frequency, Monetary Model with Python

Lesson 4: Feature Selection

Supplemental (Optional) Materials

- Learning Objectives
  - Apply filter-based, step-wise, and embedded (LASSO) methods to select features: Feature selection, Selecting meaningful features, An Introduction to Feature Selection
  - Summarize why we need to select features: Feature selection, Selecting meaningful features, An Introduction to Feature Selection
  - Explain why LASSO can select features, but Ridge Regression cannot: Ridge and Lasso Regression: L1 and L2 Regularization
- Assignment
  - Predictive Policing: Feature selection, Selecting meaningful features, An Introduction to Feature Selection

Milestone 1: Capstone - Project Definition

Supplemental (Optional) Materials

- Chapter 20 - Data Flow Diagram
- Chapter 5 - Strategies for Imbalanced Data
- Dealing with class imbalance (scroll down)
- Feature selection
- Selecting meaningful features
- An Introduction to Feature Selection

Lesson 5: Decision Trees

Supplemental (Optional) Materials

- Learning Objectives
    - Create a decision tree for performing analysis of a dataset: PDSH Chapter 5 - In-Depth - Decision Trees and Random Forests, DSS Chapter 17 - Decision Trees, Decision tree learning (scroll down)
    - Identify and select different splitting methods: PDSH Chapter 5 - In-Depth - Decision Trees and Random Forests, DSS Chapter 17 - Decision Trees, Decision tree learning (scroll down)
    - Prune a decision tree via a few methods: Decision tree learning (scroll down)
- Assignment
    - Breast Cancer Detection: PDSH Chapter 5 - In-Depth - Decision Trees and Random Forests, DSS Chapter 17 - Decision Trees, Decision tree learning (scroll down), Validation metrics

Lesson 6: Ensemble Models

Supplemental (Optional) Materials

- Learning Objectives
  - Define two different mechanisms of ensembling: bagging and boosting: Ensemble methods: bagging, boosting and stacking, Bagging and the Random Forest, Boosting, Bagging – building an ensemble of classifiers from bootstrap samples (scroll down), Leveraging weak learners via adaptive boosting (scroll down)
  - Train random forest models, and tune the hyperparameters of random forest models: PDSH Chapter 5 - In-Depth - Decision Trees and Random Forests, Decision tree learning (scroll down)
  - Explain why ensemble models perform better than base learners: Ensemble methods: bagging, boosting and stacking, PDSH Chapter 5 - In-Depth - Decision Trees and Random Forests
  - Summarize how AdaBoost works: PDSH Chapter 5 - In-Depth - Decision Trees and Random Forests, Leveraging weak learners via adaptive boosting (scroll down)
  - Train gradient boosted decision tree models, and tune the hyperparameters of the gradient boosted decision tree models: Ensemble methods: bagging, boosting and stacking, Bagging and the Random Forest
- Assignment
  - Targeted Marketing: PDSH Chapter 5 - In-Depth - Decision Trees and Random Forests, How To Visualize A Decision Tree In 5 Steps

Lesson 7: Support Vector Machines

Supplemental (Optional) Materials

- Learning Objectives
  - Distinguish the uses of SVMs as compared with linear regressions: Chapter 5 - Support Vector Machines, PDSH Chapter 5 - In-Depth: Support Vector Machines
  - Know when you would use support vector classification (SVC) as opposed to support vector regression (SVR): Chapter 5 - Support Vector Machines
  - Summarize how to tune an SVM's hyperparameters and evaluate the models for accuracy: Chapter 5 - Support Vector Machines
- Assignment
  - Age Classification of Abalone: Chapter 5 - Support Vector Machines, PDSH Chapter 5 - In-Depth: Support Vector Machines, Tuning the hyper-parameters of an estimator, sklearn.model_selection.GridSearchCV

Milestone 2: Capstone - Initial Models

Supplemental (Optional) Materials

- PDSH Chapter 5 - In-Depth - Decision Trees and Random Forests
- Bagging and the Random Forest
- Chapter 5 - Support Vector Machines
- PDSH Chapter 5 - In-Depth: Support Vector Machines

Lesson 8: Neural Networks

Supplemental (Optional) Materials

- Learning Objectives
  - Explain the foundations on which neural networks are base: DSS Chapter 18 - Neural Networks, Chapter 10 - Introduction to Artificial Neural Networks
  - Describe the structure and hyperparameters of a neural network, such as depth of architecture and learning rate: Chapter 10 - Introduction to Artificial Neural Networks, Chapter 11 - Training Deep Neural Nets, DSS Chapter 19 - Deep Learning
- Assignment
  - Wine Classifier: Chapter 10 - Introduction to Artificial Neural Networks, Chapter 11 - Training Deep Neural Nets, DSS Chapter 19 - Deep Learning, How to build your own Neural Network from scratch in Python

Lesson 9: Deep Learning I: Deep Neural Networks

- Learning Objectives
  - Differentiate among the types of deep neural networks and their uses: Chapter 11 - Training Deep Neural Nets, Chapter 13 - Convolutional Neural Networks, Chapter 14. Recurrent Neural Networks
  - Know the difference between symbolic and imperative programming: Symbolic vs. Imperative Programs
  - Explain how a convolutional neural network can perform computer vision applications: Chapter 13 - Convolutional Neural Networks
- Assignment
  - Imagine Identification: Chapter 10 - Introduction to Artificial Neural Networks, Chapter 11 - Training Deep Neural Nets

Lesson 10: Deep Learning II: Recurrent Neural Networks

- Learning Objectives
  - Explain the theory behind recurrent neural networks: Chapter 14 - Recurrent Neural Networks, DSS Chapter 21 - Recurrent Neural Networks
  - Apply recurrent neural networks for text analysis: Multi-Class Text Classification with LSTM
- Assignment
  - News Article Classification: Chapter 14. Recurrent Neural Networks, Multi-Class Text Classification with LSTM

Milestone 3: Capstone - Final Project

Supplemental (Optional) Materials

- Chapter 11 - Training Deep Neural Nets
- Chapter 13 - Convolutional Neural Networks
- Chapter 14. Recurrent Neural Networks