MA477
Theory and Applications of Data Science – Course Outline

Valmir Bucaj, PhD

Contents

DEPARTMENT OF THE ARMY
UNITED STATES MILITARY ACADEMY
West Point, New York 10996

REPLY TO
ATTENTION OF

MADN-MATH

1 Instructional Memorandum # 477-20-02

MA477 – Theory and Applications of Data Science

1.1 Purpose

This memorandum specifies materials, describes the goals and objectives, and announces policy and procedures for MA477 during AY 20-2.

1.2 Instructor Info

Instructor: Dr. Valmir Bucaj
Office Nr.: TH224
Ext: 7460
Email: valmir.bucaj@westpoint.edu

1.3 Course Website & Classroom Info & Meeting Time

- **Classroom:** TH319
- **Meeting time:** 0855 – 1010

All of the materials for this course will be published and made available in our course website, which may be found in the link below:

Course Website URL

1.4 Textbooks and Supporting Materials

For this course you are required to have the following two textbooks:

(1) Machine Learning: An Applied Mathematics Introduction by Paul Wilmott

(2) An Introduction to Statistical Learning by G. James, D. Witten et.al.

While you are welcomed to purchase both books, the second one may be downloaded for free from the author's webpage by Clicking Here!.

Extra Resources
You are also strongly encouraged to create an account with Data Camp and use it as a supporting resource. To open an account follow the link DataCamp. In addition to posting the notes on our course website, I will also post all of the JupyterNotebooks on my GitHub page, which you may find HERE!.

1.5 Technology Requirements

For the entirety of this course we will exclusively be using Python as our programming language.

While you are welcome to use a desktop Python IDE, you are strongly encouraged to use JupyterNotebooks, as this is what we will almost exclusively use throughout the course. You are strongly suggested to download Anaconda.

Anaconda is a free Python distribution which contains over 1500 open source packages, as well as Python itself, for performing scientific computation. In this case we will use this distribution, its packages, the Spyder IDE, and Jupyter Notebooks (all of this is included in this one download / install).

To install Anaconda follow the steps below:

(1) Click on the Anaconda website!

(2) Download the file that has Python 3.xx

(3) In the Programs Folder or Start-up Menue find Anaconda 3 folder

(4) Click on Anaconda Navigator

(5) Locate JupyterNotebook and click Launch

You are also required to open a GitHub account. To do so Click Here!

1.6 Course Goals

The purpose of this course is to build a solid foundation in the theoretical and applied aspects of Machine Learning as well as an appreciation for its vast applications in a variety of fields across many domains. It will introduce a variety of supervised and unsupervised machine learning techniques with applications in Python.

1.7 Course Evaluation Plan

Final course grades will follow department standing guidelines. Cadets can view their performance on all graded events in CIS.

The weight of each graded portion will roughly be as follows:

Event	**Count**	**Points**	**Weight (%)**
Exercises	3-5	150	15
Quizzes	3-5	200	20
Mini-Projects	5-7	400	40
TEE	1	250	25
Total		1000	100

1.7.1 Exercises

After completing most sections from the module **Python for Data Analysis and Visualization**, you will be assigned a short set of **JupyterNotebook** Exercisses. The purpose of these exercises will be to reinforce the concepts taught in class as well as give students an opportunity to go beyond what was taught in class.

1.7.2 Quizzes

These will be $15-20$ minute in-class concept quizzes. They will exclusively focus on theoretical concepts covered up to that point, and as such technology will not be permitted. The rough idea is for the first quiz to be administered after the section on **Regression Models**, the second after the section on **Classification Models**, the third after the chapter on **Unsupervised Learning Methods**, the fourth after **Neural Networks** and possibly a fifth somewhere in the mix as we see fit.

1.7.3 Mini-Projects

These will be individual mini JupyterNotebook Projects, and they will be assigned after most, if not each, of the Machine Learning Methods covered in class. For example, there will be a mini-project about the Python for Data Analysis and Visualization module, another one about the K-NN Regressor, and another about Linear Regression, and so on.

The purpose of these mini-projects is to provide students wit an opportunity to:

- Utilize the knowledge acquired in class up to that point
- Independently investigate a problem
- Reinforce the concepts already learned
- Be creative and show originality in thought and in implementing various ML tools.

I strongly believe that constructive competition drives progress, so in line with this mantra, the mini-projects will be of a competitive nature, as well, where your individual ranking in the competition will determine your score on the project! More specific details will be issued with each project!

1.7.4 TEE

The TEE for this course will consist of two parts:

1) An oral presentation

2) A questions and answers session

At least one week before the end of the semester you will be assigned a project with specific guidelines which will consist of your TEE for the course. On the day of the TEE you will each give a 10-15 minute presentation which will be followed up by a 10-15 minute QA session.

1.8 Professional Development

To help you achieve the student growth goals stated above, we have the following expectations for your efforts.

a. **Data Science is not a Spectator Sport.** To be successful in Data Science, we have to do the work. Often, we fall into the trap of seeing a professor or peer do a problem and think "Oh, that's how you do it" and we move on. That approach will not be good enough in this course. We will not grade you on your ability to recognize the right answer, rather on your ability to produce it. To summarize a key point from Malcolm Glidewell's book, Outliers, it takes many repetitions to get good at something. Do the reps. There's simply no substitute.

b. **Be Prepared.** Cadets have full days. Spending time with the material before class will enable your days to be more efficient and more productive. Some material in MA477 is material you may have seen in other courses. Some is relatively straight forward. Other material is challenging. With your prior preparation, your professors can provide light(er) treatment on the material you've seen before and/or is easily attainable and heavily weight class time toward the more challenging concepts. We can help you get more reps on the hard material, enabling your excellence now and in the future.

c. **Build a Team.** Ray Kroc, the founder and first CEO of McDonald's said "All of us is better than any of us." West Point and the Army are team activities. Build your Data Science team and use it to maximize your success. Study together, do exercises together, and brief each other. Seek additional resources when needed, including additional instruction (AI) from your professor. The primary purpose of AI is to address specific questions that you may have; it is not a lesson reteach.

d. **Show Your Pride.** Understand what the expectations are for each task you are given and meet those expectations to the best of your ability. Do your best work. Do it neatly.

e. **Be on Time.** When you are given a deadline, meet it.

1.9 Classroom Rules

All cadets are expected to maintain proper military bearing and appearance during instruction in accordance with appropriate regulations.

a. Respect others' rights in the classroom. No profanity, unprofessional jokes, or unprofessional computer items (audio or visual). If what you are doing is distracting anyone else in the class, what you are doing is wrong.

b. Store backpacks in the hallway.

1.10 Scholarship

This course is designed to build upon your previous instruction in mathematics and statistics as well as to explore new concepts. It is therefore appropriate (and you are encouraged) to consult sources of information beyond the course text. In doing so, keep in mind the standards of scholarly work. Documentation of your effort should follow the guidance contained in Documentation of Written Work, published by the Dean's Office, June 2015. If you are not sure about how or when to document a source, discuss the matter with your instructor prior to submitting your work.

1.11 Your Learning.

Finally, this is your education. You must take responsibility for your own learning and participate as an active learner. Your professor is here as a resource to help guide you through the educational experience this course offers, but he is only one resource. There are many others. The quantitative reasoning and problem solving skills that you continue to develop in this course will serve you in the future regardless of your major as a Cadet or your branch as an officer. Use this opportunity to develop confidence, competence, and good habits of mind. These are the ends which we will use mathematics and machine learning as a means to reach.

1.12 Chain of Command

The Department of Mathematical Sciences has an open door policy. Your initial attempt to address problems or concerns should go through the chain of command. However, if you are uncomfortable addressing an issue with your instructor, please approach any individual in the chain. If you have any concerns related to respect in the classroom, unprofessional behavior, or sexual assault or harassment, please feel free to contact the Department Head directly. Below is the chain of command for MA477:

a. **Your Instructor:** Dr. Valmir Bucaj

b. **Elective Mathematics Program Director:** COL Watts, TH239A, (845) 938-2276

c. **Head of the Department:** COL Tina Hartley, TH 238, (845) 938-5285

1.13 Course Calendar

MA477: Theory and Applications of Data Science - AY20-2 (Spring 2020) **Printed on: 3/30/2020**

AY19-2		MONDAY		TUESDAY		WEDNESDAY		THURSDAY		FRIDAY	
Block 1 - Python for Data Analysis and Visualization	1	6-Jan-20	N/A	7-Jan-20	N/A	8-Jan-20	1-1	9-Jan-20	2-1	10-Jan-20	1-2
		No Class Reorgy Week		No Class Reorgy Week		No Class		Course Introduction and Intro to Machine Learning §2.1 of ISLR and §2.1-2.4 of Wilmott			
	2	13-Jan-20	2-2	14-Jan-20	N/A	15-Jan-20	1-3	16-Jan-20	2-3	17-Jan-20	1-4
		Intro to ML Continued &Python for Data Analysis -- Numpy		No Class Honorable Living Stand Down Day				Python for Data Analysis Pandas			
	3	20-Jan-20	N/A	21-Jan-20	2-4	22-Jan-20	N/A	23-Jan-20	1-5	24-Jan-20	2-5
		No Class MLK Day		Python for Data Analysis Pandas Continued Exercise 1		No Class Study Day		No Class		Python for Data Visualization Matplotlib & Seaborn Exercise 2	
Block 2 - Regression Models	4	27-Jan-20	1-6	28-Jan-20	2-6	29-Jan-20	1-7	30-Jan-20	2-7	31-Jan-20	1-8
		No Class		RegModels - K-NN Regressor §3.1 -- §3.7 of Wilmott Exercise 3		No Class		RegModels K-NN Regressor Lab		No Class	
	5	3-Feb-20	1-9	4-Feb-20	2-8	5-Feb-20	1-10	6-Feb-20	2-9	7-Feb-20	1-11
		No Class		RegModels Linear Regression Ch. 3 of ISLR & Ch. 6 of Wilmott		No Class		RegModels Linear Regression Lab		No Class	
	6	10-Feb-20	2-10	11-Feb-20	1-12	12-Feb-20	N/A	13-Feb-20	1-13	14-Feb-20	2-11
		RegModels - Ridge & Lasso Regression §3.1 of ISLR		No Class		No Class Study Day		No Class		ClassModels - KNN Classifier Ch. 3 of Wilmott and §2.2.3 of ISLR Lab	
Block 3 - Classification Models	7	17-Feb-20	N/A	18-Feb-20	1-14	19-Feb-20	2-12	20-Feb-20	1-15	21-Feb-20	2-13
		No Class President's Day		No Class		ClassModels - Logistic Regression §4.3 of ISLR and §6.4 of Wilmott Mini-Project 1		No Class		Logistic Regression §4.3 of ISLR and §6.4 of Wilmott Lab	
	8	24-Feb-20	1-16	25-Feb-20	2-14	26-Feb-20	N/A	27-Feb-20	2-15	28-Feb-20	1-17
		No Class		ClassModels - Naïve Bayes Classifier Ch. 5 of Wilmott		No Class Study Day		Application of Naïve Bayes Method Intro to Natural Language Processing (NLP) Quiz 1		No Class	
	9	2-Mar-20	1-18	3-Mar-20	2-16	4-Mar-20	1-19	5-Mar-20	2-17	6-Mar-20	1-20
		No Class		NLP Tokenization, Tagging, Chunking Mini-Project 2		No Class		Sentiment Analysis Text Analysis and Prediction Lab		No Class	
Spring Break		9-Mar-20	N/A	10-Mar-20	N/A	11-Mar-20	N/A	12-Mar-20	N/A	13-Mar-20	N/A
		No Class Spring Break		No Class Spring Break		No Class Spring Break		No Class Spring Break		No Class Spring Break	
	10	16-Mar-20	1-21	17-Mar-20	2-18	18-Mar-20	N/A	19-Mar-20	2-19	20-Mar-20	1-22
		No Class				No Class Study Day		Decision Trees §8.1 ISLR and Ch. 9 of Wilmott		No Class	
Block 4 - Unsupervised Learning	11	23-Mar-20	1-23	24-Mar-20	2-20	25-Mar-20	1-24	26-Mar-20	2-21	27-Mar-20	1-25
		No Class		Ensamble Models & Random Forests §8.2 of ISLR		No Class		Ensemble Models Voting Classifiers		No Class	
	12	30-Mar-20	2-22	31-Mar-20	1-26	1-Apr-20	N/A	2-Apr-20	1-27	3-Apr-20	2-23
		Boosting Models Adaptive Boosting		No Class		No Class Study Day		No Class		K-Means Clustering § 10.3.1 of ISLR and Ch. 4 of Wilmott Quiz 2	
	13	6-Apr-20	1-28	7-Apr-20	2-24	8-Apr-20	N/A	9-Apr-20	1-29	10-Apr-20	2-25
		No Class		H-Means Clustering § 10.3.1 of ISLR		No Class Study Day		No Class		Principal Component Analysis § 10.2 of ISLR and § 2.5 of Wilmott	
Block 5 - ANNs and Reinforcment Learning	14	13-Apr-20	1-30	14-Apr-20	2-26	15-Apr-20	1-31	16-Apr-20	2-27	17-Apr-20 Modified	1-32
		No Class		Neural Networks Ch. 10 of Wilmott Mini-Project 3		No Class		Neural Networks		No Class	
	15	20-Apr-20	1-33	21-Apr-20	2-28	22-Apr-20	N/A	23-Apr-20	1-34	24-Apr-20	2-29
		No Class		Neural Networks Quiz 3		No Class Study Day		No Class		Special Topic TBD Mini-Project 4	
	16	27-Apr-20	1-35	28-Apr-20	N/A	29-Apr-20	1-36	30-Apr-20	N/A	1-May-20	1-37
		No Class		No Class Study Day		No Class		No Class Projects Day		No Class	
	17	4-May-20	1-38	5-May-20	2-30	6-May-20	1-39	7-May-20	N/A	8-May-20	1-40
		No Class		Special Topic TBD		No Class		No Class Study Day		No Class	
TEE	18	11-May-20	N/A	12-May-20	N/A	13-May-20	N/A	14-May-20	N/A	15-May-20	N/A
		No Class TEE Week		No Class TEE Week		No Class TEE Week		No Class TEE Week		No Class TEE Week	

KEY	No Class	Lab	Course-wide Graded Event	Quiz	Assignment or Modified Sched.

2 Intro to Machine Learning – A Brief Overview

2.1 What is ML?

Suggested Reading:

- Section 2.1 of [1]
- Chapter 1 and Sections 2.1 – 2.4 of [2].

2.1.1 Estimating the relationship between *predictor* and *response*

2.1.2 Prediction Accuracy vs. Model Interpretability

2.1.3 Supervised vs. Unsupervised Learning

2.1.4 Regression vs. Classification Models

2.2 Assessing Model Accuracy

Suggested Reading:

- Sections 2.7 and 2.11 of [2]
- Sections 2.2.1 and 2.2.2 of [1]

(a) Classification Setting

- Confusion Matrix and ROC Curve
- Cross-Validation

(b) Regression Setting

- Mean Squared Error

2.3 Bias - Variance Trade-Off

3 Python for Data Analysis and Visualization

In this chapter we will talk

3.1 Numpy

3.1.1 Numpy Arrays and Indexing

3.1.2 Numpy Operations

3.2 Pandas

3.2.1 Series

3.2.2 Data Frames

3.2.3 Missing Data, Groupby, and Basic Operations

3.2.4 Merging, Joining, and Concatenating

3.2.5 Reading and Writing Data

3.3 Matplotlib

3.3.1 Basic Plotting Methods

3.3.2 Object Oriented Plotting Methods

3.4 Seaborn

3.4.1 Distribution and Categorical Plots

3.4.2 Matrix Plots

4 Supervised Learning Methods

4.1 Regression Models

4.1.1 K-Nearest Neighbors Regressor

Suggested reading:

- Chapter 3 of [2]

4.1.2 Linear Regression – Least Squares

Suggested Reading:

- Chapter 3 of [1]
- Sections 6.1–6.3 of [2]

4.1.3 Ridge and Lasso Regression

Suggested Reading:

- Section 6.3 of [1]

4.1.4 (Optional) Random Forest Regressor

4.2 Classifications Models

Suggested Reading:

- Sections 4.1 and 4.2 of [1]

4.2.1 Logistic Regression

Suggested Reading:

- Section 4.3 of [1]
- Sections 6.4 and 6.5 of [2]

4.2.2 K-Nearest Neighbors Classifier

Suggested reading:

- Chapter 3 of [2]
- Section 2.2.3 of [1]

4.2.3 Naive Bayes Classifier

Suggested Reading:

- Chapter 5 of [2]

4.2.4 Natural Language Processing (NLP)

I. Overview of the Natural Language Tool Kit (NLTK)

- Counting Text
- Frequency Distribution
- Bigrams

II. Overview of Regular Expressions (RegExp)

III. Tokenization

IV. Tagging

V. Chunking

4.2.5 Decision Trees and Random Forests

Suggested Reading:

- Sections 8.1 and 8.2.2 of [1]
- Chapter 9 of [2]

4.2.6 (Optional) Support Vector Machine Classifier (SVC)

5 Unsupervised Learning Methods

Suggested Reading:

- Section 10.1 of [1]

5.0.1 Principal Component Analysis

Suggested Reading:

- Section 10.2 of [1]
- Section 2.5 of [2]

5.0.2 K Means Clustering

Suggested Reading:

- Section 10.3.1 of [1]
- Chapter 4 of [2]

5.0.3 H-Means Clustering

Suggested Reading:

- Section 10.3.2 of [1]

5.0.4 (Optional) T-distributed Stochastic Neighbor Embedding (t-SNE)

6 Neural Networks

Suggested Reading:

- Chapter 10 of [2]

7 Reinforcement Learning

Suggested Reading:

- Sections 11.1 – 11.17 of [2]

8 (Optional) Recommender Systems

References

[1] G. James, D. Witten, T. Hastie, R. Tibshirani *An Introduction to Statistical Learning* Springer, 2014.

[2] P. Wilmott. *Machine Learning: An Applied Mathematics Introduction*. Panda Ohana Publishing. 2019.