2143488 Big Data and Artificial Intelligence

2022 2nd Semester

Class Information

Course Title	Lecture/Lab Room	Class Date-Time
2143488 Big Data and Artificial Intelligence	Room 210, Building 2 (Class Notes: myCourseville)	Mon: 9:00 am -12:00 pm

Instructor Information

Instructor	Email	Office Location & Hours
Dr. Jing Tang	Jing.T@chula.ac.th	Room 204, Building 2 30 mins **before & after** every class session

General Information

Description: 3(3-0-6) Credit

Introduction to data science basic concepts and application of data science; data types; scale of measurement; life cycle of data science project; modelling; evaluation and deployment; exploratory data analysis; summary statistics; presentation and visualization; unsupervised methods; k-means; association rules; clustering evaluation; supervised methods; ensemble methods; classifiers evaluation and comparison; statistical modelling construction and machine learning methods.

Prerequisite: Statistics, Linear Algebra

Objectives:

Data Science is the study of the **generalizable extraction of knowledge from data**. Being a data scientist requires an integrated skill set spanning **mathematics, statistics, machine learning, databases, other branches of computer science and business** along with a good understanding of the craft of problem formulation to engineer effective solutions. This course will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset.

After successfully completing the course, students should be able to do the following:

- ❖ Understand the technical terms related to data science, big data, AI, and machine learning.
- ❖ Familiar with life cycle of data science project, and data science tools.
- ❖ Able to conduct basic data science project: data preparation, demographic data analysis, statistical analysis, forecasting, data visualization, and presentation.
- ❖ Understand core machine learning techniques: regression, time-series, decision tree, clustering, SVM, ANN, text mining, and image mining.

Primary Textbooks:

None

Course Rules

Students should follow the following course rules:

1) Follow postings and announcements made in class and/or on myCourseVille. It will be assumed that anything announced in class and/or on myCourseVille is acknowledged by all students.
2) No noise, no cellphone ring during the class.
3) Submit homework and exercises on time via myCourseVille. No late submission will be accepted. Plagiarism in any form is not allowed (result in "0").
4) No regular attendance checking. Bonus points will be given to students who answer my questions during the class, or who attend the course when I check attendance, and so on. Come to record your bonus point (1 point /.class) right after each class.
5) I reserve the right to make changes to this syllabus as needed.

Evaluation Criteria

Criteria	Points
Lab/HW (5 points / time)	30
Midterm Exam	30
Final Report or Presentation	40

- An "A" will be given to students who having total scores **more than 90 points and top 20% (max);**
- An "F" will be given to students who having total scores **less than 50 points or absent mid-term or final exams**.

Midterm (Examples)

basic definition
- big data

Predict the GDP of Thailand in the next year
- what kind of ML you need
- typically regression, linear regression
- say GDP depends on the population, population growth, land size, etc.
- say GDP depends on the historical GDP of the past year (time-series)
- use lots of x to predict y

Students who will go to masters program
- identify what kind of students will go to masters, what kind will go to work, etc.
- you may find historical data via survey and identify what are the characteristics of the students who go to the masters program
- classification problem
- use the model you trained to test the current students, etc.

Hobbies
- separate the students into groups of hobbies
- ex: sports/non-sports -> this becomes a classification problem (you know what outcomes you get)
- if you don't know what groups, (don't know what kind of things I will find, but I want to classify students into 2,3,4 groups. You don't have historical data, so this question becomes a clustering problem.

What is high-correlation?

Format: give question -> you write out the problem and the plan. check what kind of problem, what are the criteria. No coding in the midterm exam
- not lower than 85%. many students get 90%
- if you don't pay attention in the class, the score will be low.

Effort in finding the data (part of what she evaluates the student)
Choose a method
Clean data properly
Able to analyze the data
Doesn't require that the model is super good, r-squared 90%, etc.
Just need to do a reasonable process
Ex: social science question -> 20-30% r^2 is fine for the prof. need to explain the reason why the model score is low, etc.

make powerpoint, easy to understand logic
- don't want very detailed
- around 20 pages
- has key message, don't need to explain the code
- checks logic
-

Course Schedule

#	Date	Topic	Lab
1	9-Jan	Data Science/AI/Machine Learning Overview	
2	16-Jan	Data Acquisition & Preprocessing	Lab_01
3	23-Jan	Linear Regression	Lab_02
4	30-Jan	Logistic Regression	Lab_03
5	6-Feb	Time-Series Analysis	Lab_04
6	13-Feb	Decision Tree	Lab_05
7	20-Feb	K-Means Clustering	Lab_06
8	27-Feb	Association Rule	Lab_07
9	**7-Mar**	**Midterm Week (13:00-16:00)**	
10	13-Mar	ANN	Lab_08
11	20-Mar	SVM	Lab_09
12	27-Mar	Ensemble Method	Lab_10
13	3-Apr	Image Mining	Lab_11
14	10-Apr	Text Mining	Lab_12
15	17-Apr	(Holiday)	
16	24-Apr	Distributed Processing	Lab_13
17	1-May	Final Project Presentation	
18	**8-May ~**	**Final Exam Week (No Exam)**	