San José State University
College of Science / Department of Computer Science Introduction to Machine Learning, CS171-04, Fall 2022

Course and Contact Information

Instructor: Samuel Chen, Ph.D.

Office Location: TBD

Email: sam.chen@susj.edu

Office Hours: Tuesday & Thursday 9pm - 10pm (appointment only)

Class Days/Time: Tuesday & Thursday 7:30 pm -8:45 pm

Classroom: Online

Prerequisites: CS 146 Data Structures and Algorithms

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on Canvas Leaning Management System course login website at http://sjsu.instructure.com. You are responsible for regularly checking with the messaging system through MySJSU at http://my.sjsu.edu (or other communication system as indicated by the instructor) to learn of any updates.

Course Description

General: Introduction to industry commonly used machine learning algorithms such as linear regression, logistics regression, AB testing, decision tree, random forest, ridge regression, lasso regularization, K-Mean clustering, nearest neighborhood clustering, and time series analysis for classification, clustering and prediction as well as the

complete lifecycle of industry model development and validation processes. Prerequisite: CS 146 (with a grade of "C-" or better); or instructor consent.

Course Objectives

- To introduce students how Machine Learning algorithms was used in industry.
- To teach students about data preparation, data cleansing, feature engineering, and how to handle missing data
- To teach students about model fitting metrics such as loss function/ residuals/error terms
- Introduce model fitting mechanism such as maximum likelihood and gradient decent
- To teach students about model performance measurement method such as back test and cross validation
- To teach students about classification algorithms e.g., logistic regression, decision tree, random forest
- To teach students about unsupervised algorithms e.g., k-mean clustering, KNN
- To teach students about predictive modeling algorithms e.g., linear regression, generalized linear model, and time series analysis
- To teach students about Regularization ,Shrinkage Methods

- Enhance students Python programming skills and familiarity of hands on packages, e.,g pandas, numpy, sikit-learn
- To teach students about A/B Testing, p values

Course Learning Outcomes (CLO)

Upon successful completion of this course, students should be able to:

- Apply theoretical knowledge and practical skills to develop classification/predictive/clustering models
- Proficiently use the Python Jupyter Notebook to complete required tasks in model development
- Have sense of model over fitting and under fitting
- Equipped knowledge to optimize model performance
- Select appropriate machine learning algorithms to answer business questions
- Have hands on skills on machine learning algorithms
- Apply theoretical knowledge and practical skills to validate classification/predictive/clustering models

Recommended Texts/Readings

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow ,Concepts, Tools, and Techniques to Build Intelligent Systems (2nd Edition) by Aurélien Géron

(Full text available at SJSU Library)

Hands-on Time Series Analysis with Python From Basics to Bleeding Edge Techniques (1st edition) by BV Vishwas and Ashish Patel

(Online access available at SJSU Library)

Optional Texts/Readings

Essential Python (1st edition) by Sridevi Pudipeddi and Ravi Chityala

These are the BSCS Program Outcomes supported by this course:

- An ability to apply knowledge of computing and mathematics to solve problems
- An ability to analyze a problem, and identify and define the computing requirements appropriate to its solution
- An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs
- An ability to use current techniques, skills, and tools necessary for computing practice

Course Requirements and Assignments

Success in this course is based on the expectation that students will spend, for each unit of credit, a minimum of 45 hours over the length of the course (normally three hours per unit per week) for instruction, preparation/studying, or course related activities, and so on.

Homework Assignments

You are expected to learn all of the material presented in the lectures.

Written homework and project reports are also a requirement of the course. Homework and project reports must be turned in on time; **late homework and reports will NOT be accepted.** Both homework and project assignments are due at the beginning of the class period on the announced due date.

In Class Exercises, Pop Quizzes/Questions and Discussion Forum for Participation Points

Unannounced in class exercises and pop questions may be given anytime during class. The purpose of in class exercises and pop questions is to encourage you to learn, study and review the concepts and materials presented/discussed in the lecture. These will generally be problems covered in the today's or previous lecture. Another way to earn participation points is to ask or answer other students question, or share you opinion in the discussion forum in Canvas.

Midterm and Final Exams

Exams will consist of questions and problems aimed at assessing student mastery of course topics. Conceptual questions may be in the form of essay or multiple-choice format. Python code problems will require to type Python commands or select the right answer from multiple choice. Calculation problems will require you to use calculator to solve the problems of statistics or metrics that covered in class. ***You can bring 1 letter size cheat sheet for MidTerm andFinal.*** All exams are closed book and note.If you are unable to attend any one of the exams, arrangements may be made only if you have a legitimate reason. You need to inform your instructor ahead of time and have written documentation available. If you are unable to attend the exam due to illness or emergency, you also need to inform your instructor **before the exam** and bring documentation afterwards to request a make-up exam, or the points for that exam will be allocated to other exams.

Model Development and Validation Project

Students can either form a team with 2 or 3 members or individually develop classification/predictive/clustering/knn models. Apply the learned knowledge starting pull raw data, data cleansing, provide descriptive statistics of variables, perform feature engineering, perform bi-variate analysis (correlation analysis), model fitting, back test or cross validation, provide model fitting metrics, and visualize the output. Each team should submit the final project including 2 documents:

Model White Paper

White paper should provide the business background (business questions being answered), data source, data field description, machine learning algorithm being used, the setup of cross validation or back test, model performance, and how can the manager make decision based on this data driven analysis.

Python Code Runbook

The python code runbook should be well structured by sections and clearly describe what each section of code is doing. Good comments of code is highly suggested. The code should be able to be run through without syntax error, logic error, and not counter business sense.

By submitting/presenting a project, team members attest that they all participated in the conceptualization and accomplishment of the project. It is incumbent on team members to assure that **each team member MUST contribute in writing program code and documents**, no one on the team "free rides" through the project. If problems arise during the term, upon consultation with team members, the instructor will remove non-participating team members from their teams. Individuals removed from teams will not receive points on the team assignment.

Available Software and useful Links

- Anaconda individual edition with Python 3.8, at https://www.anaconda.com/products/individual
- Python Tutorial: https://docs.python.org/3/tutorial/index.html
- Google Corelab: https://colab.research.google.com/notebooks/welcome.ipynb

Source of Sample Data

- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/
- Kaggle : https://www.kaggle.com/datasets

Grading Information

Determination of Grades

The components of the final grade will be distributed as follows:

- **Class Participation: 80 points** (in class exercises, pop quiz, discussion forum, etc.)
- **HW Assignments: 320 points** (4 Individual HWs)
- **Midterm exam: 200 points**
- **Model Development and Validation term project: 200 points** (Team with peer evaluations)
- **Final exam: 200 points** (Comprehensive)

Digit number grades will be assigned according to the following policy:

960 ~1000 ---- A+
930 ~ 959 ---- A
900 ~ 929 ---- A-
860 ~ 899 ---- B+
830 ~ 859 ---- B
800 ~ 829 ---- B-
760 ~ 799 ---- C+
730 ~ 759 ---- C
700 ~ 729 ---- C-
660 ~ 699 ---- D+
630 ~ 659 ---- D
600 ~ 629 ---- D-
0 ~599 ----F

Each assignment, project, and exam will be scored (given points) but not assigned a letter grade.Final individual class letter grades will be assigned based on the class curve. Your final class grade can be adjusted up or down depending on your level and quality of class / project performance.

Online Class Protocol and Other Notes

- **Absences in attending anyone of the first two lectures will be instructor-dropped out from the class.**

- Students are required to have an electronic device (laptop, desktop or tablet) with a camera and built-in microphone. Students are responsible for ensuring that they have access to reliable Wi-Fi during tests. If students are unable to have reliable Wi-Fi, they must inform the instructor, as soon as possible.

- All pop quizzes and exams will be proctored in this course through Respondus Monitor (with eye- tracking) and LockDown Browser. A webcam during exams is required. Please note it is the instructor's discretion to determine the method of proctoring. If cheating is suspected the proctored videos may be used for further inspection and may become part of the student's disciplinary record. Note that the proctoring software does not determine whether academic misconduct occurred, but does determine whether something irregular occurred that may require further investigation. Students are encouraged to contact the instructor if unexpected interruptions (from a parent or roommate, for example) occur during an exam.

- There will be no Zoom lecture recordings for later review/study. Recording a lecture is prohibited. Students are prohibited from recording class activities (including class lectures, office hours, advising sessions, etc.), distributing class recordings, or posting class recordings. Materials created by the instructor for the course (syllabi, lectures and lecture notes, presentations, etc.) are copyrighted by the instructor. This university policy (S12-7) is in place to protect the privacy of students in the course, as well as to maintain academic integrity through reducing the instances of cheating. Students who record, distribute, or post these materials will be referred to the Student Conduct and Ethical Development office. Unauthorized recording may violate university and state law. It is the responsibility of students that require special accommodations or assistive technology due to a disability to notify the instructor.

- You will be called in most class sessions for pop questions and to discuss material contained in lectures by using Random Roster Checker.

- **Plagiarism/Cheating will not be tolerable: 'F' will be given to your FINAL COURSE GRADE and will be reported to the Department and the University. (please be noted: obtaining HW solutions from someone or giving/showing your HW solutions to someone is also treated as plagiarism/cheating.)**

- **Attendance is crucial to doing well on pop quizzes, assignments and examinations.**

- **Students are responsible for all materials distributed on Canvas and discussed in the class.**
  Attendance: University policy F69-24 at http://www.sjsu.edu/senate/docs/F69-24.pdf states that students should attend all meetings of their classes, not only because they are responsible for material discussed therein, but because active participation is frequently essential to insure maximum benefit for all members of the class. Consent for Recording of Class and Public Sharing of Instructor Material: University Policy S12-7, http://www.sjsu.edu/senate/docs/S12-7.pdf, requires students to obtain instructor's permission to record the course: Common courtesy and professional behavior dictate that you notify someone when you are recording him/her. You **must** obtain the instructor's permission to make audio or video recordings in this class. Such permission allows the recordings to be used for your private, study purposes only. The recordings are the intellectual property of the instructor; you have not been given any rights to reproduce or distribute the material. Course material cannot be shared publicly without his/her approval. **You are not allowed to publicly share or upload instructor generated material for this course such as exam questions, lecture notes, or homework solutions without instructor consent.**

MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found at **Canvas** of SJSU One. **You are responsible for regularly checking with the email system and Canvas through One.SJSU at http://one.sjsu.edu to learn of any updates.**

University Policies

Per University Policy S16-9, university-wide policy information relevant to all courses, such as academic integrity, accommodations, etc. will be available on Office of Graduate and Undergraduate Programs' Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/" Make sure to review these policies and resources.

CS171_04-Introduction to Machine Learning, Fall 2022, Course Schedule

Tentative Course Schedule (This schedule is subject to change with fair notice.)

Week	Date	Topics, Readings, Assignments, Deadlines
1	8/23	Motivation, Orientation/Syllabus, Course Introduction, Prerequisites Check (Student's Information Due)
1	8/25	Introduction to Python Anaconda/Jupyter Notebook Comparison of AI, Machine Learning, Data Mining, and other Applied Mathematics Specialization (Statistics, Econometrics)
2	8/30	Python Programming Review , Data Structures,Project team assignment and formation, 3 students per team
2	9/1	Pandas, Data Wrangle, Summary Statistics, Data Visualization
3	9/6	Numpy
3	9/8	Data Preparation, Data Cleansing, Data Normoalization (HW1 Assignment)

4	9/13	Linear Regression :Basic Assumptions of Linear Regression,Bi-Varitae/Correlation Analysis
4	9/15	Linear Regression: Model Fitting Statistics, P- Value Lookup, R Square, Residual Diagnostics, ANOVA table, AB Test
5	9/20	Logistic Regression , Feature Engineering, Confusion Matrix (HW1 Due)
5	9/22	Confusion Matrix continued, Sigmoid Function, Odds, Odds Ratio (HW2 Assignment)
6	9/27	Cost(Loss) Function, Gradient Descent
6	9/29	Gradient Decent (Batch/Stochastic/Mini Batch), Learning Curve
7	10/4	Over Fit, Under Fit, Learning Curve demonstrated by Polynomial Regression,
7	10/6	Regularization ,Shrinkage Methods
8	10/11	Lasso , Ridge Regression and Elastic Net (HW2 Due)
8	10/13	Dimension Reduction Method (HW3 Assignment)
9	10/18	Mid-Term
9	10/20	Principal Component Analysis, PCA for Compression, Randomized PCA, Incremental PCA
10	10/25	Another classification algorithm — Decision Tree
10	10/27	Random Forest (Boosting and Bagging)
11	11/1	Ensemble Model, Boosting, Probability Theory: Law of Large Number, Central Limit Theorem
11	11/3	Univariate Time Series - Exponential Smoothing(HW3 Due)
12	11/8	Univariate Time Series - ARIMA
12	11/10	Time Series Regression - SARIMA w/ Exogenous Variable (HW4 Assignment)

13	11/15	Final Project Guideline and Discussion
13	11/17	Unsupervised Algorithm - Clustering, KMean
14	11/22	KNN (HW4 Due)
14	11/24	Thanksgiving
15	11/29	Final Project Presentation
15	12/1	Final Project Presentation
16	12/6	Final Project Presentation
16	12/8	Final Exam