San Jose State University
Department of Computer Science
CS 271, Topics in Machine Learning, Spring 2019

- **Course and Contact information**
  - **Instructor:** Mark Stamp
  - **Office Location:** MH 216
  - **Telephone:** 408-924-5094
  - **Email:** mark.stamp@sjsu.edu
  - **Office hours:** Tuesday & Thursday, noon - 1:15pm
  - **Class Days/Times:** Tuesday & Thursday, 10:30 - 11:45pm
  - **Classroom:** MH 233
  - **Prerequisites:** CS 149

- **Course Description**
  - Topics in machine learning. The following machine learning techniques and related topics are covered in detail: hidden Markov models (HMM), profile hidden Markov models (PHMM), principal component analysis (PCA), support vector machines (SVM), clustering, data analysis, backpropagation and selected topics in neural networks. Illustrative applications of each of these major topics are provided, with most of the applications drawn from the field of information security. In addition, the course will include an overview of each of the following topics: k-nearest neighbor, boosting/AdaBoost, random forests, linear discriminant analysis (LDA), naive Bayes, with additional topics as time permits. Prerequisite: CS 149.

- **Learning Outcomes**
  - The focus of this course will be machine learning, with illustrative applications drawn primarily from the field of information security. After completing this course students should have a working knowledge of a wide variety of machine learning topics, and have a good understanding of how to apply such techniques to real-world problems.

- **Required Texts/Readings**
  - The primary text will be ***Machine Learning with Applications in Information Security*** (https://www.crcpress.com/Introduction-to-Machine-Learning-with-Applications-in-Information-Security/Stamp/p/book/9781138626782), by Mark Stamp, published by Chapman Hall/CRC in 2017. This book covers several machine learning techniques in detail, and includes a large number of illustrative applications. Many of the applications are from information security, including a variety of topics related to malware, intrusion detection (IDS), spam, and cryptanalysis, among others.

  - Additional relevant material:
    - PowerPoint slides at http://www.cs.sjsu.edu/~stamp/ML/powerpoint
    - Current semester lecture videos are available at http://www.cs.sjsu.edu/~stamp/ML/lectures/CS271_Spr19/. If you are asked to login to access the videos, both the username and password are "infosec". **Note**: The instructor hereby gives students permission to record his lectures (audio and/or video). At least with respect to this class, your instructor has nothing to hide.
    - Class-related discussion will be posted on Piazza at https://piazza.com/class/jr6y6fuchq819a. You are strongly encouraged to participate by asking questions, as well as by responding to questions that other students ask. At the start of the semester, you should receive an email asking you to join this discussion group—if not, contact your instructor via email.

  - The applications parts of this course are essentially self-contained, but for additional background information on the security-related topics, the following resources are recommended.

- *Computer Viruses and Malware*, John Aycock, Springer 2006. Many of the applications we discuss are related to malware. Aycock's book is easy to read and in spite of being fairly old, it provides a good foundation for malware research.
- *Information Security: Principles and Practice*, Mark Stamp, Wiley 2011. If you have not taken CS 265, you should do so. You can refer to this fine book if you have questions about security-related topics during this course.
- Open Malware (at http://www.offensivecomputing.net/) includes a large collection of samples of live malware.
- VX Heavens (at http://vx.netlux.org/) is a source for "hacker" type of information on viruses. Malware samples are also available.
- Journal of Computer Virology and Hacking Techniques (at http://www.springer.com/computer/journal/11416) is a journal for malware-specific research papers. There are also several good conferences that focus on malware and/or machine learning applications in information security.
- Recent masters project reports (at http://www.cs.sjsu.edu/~stamp/cv/mss.html#masters). Most of these projects involve applications of machine learning to malware or other topics in information security.

- **Course Requirements and Assignments**
  - SJSU classes are designed such that in order to be successful, it is expected that students will spend a minimum of forty-five hours for each unit of credit (normally three hours per unit per week), including preparing for class, participating in course activities, completing assignments, and so on. More details about student workload can be found in University Policy S12-3 at http://www.sjsu.edu/senate/docs/S12-3.pdf.

  - Schedule
    - Week 1 --- Introduction and overview
    - Week 2 --- Hidden Markov Models
    - Week 3 --- Data Analysis
    - Week 4 --- Applications of Hidden Markov Models
    - Week 5 --- Profile Hidden Markov Models
    - Week 6 --- Applications of Profile Hidden Markov Models
    - Week 7 --- Principal Component Analysis
    - Week 8 --- Applications of Principal Component Analysis
    - Week 9 --- Support Vector Machines
    - Week 10 --- Applications of Support Vector Machines
    - Week 11 --- Clustering
    - Week 12 --- Clustering Applications
    - Week 13 --- k-Nearest Neighbor, Neural Networks, Boosting/AdaBoost, Random Forests
    - Week 14 --- Linear Discriminant Analysis, Naive Bayes, Regression Analysis, Conditional Random Fields
    - Week 15 --- Project presentations

  - Homework is due *typewritten* (include source code, but not executable files) by class starting time on the due date. Each assigned problem requires a solution and an explanation and work detailing how you arrived at your solution. Cite any outside sources used to solve a problem. When grading an assignment, I may ask for additional information. Note that a *subset* of the assigned problems will typically be graded.

    Homework must be submitted via email before the start of class on the due date. Be sure to have an extra copy of your homework with you in class, and be prepared to discuss your solutions. Your written solutions must be in a pdf file. Submit any source code or other attachments in separate files (i.e., no code in the solution itself). You must provide enough discussion of your solution so that the grader can understand your solution, and so that the grader can be sure that you understand your solution. Put your written solution and any relevant source code in a folder named "yourlastname". Then zip your homework folder and submit the file yourlastname.zip via email to cs271.spring19@gmail.com. The subject line of your email ***must*** be of the form:

```
CS271HMK assignmentnumber yourlastname last4digitofyourstudentnumber
```

The subject line must consist of the four identifiers listed. There is no space within an identifier and each identifier is separated by a space.

- Assignment 0: Due Tuesday, August 28
  For this assignment, turn in a hardcopy of your solutions at the start of class.
  1. Read A Revealing Introduction to Hidden Markov Models (at https://www.cs.sjsu.edu/~stamp/RUA/HMM.pdf) and do the following.
     a. Briefly (1 paragraph) summarize how an HMM is trained.
     b. How is a trained HMM used to score a sequence?
     c. Very briefly explain how an HMM and dynamic program differ.
     d. Why is it necessary to scale the values of the matrices when training an HMM?
  2. Read the article "Models will rule the world" (I will give you a hardcopy of the article on the first day of class) and do the following.
     a. In one paragraph, summarize the authors' main points.
     b. Write second paragraph discussing what you most agree with and anything that you disagree with in this article.

- Assignment 1: Due Tuesday February 5
  Chapter 2, problems 1, 2, 3, 10. For problem 10 you must use HMM code that you have written entirely on your own, following the algorithms given in your textbook.

- Assignment 2: Due Tuesday February 12
  Chapter 2, problems 11, 14, and 15 parts a thru c. For problems 11 and 14, use 27 symbols (lower-case A thru Z and word-space) instead of 26 as specified in these problems. For problem 15, there are only 26 plaintext symbols (lower-case A thru Z), so you will need to use N=26. You will receive 10 points extra credit if you use your own HMM code to solve at least 2 of the 3 problems in this assignment. You can also earn 10 points extra credit if you solve problem 15, part d. If your code is too slow and/or buggy, you may use this reference HMM implementation (https://www.cs.sjsu.edu/~stamp/RUA/HMM_ref.zip).

- Assignment 3: Due TBD
  Chapter 3, problems TBD

- Assignment 4: Due TBD
  Chapter 4, problems TBD

- Assignment 5: Due TBD
  Chapter 5, problems TBD

- Assignment 6: Due TBD
  Chapter 6, problems TBD

- Assignment 7: Due TBD
  Chapter 7, problems TBD

- Assignment 8: Due TBD
  Problems TBD

- Assignment 9: Due TBD

        Problems TBD

        - Assignment 10: Due TBD
          Problems TBD

    - NOTE that 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. Attendance per se shall not be used as a criterion for grading."

- **Grading Policy**
    - Test 1, 100 points. Date: TBD.
    - Homework, quizzes, class participation and other work as assigned, 100 points. A subset of the assigned problems will be graded.
    - Machine Learning Project, 100 points. You must obtain approval for your project proposal from me (via email) prior to the start of class on Thursday, February 21, and you must be prepared to give a brief presentation of your proposed topic on that day. A written project report is due Tuesday, April 30 and project presentations will begin on that day (or shortly thereafter).
    - Final, 100 points. Date: Thursday, May 16 from 9:45am to noon. The official finals schedule is here: http://info.sjsu.edu/static/catalog/final-exam-schedule-spring.html
    - Semester grade will be computed as a weighted average of the major scores listed above.
    - ***No*** make-up tests or quizzes will be given and ***no*** late homework or project (or other work) will be accepted.
    - Grading Scale:

Percentage	**Grade**
92 and above	A
90 - 91	A-
88 - 89	B+
82 - 87	B
80 - 81	B-
78 - 79	C+
72 - 77	C
70 - 71	C-
68 - 69	D+
62 - 67	D
60 - 61	D-
59 and below	F

    - Note that "All students have the right, within a reasonable time, to know their academic scores, to review their grade-dependent work, and to be provided with explanations for the determination of their course grades." See University Policy F13-1 at http://www.sjsu.edu/senate/docs/F13-1.pdf for more details.

- **Guest Lectures**
    - TBD
        - Date: TBD
        - Time: TBD
        - Location: TBD

        - Topic: TBD
        - Abstract: TBD

    - TBD
        - Date: TBD
        - Time: TBD
        - Location: TBD
        - Topic: TBD
        - Abstract: TBD

- **Classroom Protocol**
    - Keys to success: Do the homework, complete a good project, and attend class

    - **Wireless laptop is *required***. Your laptop must remain closed (preferably in your backpack and, in any case, not on your desk) until I inform you that it is needed for a particular activity

    - **Cheating** will not be tolerated, but working together is encouraged

    - Student must be respectful of the instructor and other students. For example,
        - No disruptive or annoying talking
        - Turn off cell phones
        - Class begins on time
        - Class is not over until I say it's over

    - Valid picture ID required at all times

    - The last day to drop without a "W" grade is Tuesday, February 5, and the last day to add is Tuesday, February 12

- **University Policies**
    - Office of Graduate and Undergraduate Programs maintains university-wide policy information relevant to all courses, such as academic integrity, accommodations, etc. You may find all syllabus related University Policies and resources information listed on GUP’s Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/