San José State University
Department of Mathematics & Statistics
Math 251 Statistical and Machine Learning Classification
Fall 2018

Course and Contact Information

Instructor:	Guangliang Chen
Office Location:	MQH 417
Telephone:	(408) 924-5131
Email:	guangliang.chen@sjsu.edu
Office Hours:	MW 10:45-11:45am, TR 9:50-10:20am, and by appointment
Class Days/Time:	MW 9-10:15am
Classroom:	Clark Hall 111 (Incubator Classroom)
Prerequisites:	Math 32, Math 129A, Math 164, Math 267A or instructor consent. Math 267A may be taken concurrently.

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on my faculty web page at http://www.sjsu.edu/faculty/guangliang.chen/ and 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

Dimensionality reduction, instance-based classification, discriminant analysis, logistic regression, support vector machine, kernel methods, ensemble learning, neural networks and deep learning, classification of nonnumeric data. 3 units.

Course Goals

- Introduce the machine learning field of classification and its applications
- Present the main ideas and necessary theory of major classification methods in the literature
- Teach how to use specialized software to perform classification tasks while adequately addressing the practical challenges (e.g., parameter tuning, memory and speed)
- Provide students with valuable first-hand experience in handling big, complex data

Course Learning Outcomes (CLO)

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

- Use the terminology associated with classification
- Determine the nature of the various classifiers (linear or nonlinear, distribution-based or optimization-based, etc.)

- State clearly the mathematical assumption and objective of each classifier
- Apply dimensionality reduction techniques (such as PCA and LDA) to preprocess the data
- Perform classification on data sets assisted with software
- Properly set the parameters associated to each classifier
- Assess the classification output through measures such as confusion matrix, training/testing error

Required Texts/Readings

Required Textbooks

- James, Witten, Hastie and Tibshirani (2015), "An Introduction to Statistical Learning with Applications in R", 6th edition, Springer. ISBN 978-1-4614-7137-0, ISBN 978-1-4614-7138-7 (eBook), DOI 10.1007/978-1-4614-7138-7. Freely available online at http://www-bcf.usc.edu/~gareth/ISL/
- Nielson (2015), "Neural Networks and Deep Learning", Determination Press. Freely available online at http://neuralnetworksanddeeplearning.com/

Recommended Readings

- Hastie, Tibshirani, and Friedman (2009), "*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*", 2nd edition, Springer-Verlag. Freely available online at http://statweb.stanford.edu/~tibs/ElemStatLearn/index.html

Other resources such as tutorial papers and slides will be provided from time to time in class to assist in learning the material.

Other technology requirements / equipment / material

The course will make intensive use of specialized software such as MATLAB, R and Python to perform various computing tasks in the setting of big data. Therefore, familiarity with at least one of the programming languages is required.

Students taking this course will be asked to use the following data for practice:

- MNIST Handwritten Digits (available at http://yann.lecun.com/exdb/mnist/), which consists of 70,000 digital images of size 28x28 of handwritten digits 0…9 collected from about 250 people
- USPS Zip Code Data (available at http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html), which consists of 9,300 size 16x16 grayscale images of handwritten digits scanned from envelops
- 20 Newsgroups Data Set (available at http://qwone.com/~jason/20Newsgroups/), consisting of about 19,000 text documents that are divided into 20 groups (according to their topics)

Smaller data sets such as those from the UCI Machine Learning Repository (at http://archive.ics.uci.edu/ml/) will also be used for teaching demonstration and homework assignments.

Course Requirements and Assignments

Course requirements include class attendance, participation in discussions, weekly homework and reading assignments, and a midterm exam.

You are expected to attend all classes and actively participate in classroom discussions which often lead to a deeper understanding of the concepts and are also strongly associated with course grade.

The homework assignments will typically involve both theory and programming, which are either to implement by yourself a classifier learned in class or to learn how to use an existing function/package. Detailed instructions about homework will be provided in class.

The students may collaborate on homework but must write independent codes and solutions. Copying and other forms of cheating will not be tolerated and will result in a zero score for the homework (minimal penalty) or a failing grade for the course, possibly combined with other disciplinary actions from the university.

Final Examination or Evaluation

The course will end with a final project to be selected between each individual student and the instructor. The students will need to give a 10-minute oral presentation to report their findings and meanwhile write a report of at least 5 pages. Both the presentation and report will be graded based on clarity, depth, completeness, and originality. More details will be given near the end of the semester.

Grading Information

You must submit homework on time to receive full credit. No make-up exam will be given if you miss the midterm exam (unless you have a legitimate excuse such as illness or other personal emergencies and can provide documented evidence).

You must show all your work for both homework and test. Note that it is your work (in terms of correctness, completeness, and clarity), not just your answer, that is graded. Thus, correct answers with no or poorly written supporting steps may receive very little credit.

The weights in determining the semester average are:

- Homework (weekly): 40%
- Midterm (Oct 15): 30%
- Final project (Dec 10 and 12): 30% (10% oral, 20% report)

I expect to use the following cutoffs for assigning your course grade (I reserve the right to slightly adjust these percentages in order to better reflect the actual distribution of the class in the end):

- A+: 96%, A: 93-96%, A-: 90-93%
- B+: 85-90%, B: 80-85%, B-: 75-80%
- C+: 72-75%, C: 68-72%, C-: 65-68%
- D: 55-65%
- F: <55%

Classroom Protocol

- The class starts on time, so do not be late.
- If you miss a class, you are responsible for finding out what's said/done in that class (such as new announcement, deadline change, etc.) and responding accordingly.
- Please make sure to turn off or mute your cell phone during class.
- Please do not perform irrelevant or distracting activities in class.
- Academic dishonesty at any level is not tolerated and will be surely reported to the Office of Student Conduct (per SJSU policy).

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/"

Tentative Course Schedule

This schedule is subject to change with fair notice which will be made in class if there is a delay in progress.

Week	Date	Topics	Notes
1	Aug 22	Review of linear algebra	
2	27	Course introduction	
2	29	Principal Component Analysis I	
3	Sep 5	Principal Component Analysis II	
4	10	kNN classification I	
4	12	kNN classification II	
5	17	Bayes classifiers I	
5	19	Bayes classifiers II	
6	24	Discriminant analysis I	
6	26	Discriminant analysis II	
7	Oct 1	Logistic regression I	
7	3	Logistic regression II	
8	8	Multiclass extensions	
8	10	(buffer)	
9	15	Midterm	
9	17	Support vector machine I	
10	22	Support vector machine II	
10	24	Support vector machine III	
11	29	Support vector machine IV	
11	31	Classification trees	
12	Nov 5	Ensemble learning I	
12	7	Ensemble learning II	
13	14	Ensemble learning III	
14	19	Neural networks I	
15	26	Neural networks II	
15	28	Neural networks III	
16	Dec 3	Introduction to deep learning	
16	5	Last class	
17	10	Project presentations	
17	12	Project presentations, cont’d	