Math 285 – Classification with Handwritten Digits

Time and location: TTh 1:30-2:45pm, Clark Hall 111 (Incubator Classroom)

Instructor: Dr. Guangliang Chen

Contact: Office: MH 417 Phone: 4-5131 Email: *guangliang.chen@sjsu.edu*

Webpage: *http://www.math.sjsu.edu/~gchen/Math285S16.html*

Office hours: 2:45-3:30pm TTH, 2-3 W, and by appointment

Course description: Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set whose category membership is known. An example would be labeling a given email as "spam" or "non-spam", using examples from each class. Classification is one of the most important research fields in machine learning, with still lots of ongoing research nowadays. In this course we will survey different kinds of classification algorithms in the literature, including many from this century. Below is a list of topics to be covered in this course:

- ***Dimensionality reduction techniques****: principal component analysis*
- ***Instance-based classifiers****: kmeans, k nearest neighbor (kNN) classifiers, and their variants*
- ***Distribution-based classifiers****: Bayes classifiers*
- ***Linear classifiers****: linear discriminant analysis, logistic regression, and support vector machines*
- ***Kernel-based nonlinear classifiers****: linear classifiers combined with kernels*
- ***Ensemble methods****: trees, bagging, random forests, and boosting*
- ***Perceptron and neural networks***
- ***Classifiers that can handle categorical data***

Because this is a new course, the instructor reserves the right to change/add/remove topics at any time during the semester.

Data sets: The MNIST handwritten digits, available at *http://yann.lecun.com/exdb/mnist/*, is a benchmark data set used by many researchers for testing and comparing their algorithms. It consists of 60,000 images of size 28x28 of handwritten digits 0…9 for training and 10,000 for testing; see the figure at the beginning of the syllabus for some examples in the training set.

The MNIST dataset is very easy to understand and use; its size and dimension are both quite large; it is also difficult enough in terms of classification. Therefore, throughout the semester you will be asked to implement and try various classifiers on this data set, report your results, and compare different classifiers in order to fully understand their performance.

Additional data used in this course includes

- Small toy data sets created by the instructor (mainly for demonstrating ideas and testing algorithms)
- Real data in other databases (for further testing of various algorithms), such as
    - Other handwritten digits, such as USPS Zip Code Data (http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html), and
    - UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html).

You are also welcome to suggest new data that is appropriate for use in this course.

Prerequisites: Math 164 and Math 129A. The course introduces cutting-edge machine learning research that is based on advanced linear algebra and statistics, so a grade of B or better in each course is required.

Programming language: Familiarity with at least one of MATLAB / R / Python is required. The course will have an extensive computing component, and students are expected to implement many of the ideas using a language they are familiar with and test them on various data sets.

Required textbook: None, but we will cover material from various sources (websites, papers, textbook chapters, instructor's notes, etc.) and reading material will be provided from time to time in class.

Below are some optional textbooks you may use as references:

- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, by Hastie, Tibshirani, and Friedman, Springer. Freely available at *http://statweb.stanford.edu/~tibs/ElemStatLearn/*
- *Introduction to Data Mining*, by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar 1st ed, Addison Wesley, 2006, ISBN: 978-0321321367. Available at *http://www-users.cs.umn.edu/%7Ekumar/dmbook/index.php*

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

- Understand and appreciate the classification problem and its various applications
- Learn many of the existing classifiers and get familiar with their implementation
- Apply the algorithms and software to real data
- Gain valuable first-hand experience in big data that is desired by many companies

Requirements and grading: Course requirements will include around 6 homework assignments and two projects: midterm and final (you will be divided into groups of sizes 2 or 3 to work on the projects).

The homework will typically involve extensive programming, either to implement by yourself a classifier learned in class or to learn how to use an existing function/package, with the ultimate goal of testing the newly learned classifier on the MNIST handwritten digits. Note that

- You may collaborate on homework but you 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 by other disciplinary actions from the university.
- You must prepare your homework in presentation format using PowerPoint or LaTex that contains all necessary details such as software used, parameter values, accuracy, and running time. You must also include the original scripts to support your results.
- You must submit homework on time in order to receive full credit. Late homework submitted within a day will still be accepted but will be discounted for only half of the points (regardless of the reason).

There is a distinct midterm project after each major classifier (or class of classifiers) is taught, to be completed by a distinct group of students. In this project you and your group will need to carefully summarize the idea and steps of the algorithm(s), describe how it is applied to the MNIST digits, and report the best possible results with that kind of classifier. You will need to make a poster to be displayed on the wall of the classroom.

The final project will be selected between your group and the instructor. It can be either about a new method, or about a novel application of some classifier learned in class. More detail will be given later but you will have at least one month to work on the project. In the end, you will need to either make a poster presentation or give a 15-minute oral presentation in class to report your findings. Regardless which format you choose, your presentation will be graded based on *clarity*, *depth*, *completeness*, and *originality*.

The weights in determining the semester average are:

- **Homework**: 50%
- **Midterm project**: 20%
- **Final project**: 30%

Letter grades will be computed from the semester average. Maximum cutoffs for A, B, C and D grades are 90%, 80%, 70%, and 60%, respectively. These bounds may be moved lower at the instructor's discretion.

On the weekly basis, you are expected to spend at least 6 hours outside class to read the assigned material and do the homework.

Special accommodations: If you anticipate needing any special accommodation during the semester (for example you have a disability registered with SJSU's Accessible Education Center), please let me know as soon as possible.

Instructor feedback: This is an experimental course in data science, being taught at SJSU for the first time. Your (early) feedback is encouraged and greatly appreciated, and it will be seriously considered by the instructor for improving the course experience of you and your classmates. Please submit your anonymous feedback through this google page: *http://goo.gl/forms/f0wUD5aZSK*.