Math 285 Classification with Handwritten Digits – First day of class, Spring 2016

Dr. Guangliang Chen

Dept. of Math and Statistics
San José State University

How did this course originate?

- Developed from last fall's Math 203 CAMCOS (whose theme was classification)
- Based on a current Kaggle competition: *Digit Recognizer*
  - Kaggle is a Silicon Valley start-up and Kaggle.com is its online platform hosting many data science competitions.
  - It uses a *crowdsourcing* approach that relies on the fact that "there are countless strategies that can be applied to any predictive modelling task and it is impossible to know at the outset which technique or analyst will be most effective".

How do Kaggle competitions operate?

- Companies, with the help of Kaggle, post their data as well as a description of the problem on the website;
- Participants (from all over the world) experiment with different techniques and submit their best results to a scoreboard to compete;
- After the deadline passes, the winning team receives a cash reward (which could be as much as several millions) and the company obtains "a worldwide, perpetual, irrevocable and royalty-free license".

Potential benefits of participating in a Kaggle competition

- Experience with large, complex, interesting, real data
- Learning (new knowledge and skills)
- Become a part of the data science community
- Cash prize
- Could get you a job

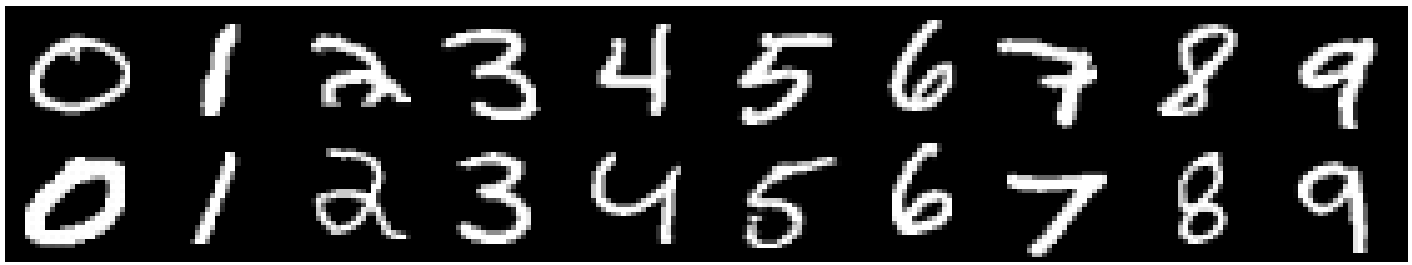
What is classification?

Classificiation is a major *machine learning* field that studies how to assign labels to new data (**test set**) based on labeled data (**training set**).

- Lots of applications, e.g., spam email detection, digit recognition, face recognition, and document classification
- Classification is an example of *supervised learning*; in contrast, clustering is unsupervised (focus of last semester's 285)

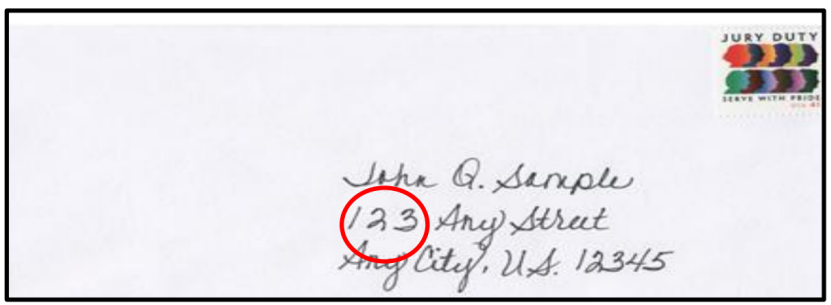
The digit recognition problem

Given a set of training examples, determine what digits the test images contain by machine:



Potential Applications

- **Banking**: Check deposits
- **Surveillance**: license plates
- **Shipping**: Envelopes/Packages



Our main data set

The MNIST database of handwritten digits, formed by Yann LeCun of NYU, has a total of 70,000 examples from approximately 250 writers:

- The images are 28×28 in size
- The training set contains 60,000 images while the test set has 10,000
- It is a benchmark dataset used by many people

Why MNIST?

- Famous
- Simple to understand and use
- Yet difficult enough for classification
  - Big data (large size and high dimensionality)
  - 10 classes in total $(0, 1, \ldots, 9)$
  - Great variability (due to different ways people write)
  - Nonlinear separation

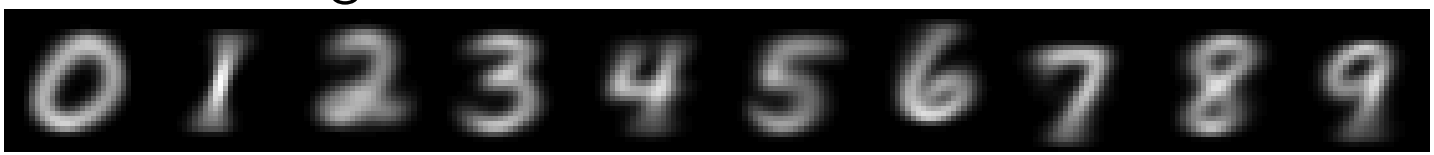
- Well studied (thus lots of resources available)
  - The Kaggle competition page (`https://www.kaggle.com/c/digit-recognizer`)
  - Lecun's page (http://yann.lecun.com/exdb/mnist/)
  - CAMCOS course page from last semester (`http://www.math.sjsu.edu/~gchen/math203.html`)

Representation of the digits

- The original format is matrices (of size 28×28)
- They can also be turned into vectors (784 dimensional)

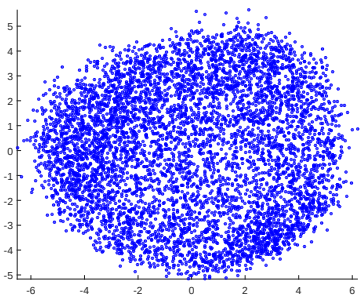
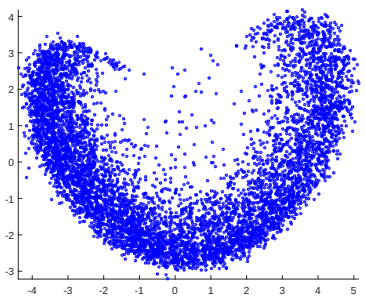
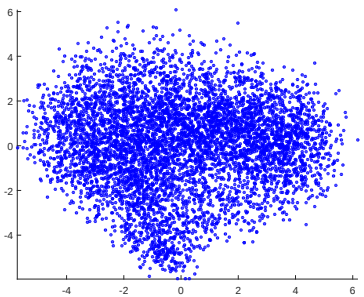
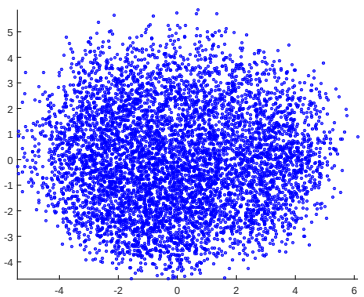
Visualization of the data set

1. The "average" writer



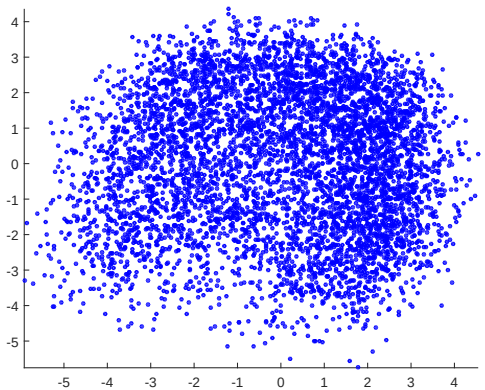
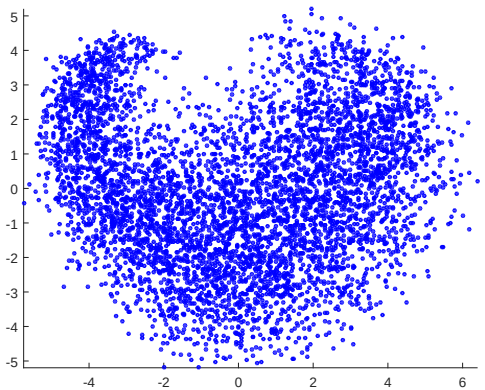
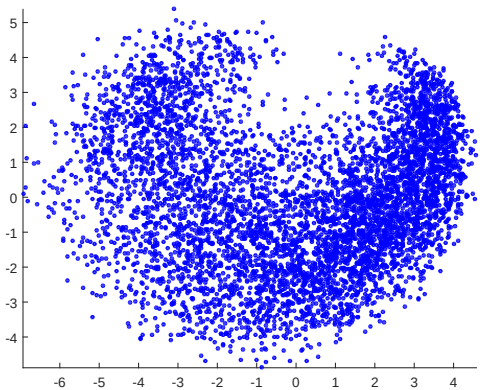
2. The full appearance of each digit class

0 - 3

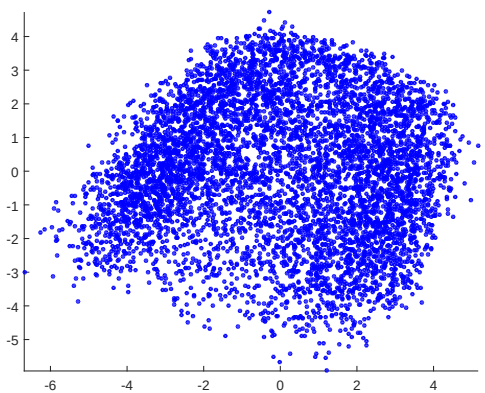
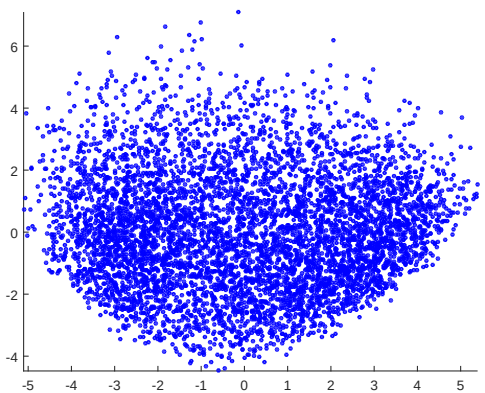
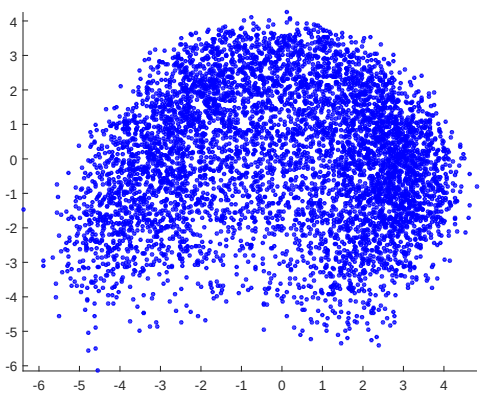





(cont'd on next page)

4-6

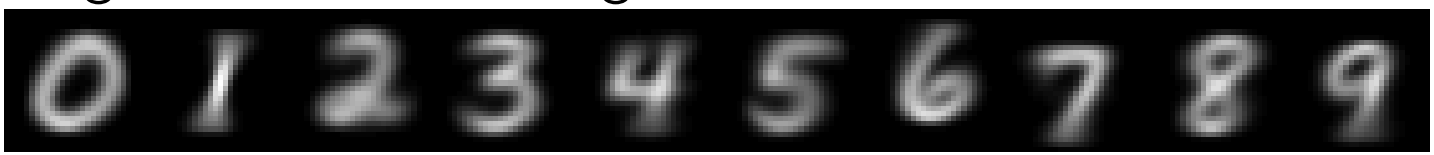




7-9

A very first attempt at classification

Assign labels to test images based on the most similar centers.

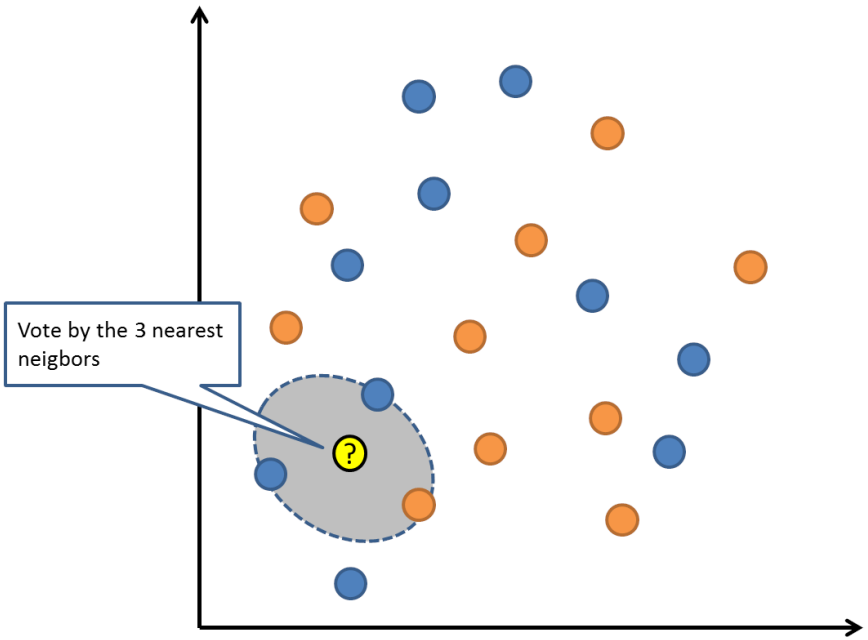


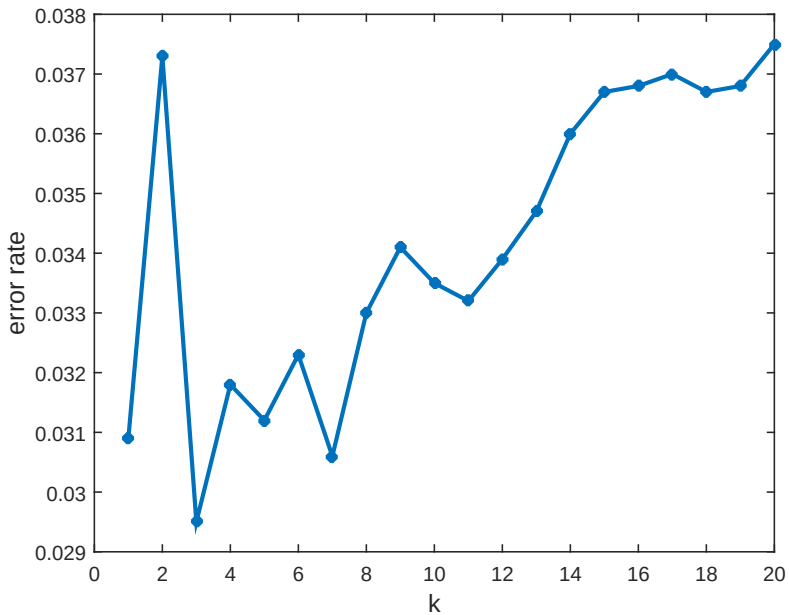
Call this classifier (global) k*means*.

How well does it do: 18.0% error rate.

A second attempt

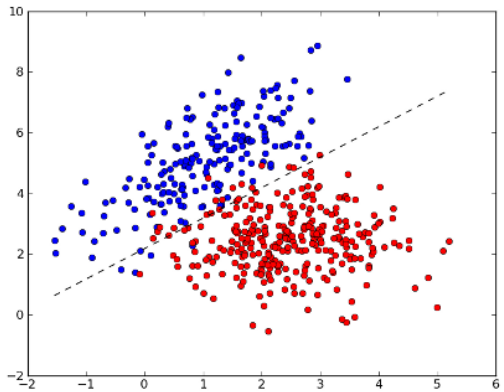
The k nearest neighbors (kNN) classifier assigns class label based on the k closest examples around a test point





Which other classifiers will we cover?

- Linear classifiers, such as Logistic regression, LDA, and SVM



- Nonlinear classifiers
- Tree classifiers
- Neural networks

Additional data to be used

- Small toy data sets created by the instructor
- Real data in other databases, such as
  - USPS Handwritten digits (`http://statweb.stanford.edu/~tibs/ElemStatLearn/data.html`),
  - UC Irvine Machine Learning Repository (`http://archive.ics.uci.edu/ml/datasets.html`).

Necessary technical background to take this course

- Strong linear algebra knowledge and skills (129A)
- Know probability and statistics well (163 and 164)
- Excellent programming skills (in at least one of Matlab, R, Python)

Characterwise expectation

- Hard work
- Independent thinking
- Team player
- Active learner
  - Eager to explore new things
  - Willing to go beyond requirements

Requirements of this course

- About 6 homework assignments (50%)
- A midterm project (20%)
- A final project (30%)

Homework policy

- You may collaborate on homework but you must write independent codes and solutions.
- You must prepare your homework solutions in presentation format using PowerPoint or LaTex.
- You must submit homework on time in order to receive full credit.

Is there a required textbook?

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

Below is an excellent book that you may use as reference:

- *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/`

The MATLAB License issue

Unfortunately, MATLAB could not be installed on the computers in this classroom. Here are some alternative options:

- A freely available 30-day trial version of the newest MATLAB with all toolboxes at `mathworks.com`
- The computer lab in MacQuarrie Hall 221 (limited access)
- Consider buying MATLAB student license (platform+all toolboxes, \$99.99/year).
- Talk to me if you still need help.

Some final reminders

- This is an experimental course (subject to change as needed)
- This is a demanding/challenging course
- This is also a highly rewarding course
- This is not a traditional course (essentially projects based)
- This is not an ordinary classroom!

Assignments

- Install/check software (MATLAB, R, Python).
- Download the MNIST data set from Lecun's page (and use the scripts that I provide to process them into .mat format)
- Perform very preliminary experiments (e.g., you need to be able to load the data). Try to understand the data as much as you can.
- Explore Kaggle.com and the competition pages

Questions?