Classification with Handwritten Digits

—Midterm Project Posters Session

Dr. Guangliang Chen

Dept. of Math & Statistics

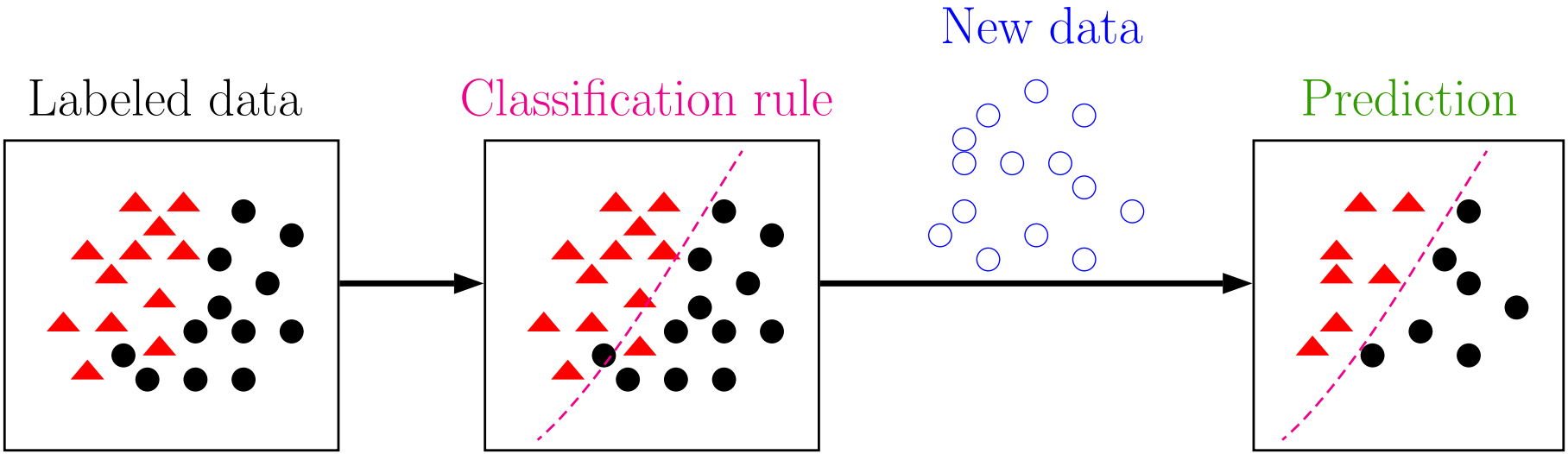
May 10, 2016

Rational of this course

- Learn a subject (**classification**)
- ...through an application (**digits recognition**)
- ...with a benchmark dataset (**MNIST Handwritten Digits**)
- ...using one technical computing language (**MATLAB / Python**)

What is classification?

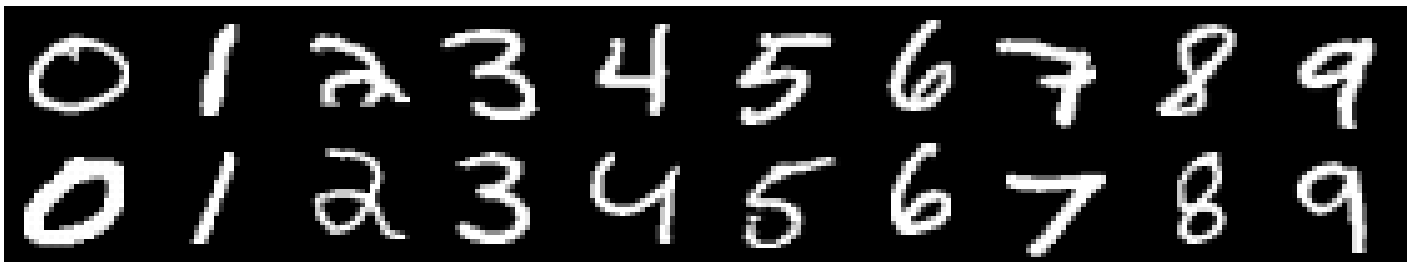
Classification is a *machine learning* problem about how to assign labels to new data based on a given set of labeled data.



Handwritten digits recognition

We teach classification mainly based on the the digit recognition problem:

Given a set of training examples



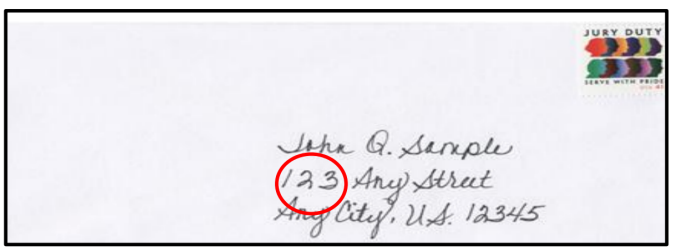
determine what digits the test images contain by machine:

Why digit recognition?

Simple, intuitive to understand, yet practically important

Potential Applications

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



Our main data set: MNIST handwritten digits

It is a benchmark data set in machine learning, consisting of 70,000 handwritting examples collected from approximately 250 writers:

- The images are black/white and 28×28 in size
- The data set is divided into two parts: 60,000 for training and 10,000 for testing

Why MNIST?

- Easy to use, yet difficult enough for classification
  - Big data (large size, high dimensionality, 10 classes)
  - Great variability (due to different ways people write)
  - Nonlinear separation between the classes
- Well studied (lots of learning resources available):
  - It is used by an ongoing Kaggle competition
  - Math 203 CAMCOS last fall at SJSU

Classifiers covered in this course

- Dimensionality reduction: PCA, Fisher's discriminant analysis, 2DLDA
- Instanced-based classifiers: kNN, $k$means
- Maximum a posteriori classification: LDA/QDA, Naive Bayes
- Logistic regression
- Support vector machine
- Ensemble methods: trees, bagging, random forest, and boosting
- Neural networks

List of posters today

- Instanced-based classifiers (by Yu Jung Yeh and Yi Xiao)
- Discriminant analysis (by Shiou-Shiou Deng and Guangjie He)
- Two dimensional LDA (by Xixi Lu and Terry Situ)
- Logistic regression (by Huong Huynh and Maria Nazari)
- Support vector machine (by Ryan Shiroma and Andrew Zastovnik)
- Ensemble methods (by Mansi Modi and Weiqian Hou)
- Two dimensional PCA (by Yijun Zhou)

Thursday: Final project posters

Thank you all for coming today!

If you wish to learn more,

- visit: `http://www.math.sjsu.edu/~gchen/Math285S16.html`, or
- email: `guangliang.chen@sjsu.edu`