CS189 Machine Learning Note

> *Date: Jan 18, 2023*
>
> *Author: Leo-Adventure*

Lecture1

Notice

Website: http://people.eecs.berkeley.edu/~jrs/189/

Professor: Jonathon Shewchuk

Discussion Section: Tue, and Wed begin next week

Homework1 Due: Jan 25, 11:59 pm

Prerequisite:

- vector calculus
- linear algebra
- probability
- plentiful programming experience

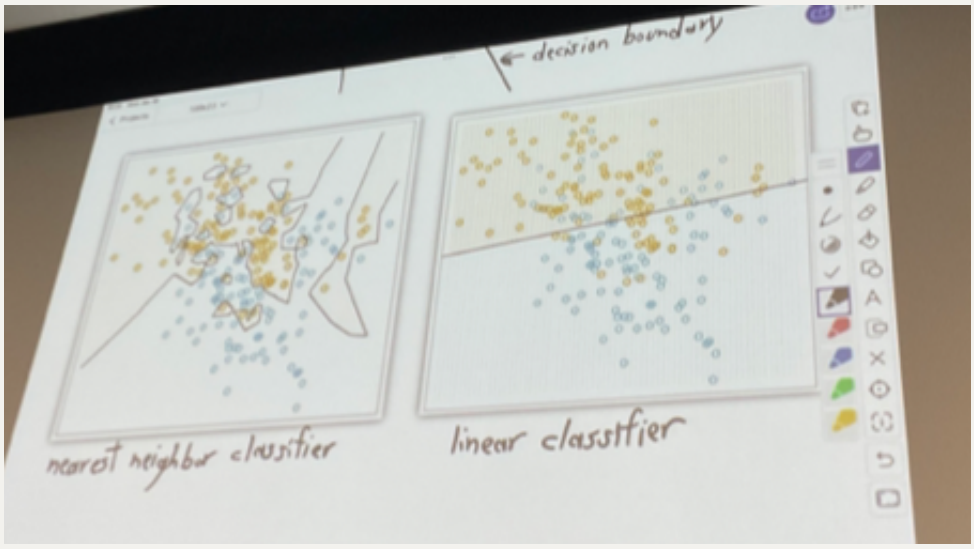
Grading:

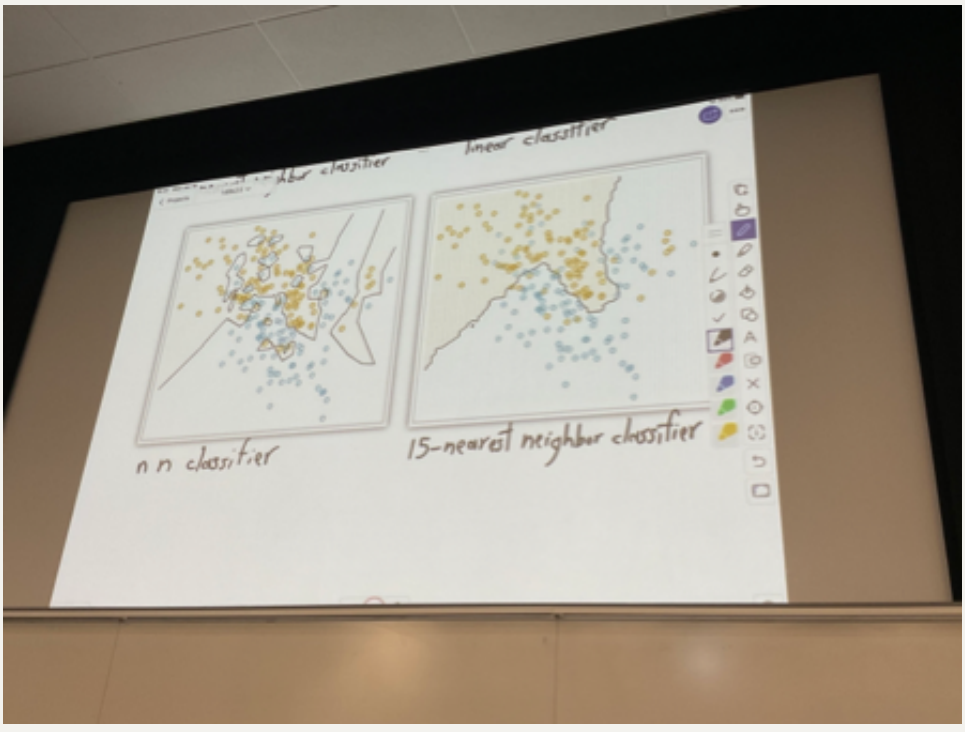
- 40% 7 homework; late policy: 5 slip days in total
- 20% midterm(Monday March 20, 7:00~8:30 pm Wheeler Auditorium)
- 40% Final exam(Friday May 12, 3:00~6:00pm)

Introduction and Definitions

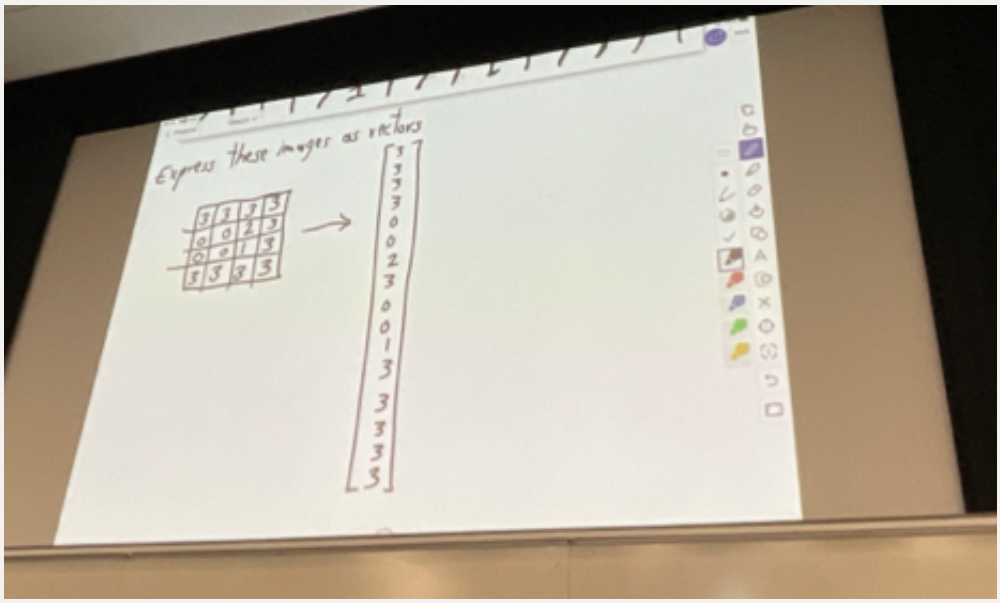
Core material

- Finding patterns in data and using them to make a prediction
- Models and statistics help us understand patterns
- Optimizations algorithms "learn" the pattern





Digits classification: to express the images as vectors



The linear decision boundary is a hyperplane

For testing and validation:

- Train a classifier
- Test the classifier

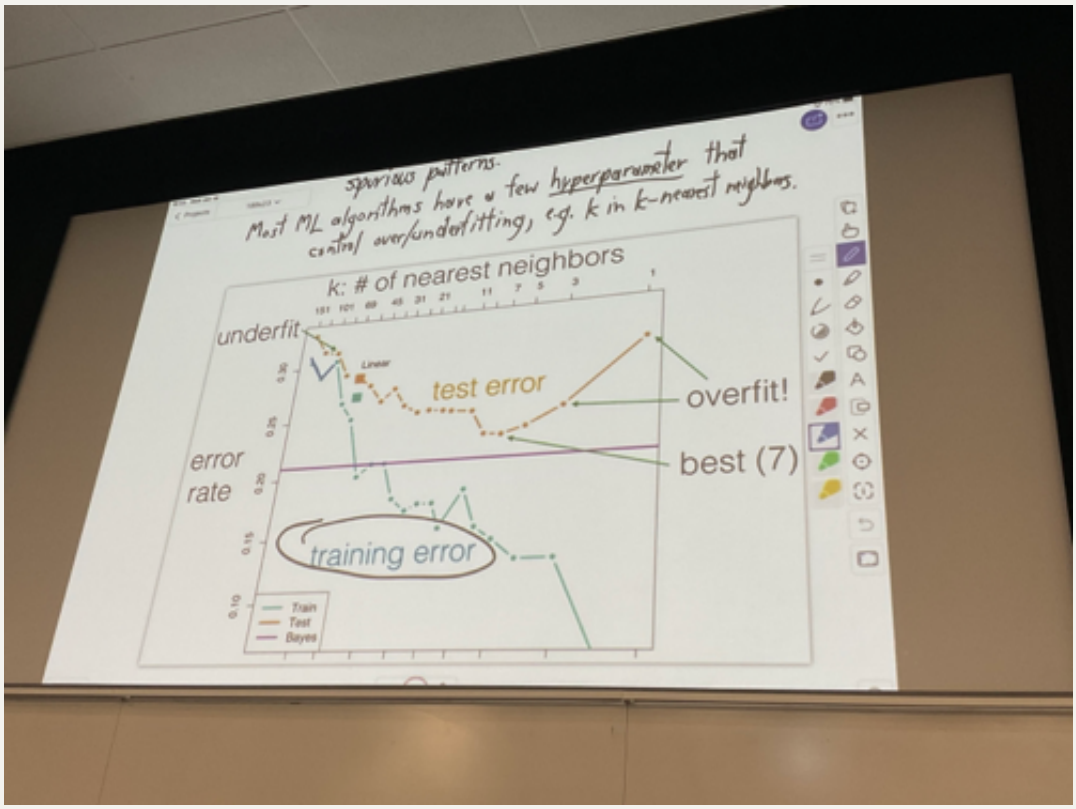
Two kinds of errors:

1. Training set error: fraction of training images not classified correctly
2. Test set error: fraction of misclassified new images that were not seen during training

Outliers: Points whose labels are atypical

Overfitting: When the test error deteriorates because the classifier becomes too sensitive to outliers or other spurious(fake) patterns.

Hyperparameters: Look for the optimal point



We select them by validation:

- Hold back a subset of the labeled data, called the validation set
- Train the classifier multiple times with different hyperparameters setting
- Choose the setting that works best on the validation set

Three sets:

- Training Set: Used to learn model weights
- Validation Set: Used to tune hyperparameters and choose among different models
- Test Set: Used to find an evaluation of the model and keep it in a vault. **Run once at the end**

Kaggle.com:

- Run ML competitions, including our homework
- We use two data sets.
-
    - "public" set labels available during the competition
    - "private" set labels known only to Kaggle