The Kaggle Competitions: An Introduction to CAMCOS Fall 2015

Guangliang Chen

Math/Stats Colloquium
San Jose State University

August 26, 2015

Outline

- Introduction to Kaggle
- Description of projects
- Summary

CAMCOS in Fall 2015: A glance

- Made possible by a proposal by Dr. Bremer (no outside sponsor this time)
- Theme of the program is **data science** (we all like data science)
  - Many online courses
  - Many universities start to offer a degree in this field
  - High demand, from both industry and academia, for graduates in data science is projected
- Projects of this CAMCOS are selected from the online competitions at *Kaggle.com*

Basic facts about Kaggle

- Kaggle is a Silicon Valley start-up and *Kaggle.com* is its online platform hosting many data science competitions.
- Founded by Anthony Goldbloom in 2010 in Melbourne, and moved to San Francisco in 2011.
- It uses a crowdsourcing approach which 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*.
- Hal Varian, Chief Economist at Google, described Kaggle as "a way to organize the brainpower of the world's most talented data scientists and make it accessible to organizations of every size".

How it works

- 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 world-wide, perpetual, irrevocable and royalty-free license".

Achievements and impact of Kaggle

- Kaggle claims 358,225 data scientists on its jobs board (picture on next slide)
- Customers include many big companies and organizations such as NASA, Merck, GE, Microsoft, Facebook, Allstate and Mayo Clinic
- It has advanced the state of the art in different fields, such as HIV research, traffic forecasting and mapping dark matter.
- It has lead to academic papers and continued interest to further innovate.

Data Science Jobs Board

Hiring?

Access 358225 data scientists

Kaggle is the world's largest community of data scientists, statisticians, and machine learning engineers. Kagglers demonstrate the skills to solve the toughest problems across many industries.

Create a Job Listing

Seeking?

Browse 1032 careers

The jobs board sources career openings for data professionals like you. Subscribe to be notified of new opportunities in data science, machine learning, statistics, and other analytics jobs.

Search our listings

Subscribe | Follow @KaggleCareers

1,032 topics, 1,277 posts | Views

Hotels.com - Data Scientist (London, UK)
16 hours ago | 231

Potential benefits of participating in the Kaggle competitions

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

Back to CAMCOS

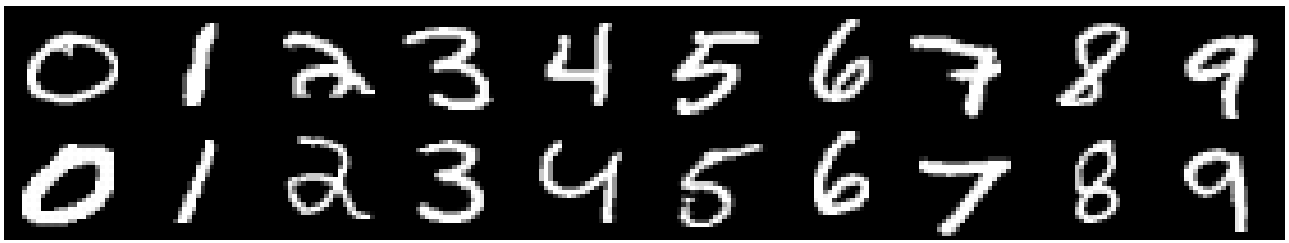
The projects of this CAMCOS are selected from Kaggle competitions:

- **Project 1: Digit recognizer**
  - Duration: July 25, 2012 – December 31, 2015
  - Award: knowledge (no cash prize)
- **Project 2: Springleaf marketing**
  - Duration: August 14 – October 19
  - Award: $100,000

Project 1: Digit recognition

Given an image of a handwritten single digit, determine what it is by machine:

- Training images must be given;



- Need to learn a rule (classifier) and apply it to new images

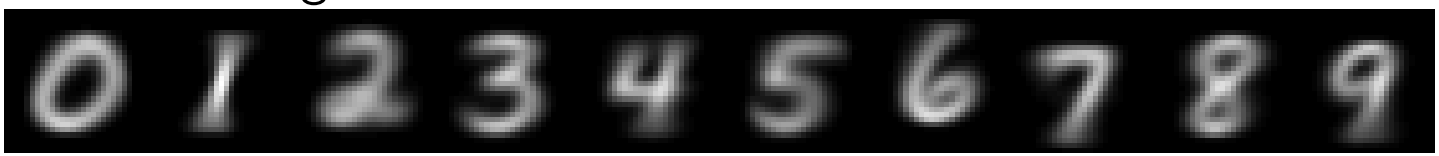
MNIST handwritten digits

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 to test their algorithms

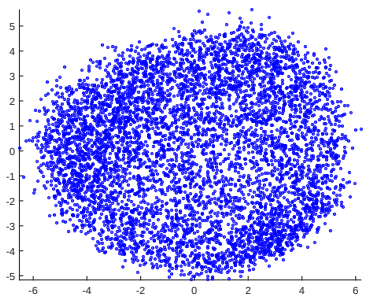
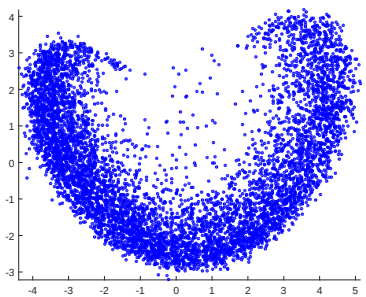
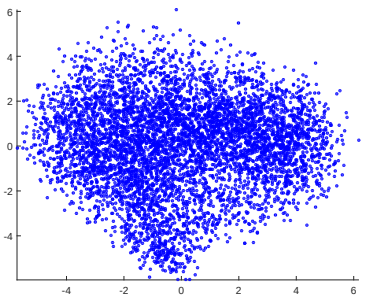
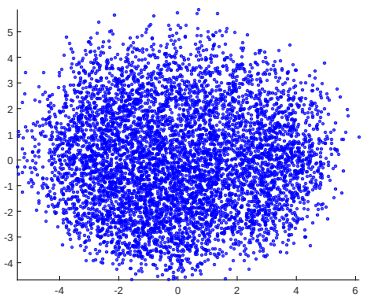
Visualization of the data set

1. The "average" writer



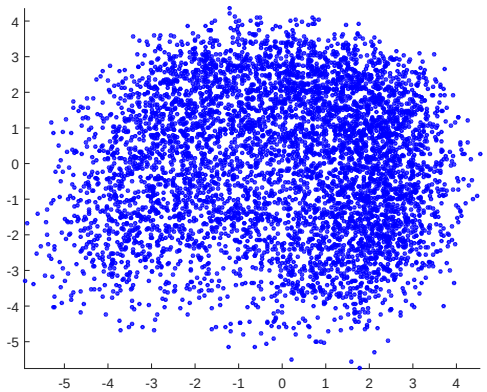
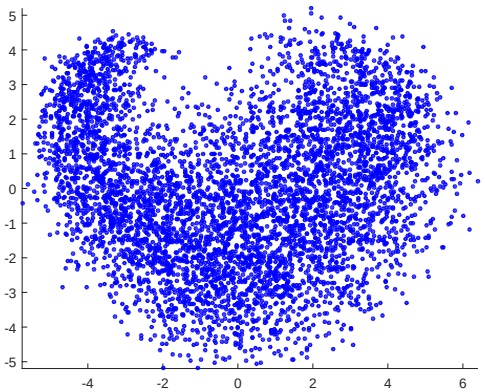
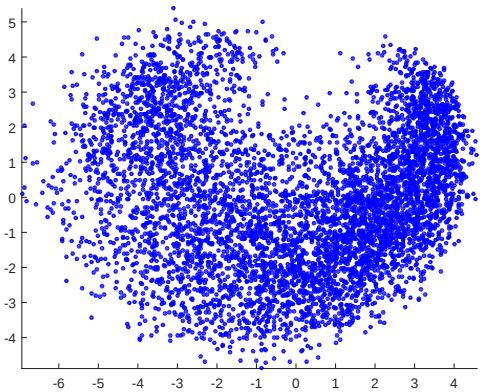
2. PCA plot of each digit cloud

0 - 3

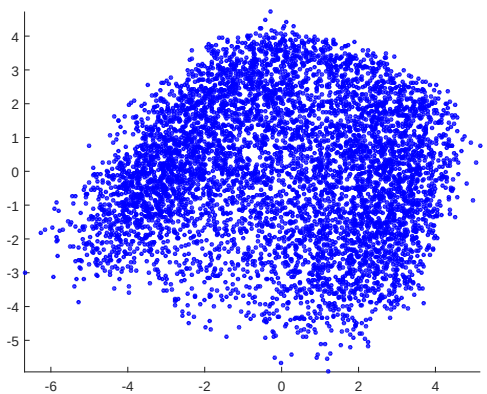
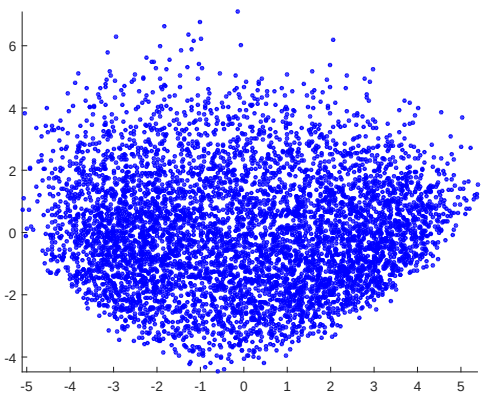
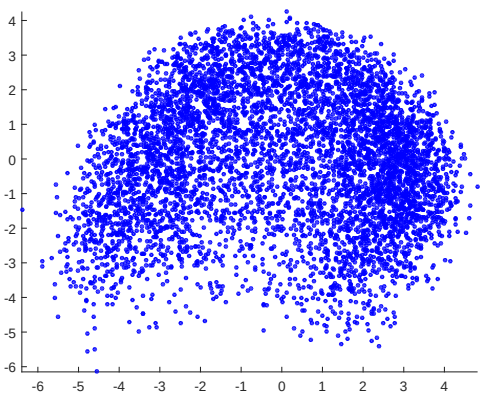





(cont'd on next page)

4-6

7-9

The general classification problem

Given data and their class labels $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{1, \ldots, J\},\ 1 \leq i \leq n$, find a function f (in some function space) by minimizing

$$\sum L(y_i, f(\mathbf{x}_i))$$

where L is a loss function (e.g., ℓ_1 or ℓ_2 distance)

- It is an instance of supervised learning.
- Statistically, this is a regression problem (with categorical outcomes), often done with logistic regression.
- Lots of applications: document classification, spam email detection, etc.

Some classifiers from the literature

- Nearest subset classifiers: $k$means
- Nearest neighbors classifiers: kNN
- Linear classifiers, such as
  - Logistic regression
  - Naive Bayes classifier
  - Linear discriminant analysis (LDA)
  - Support vector machine (SVM)
- Other: Decision trees, perceptron, neural networks, etc.

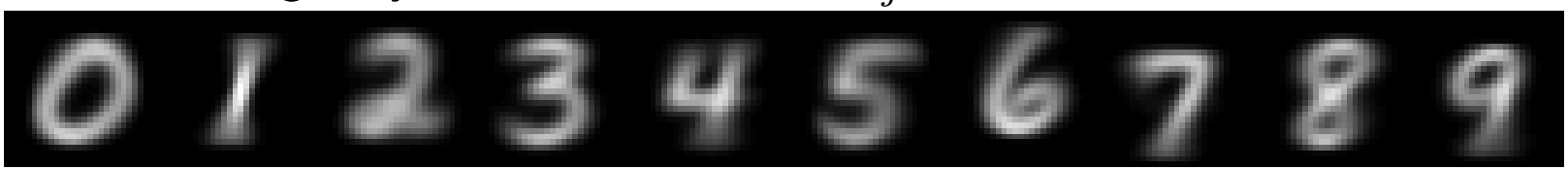
Nearest subset classifiers

The idea is to assign a new point to the "closest" class of training points:

$$\hat{j} = \text{argmin}_{1 \le j \le J} \, \text{dist}(\mathbf{x}, \mathcal{C}_j)$$

by using some kind of distance metric:

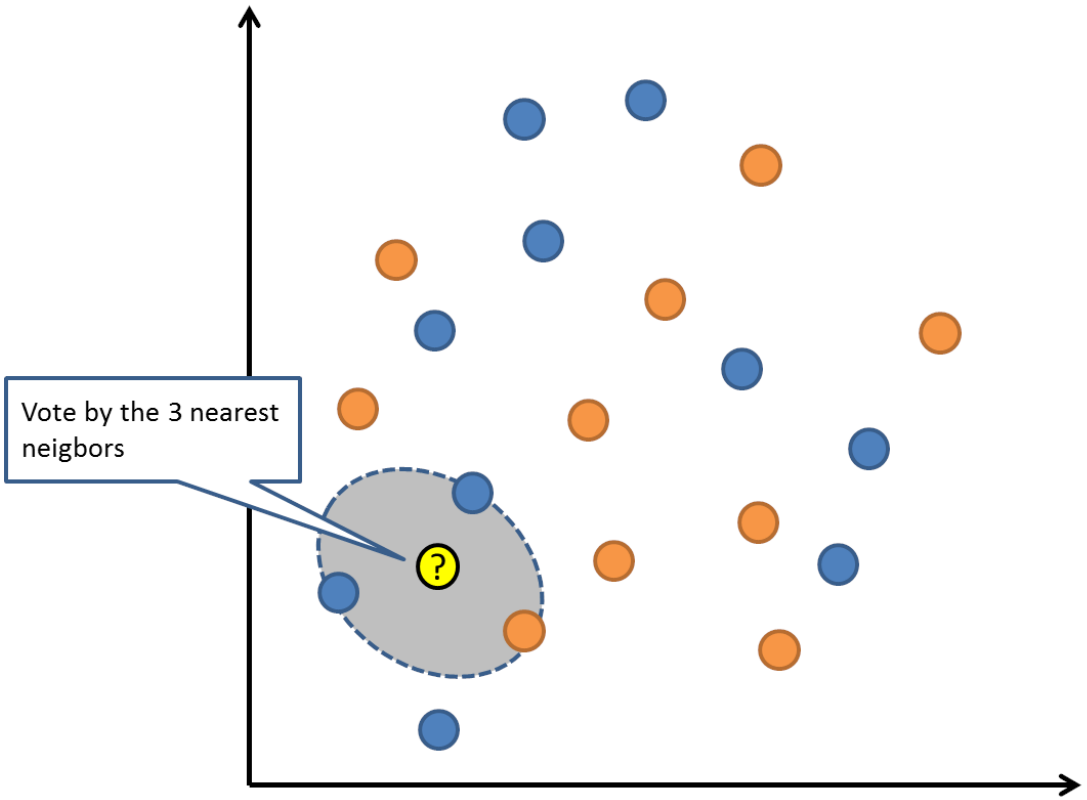
- $k$means: using only the center of each $\mathcal{C}_j$



- Local $k$means: using the center of the k closest points from each $\mathcal{C}_j$, where $k \in \mathbb{Z}^+$

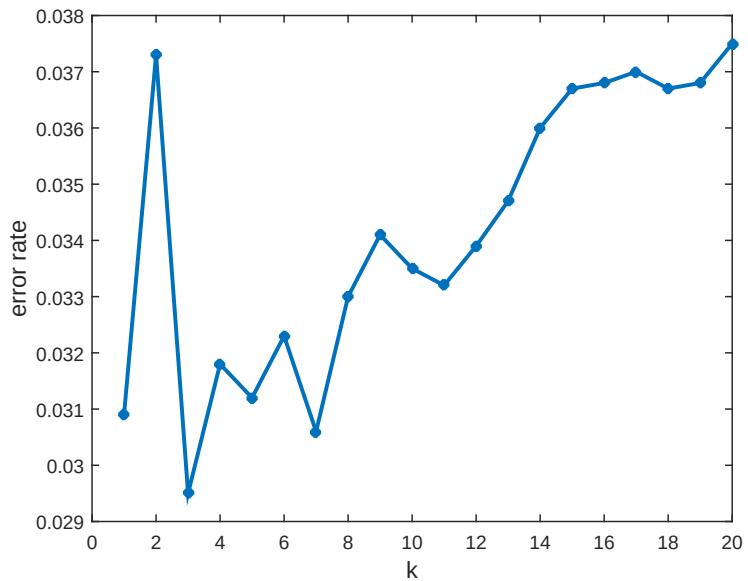
Nearest neighbors classifiers

kNN assigns class label based on the k closest points around a new point



Some quick experimental results

- The error rate of the global $k$means classifier is 18.0%.
- The error rate of the local $k$means classifier (for $k = 1$) is 3.1%.
- Error rate of the kNN classifier (for different k) is shown below:



Comments on the $k$means/kNN classifiers

- Instance-based learning (or lazy learning)
- Simple to implement
- Algorithmic complexity only depends nearest neighbors search
- The choice of k is important
- Cannot handle skewed distributions

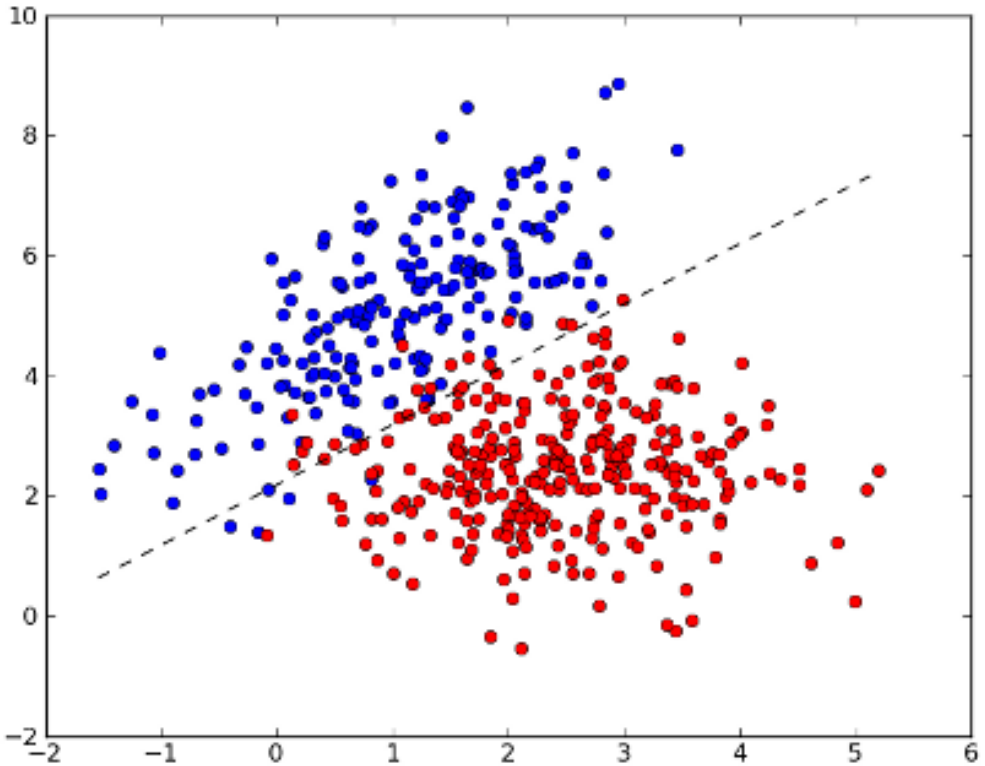
Linear classifiers

- For two classes, linear classifiers typically have the following form

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{w}^T\mathbf{x} - \mathbf{b} > 0; \\ 0, & \text{otherwise} \end{cases}$$

  where $\mathbf{w}, \mathbf{b}$ are learned from training samples.

- The above rule is equivalent to using a hyperplane as the classification decision boundary.



Building linear classifiers

There are two classes of methods for training $\mathbf{w}, \mathbf{b}$:

- Distribution-based (statistical methods): to model conditional density functions $P(\mathbf{x} \mid \mathcal{C}_j)$
  - Linear discriminant analysis (LDA): assuming Gaussian conditional distributions and performing a likelihood ratio test (when having only two categories)
  - Naive Bayes classifier: using Bayes rule $P(\mathcal{C}_j \mid \mathbf{x}) \propto P(\mathcal{C}_j)P(\mathbf{x} \mid \mathcal{C}_j)$ and selecting priors $P(\mathcal{C}_j)$

- Optimization-based (discriminative methods): to solve

$$\min_{\mathbf{w},\mathbf{b}} R(\mathbf{w}) + \gamma \sum L(y_i, 1_{\mathbf{w}^T\mathbf{x}_i - \mathbf{b}})$$

  where

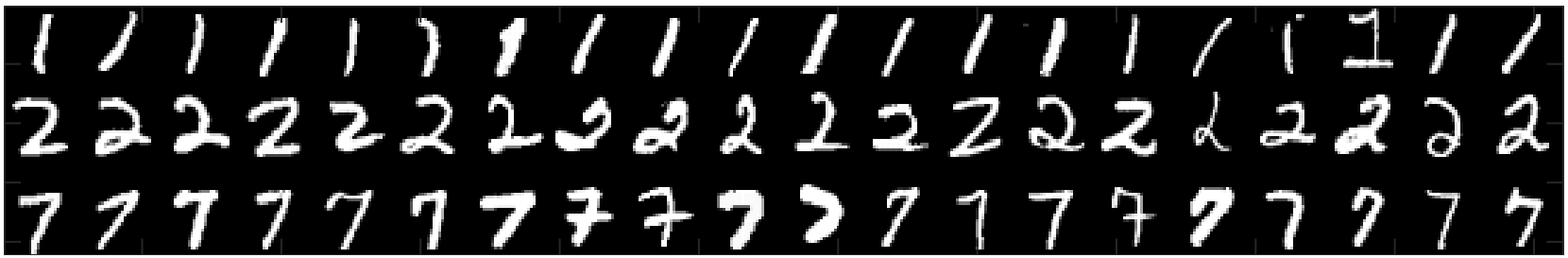
  - $R(\mathbf{w})$: regularization term
  - $L(y_i, 1_{\mathbf{w}^T\mathbf{x}_i - \mathbf{b}})$: loss of the prediction
  - γ: tradeoff constant

  Examples of this class include

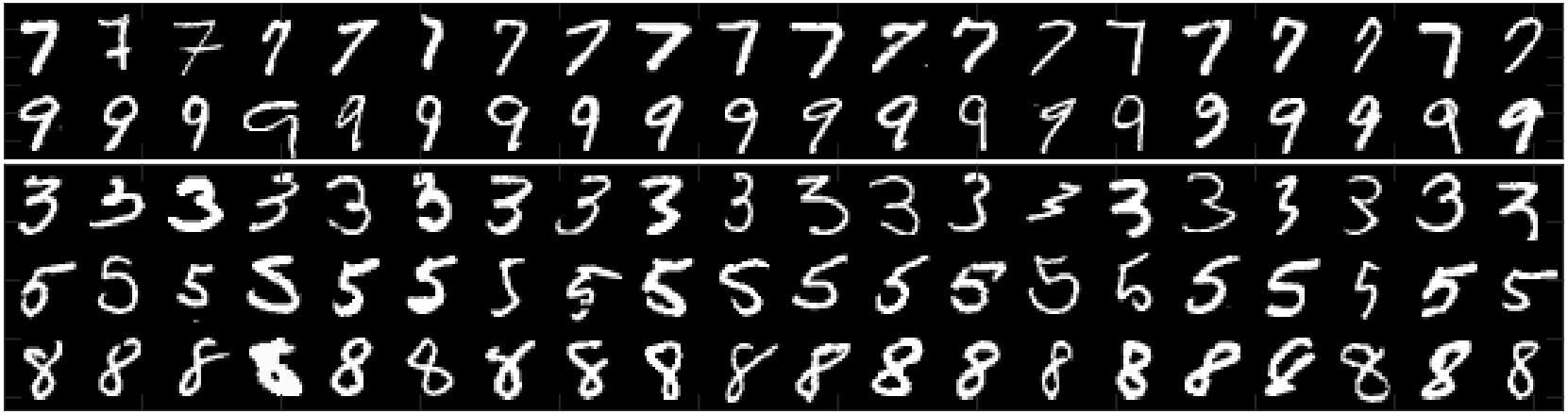
  - Support vector machine (SVM)
  - Perceptron

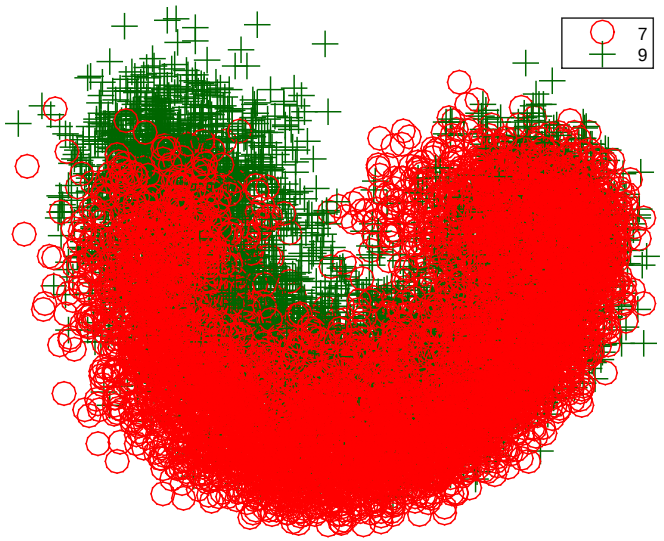
Challenges of the project

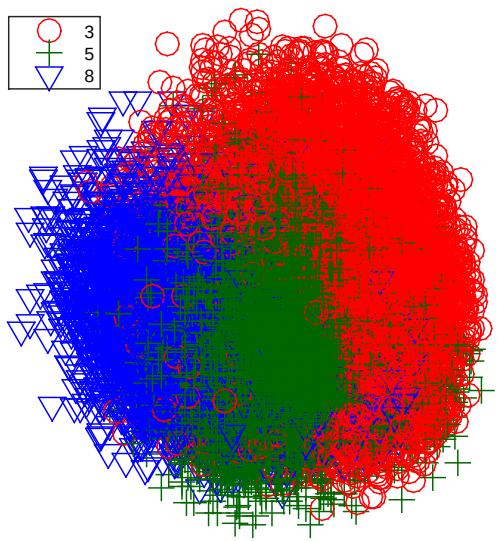
- Large amount of high dimensional data (60000×784)
- Great variability in the ways people write the digits (i.e., strong noise)



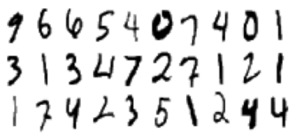
- Similar digits, e.g., $\{7, 9\}$ and $\{3, 5, 8\}$




7
9


3
5
8

- Lots of classifiers to try (and beat)



Knowledge • 651 teams

Digit Recognizer

Wed 25 Jul 2012 — Thu 31 Dec 2015 (4 months to go)

Dashboard ▼ | Public Leaderboard - Digit Recognizer

This leaderboard is calculated on approximately 25% of the test data.
The final results will be based on the other 75%, so the final standings may be different.

See someone using multiple accounts?
Let us know.

#	Δ1w	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	11245	1.00000	2	Wed, 08 Jul 2015 19:00:30 (-0h)
2	—	morpheushoc	1.00000	2	Mon, 13 Jul 2015 23:24:13
3	—	bell	1.00000	2	Fri, 31 Jul 2015 08:02:36
4	—	Bohdan Pavlyshenko	1.00000	1	Fri, 07 Aug 2015 09:57:04
5	—	gokumd	0.99986	4	Thu, 09 Jul 2015 19:08:26
6	—	zhxfl	0.99957	1	Wed, 29 Jul 2015 08:38:03
7	—	Fabio Capela	0.99914	17	Sat, 01 Aug 2015 19:01:06
8	↑22	shimtak	0.99700	4	Wed, 19 Aug 2015 13:20:25

Why you want to work on this project

- Data format is simple (easy to get started)
- Data set is well understood (as it has been extensively studied)
- The competition will provide tutorial to help you
- Lots of existing algorithms in the literature (good chance to learn)
- Can develop a solid background in classification

Project 2: Springleaf marketing

First, some background information:

- Springleaf is a company that operates in the financial services industry and does business in consumer lending
- Direct offers mailed to potential customers provide great value to the customers and it is an important marketing strategy used by Springleaf
- They want to improve their strategy to better target customers who truly need loans and seem to be good candidates
- They hosted this competition by providing training data and asking you to predict which customers will respond to a direct mail offer

Description of the data

Both the training and test data sets are 920 mb, in csv format:

- Each row corresponds to one customer (>145,000 customers only in the training set)
- The columns represent the anonymized customer information (a mix of continuous and categorical variables): "ID", "VAR_0001", "VAR_0002", ..., "VAR_1025"
- The response variable is binary and labeled "target"
- There are many missing values

Challenges of this project

- You need to be able to open/load the data files (enormous amount of complex business data)
- Need to deal with categorical variables
- Need to handle missing values
- Need to do feature selection (and get rid of lots of redundant information)
- Need to build a good classifier

- No need to win the competition (time is short and you are competing with 590+ teams)

$100,000 • 590 teams

Springleaf Marketing Response

Merger and 1st Submission Deadline

Fri 14 Aug 2015 — Mon 19 Oct 2015 (54 days to go)

Dashboard ▼ | Public Leaderboard - Springleaf Marketing Response

This leaderboard is calculated on approximately 30% of the test data.
The final results will be based on the other 70%, so the final standings may be different.

See someone using multiple accounts?
Let us know.

#	Δ2d	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Abhishek *	0.80254	18	Mon, 24 Aug 2015 13:58:17 (-3d)
2	↑4	Ilias Mir *	0.80133	29	Wed, 26 Aug 2015 01:27:11
3	↑2	YetiMan *	0.80130	33	Wed, 26 Aug 2015 00:21:20
4	↓2	Alexander Larko *	0.80130	21	Sun, 23 Aug 2015 13:49:14
5	new	clobber *	0.80128	1	Mon, 24 Aug 2015 07:00:40
6	↓2	Wenlong Shen	0.80112	21	Wed, 26 Aug 2015 02:05:34
7	↓4	Euclides Filho	0.80099	2	Thu, 20 Aug 2015 18:52:52
8	↑4	touzanis	0.80096	5	Mon, 24 Aug 2015 17:09:22

Why you want to work on this project

Because it is challenging!

Characteristics of an ideal candidate

- Good linear algebra knowledge (have taken 129A)
- Know probability and statistics well (have taken 163 and 164)
- Excellent programming skills (in Matlab, R, Python)
- Hard work
- Team player
- Eager to learn

Thank you for your attention!

- Introduction to the Kaggle competitions
- Description of course projects (both are about classification)
  - Digit recognition
  - Springleaf marketing
- Thanks to Woodward Foundation for support
- Contact: *guangliang.chen@sjsu.edu*

Questions?