

Foundations of Data Analytics (MSBD 5001, Fall 2018)

Kai Chen & Cecia Chan
CSE, HKUST

Outline

- Course logistics and pre-requisites
- MSBD 5001: learn about data science
- Course contents and tentative schedule
- Grading scheme.

Logistics

- Course webpage: on canvas
 - <https://canvas.ust.hk/> and go to MSBD 5001
 - A platform to post notifications and lecture materials
- Instructors:
 - Prof. Kai Chen, CSE, HKUST
 - Dr. Cecia Chan, CSE, HKUST
- TA:
 - Jiacheng Xia, CSE, HKUST

Logistics (cont'd)

- Lectures: Wed 7:30 – 10:20 pm, CYT G010
- Time divided into 3 50-minute sessions
 - Session 1 and 2: lecture contents, theory part
 - Session 3: Programming tutorial on course relevant topics
 - Please bring your **charged** laptops for tutorials
 - No requirements on your OS (Windows, Mac, Linux OK)
- 10 minute break between sessions

Prerequisites

- Basic programming
 - Knowledge of data structures, algorithm, object-oriented programming
- Multivariable calculus
 - Derivatives, gradients, tangent planes
- Linear algebra
 - Basic operations on vectors, matrices
 - Eigenvectors, matrix rank
- Probability & statistics

Course Goal

- Help students to learn the basic techniques in doing data analytics.
 - "What do I need to learn to be a data scientist?"
- Consider the course if you
 - Are familiar with most of pre-requisites
 - Want to learn about what data scientists do
- Introductory level, need to learn more by yourself

Requirements for Data Scientist

- Pre-process the data
- Build good models
- Focus on questions, not methods
- Know how to interpret the results

Preprocess & analyze the Data

- Pre-processing the data
 - Data cleansing, data integration
- Build good models
 - Models should describe data accurately
 - Can use statistical & machine learning tools
 - Both theory & programming knowledge required

Explain Your Findings

- Focus on questions, not methods
 - Given the data, what questions are interesting?
 - For questions, what tools are available?
- Know how to interpret the results
 - Explain black box models to non-computer experts

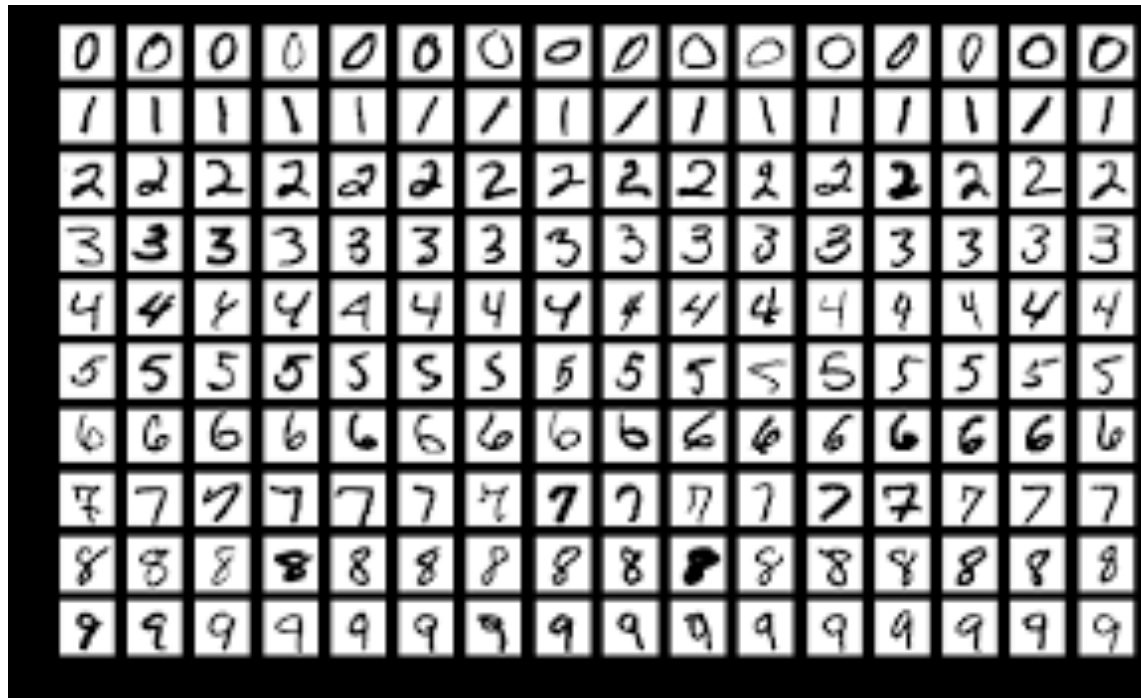
Key skills for data scientists

- Theory knowledge:
 - Practical machine learning & data mining techniques
 - Data integration
- Programming knowledge:
 - Programming tutorials
 - A deep learning tutorial with TensorFlow
- How to ask questions & present the results:
 - Individual & group projects

Theory Part: Machine Learning

- Classification
 - Data \rightarrow classes
- Regression
 - Predicting a numeric value
- Clustering
 - Discovering structures in data
- Many more

Classification: Digit Recognition



How can a
computer tell
0,1,2,...,9?

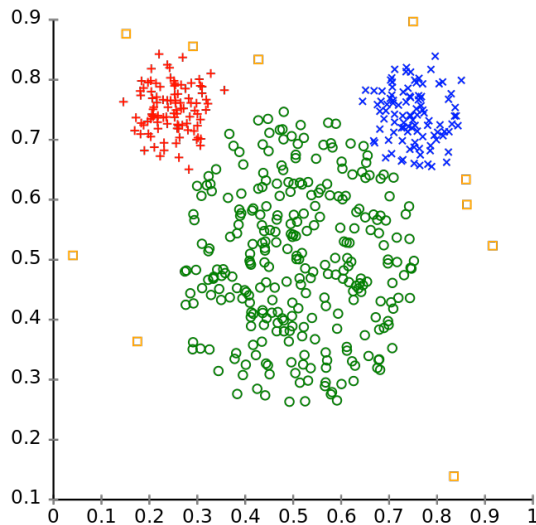
Regression: predicting stock value



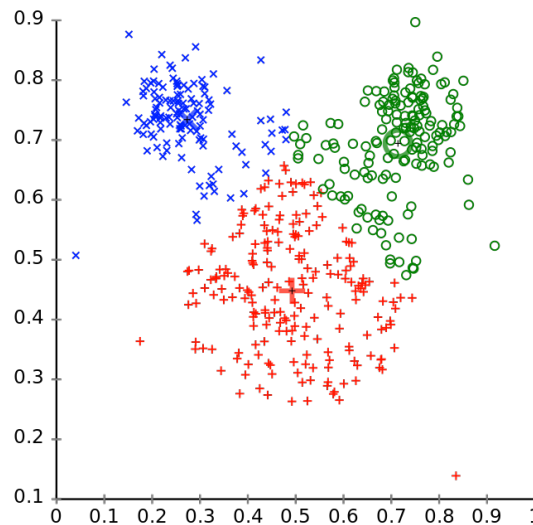
Clustering

Different cluster analysis results on "mouse" data set:

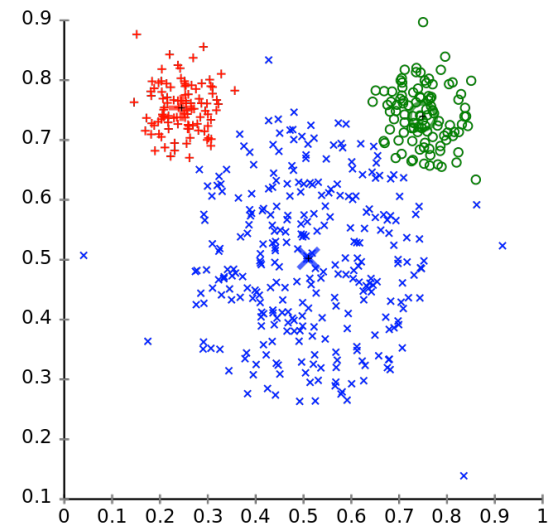
Original Data



k-Means Clustering



EM Clustering



Given unlabeled data, can you put similar data together?

Data Integration

- Databases are great: they let us manage huge amounts of data
- In reality, data sets are often created independently
 - Only to discover later that they need to combine their data!
 - At that point, they're using different systems, different schemata and have limited interfaces to their data.
- The goal of data integration: tie together different sources, controlled by many people, under a common schema.

Data Integration: A Higher-level Abstraction

Query

Mediated Schema

Semantic Mappings

- Independence of:
- source & location
 - data model, syntax
 - semantic variations
 - ...

S1

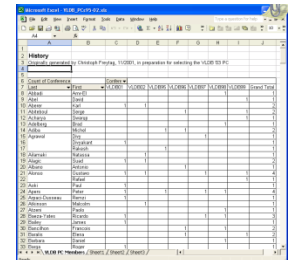
SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad
...

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

S2

```
<cd> <title> The best of ... </title>
<artist> Carreras </artist>
<artist> Pavarotti </artist>
<artist> Domingo </artist>
<price> 19.95 </price>
</cd>
```

S3



Data Cleansing

- What if the data has problems such as
 - Dummy values?
 - Absence of data?
 - Contradicting data?
 - Non-unique identifiers?
- Data cleansing helps to "clean" the data for later usage

Programming tools

- Python
 - Popular, easy to use programming language
- Sklearn
 - Python built-in libraries for data analytics
 - Will introduce parts relevant to lecture
- TensorFlow
 - Popular machine learning framework
 - Deep learning workshop (week 5)

Project tools

- Kaggle: Webpage for data science projects
 - Individual project to be familiar with competitions
 - Present your individual project on Kaggle
- GitHub: Open source platform
 - Share your individual project on Github

Presenting your projects

- It's important to explain what you did!
- Week 7: proposal presentation
- Week 11-12: project presentation
- Submit your final project report after week 12
- Details TBD

Wrap-up



Tentative schedule

Week	Lecture content	Tutorial
1	Course intro, math review	Python
2-4	Machine learning & data mining techniques	Sklearn, Kaggle
5	Deep learning workshop	TensorFlow
6	Midterm exam	
7	Public holiday	
8	Proposal presentation	
9-11	Data pre-processing	Github; other relevant topics
12-13	Project presentation	

Evaluation

- Midterm exam (week 6), in-class, 40%
- Individual project 20%
 - Kaggle in-class competition
- Group project 40%
- Details of project requirements will be released later
- No final exams

Midterm (40 pts)

- Week 6, in-class (Oct. 10)
- Closed book, closed notes
- Based on the contents in lecture 1~5

Individual project (20 pts)

- In-class Kaggle competition.
 - submit a valid result (5 pts)
 - Code quality on Github (5 pts)
 - Posting solutions on Kaggle (5 pts)
 - Ranking (5 pts)
- Bonus available
 - Leader-of-week (every Tuesday): 1 pt, may repeat
 - Top-3 of the competition: 5 pts, needs to do presentation in class.

Group projects (40 pts)

- Use the dataset we provide
- Define a problem, solve the problem
- In-class presentations, reports
 - Both proposal and final report
 - Details to be released later
- Bonus available: Up to 5 pts. We hope to see novel ideas in your project!

Bonus points

- One may get bonus on different parts at the same time
- Maximum grade may exceed 100

References

- Two books for your optional reading:
 - T. Hastie, ***The Elements of Statistical Learning***, Springer, 2009
 - ***Big Data Integration***, Morgan & Claypool, 2015, Luna Dong, Divesh Srivastava
- Programming contents available on the internet