Data Mining & Machine Learning

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA

ILLINOIS TECH

College of Computing

Schedule: Before Midterm Exam

- Data, Data Types
- Data Preprocessing
- Classification
  - KNN Classifier, Naïve Bayes Classifier
  - Decision Trees, SVM, Logistic Regression
  - Ensemble Classifier
  - Binary, Multi-Class and Multi-Label classifications
- Clustering
  - K-Means, Density-based Clustering, Hierarchical Clustering

Schedule: After Midterm

- Unsupervised Learning
  - Association Rules
  - Outlier Detections
- Semi-Supervised Learning
- Neural Networks
- Advanced Topics

Schedule

- Midterm Review
- Instructions for Final Project

Schedule

- Midterm Review
- Instructions for Final Project

Midterm Exam

- Live Students take exams in main campus, SB-107.
- Remote students take exams in India/China
  - Contact your local office for details
- Online students take exams at your selected locations
  - Rice Campus, IIT online will notify you the rooms
  - Main Campus, SB-107
  - US cities, if we have IIT office there
  - ProctorU (you need to pay a fee for it)

Midterm Exam

- Time: Oct 27, 11:30 AM – 12:50 PM
- Location: SB-107
- What you cannot bring
  - Books, Computers, Laptops
- What you can bring
  - One-page note, you can put anything on it
  - Calculator
  - Pen

Schedule: Before Midterm Exam

- Data, Data Types
- Data Preprocessing
- Classification
  - KNN Classifier, Naïve Bayes Classifier
  - Decision Trees, SVM, Logistic Regression
  - Ensemble Classifier
  - Binary, ~~Multi-Class and Multi-Label classifications~~
- Clustering
  - K-Means, ~~Density-based Clustering~~, Hierarchical Clustering

Midterm Exam Review

- Questions will be similar to assignments
  - No questions related to coding or Python
  - Similar to the calculation-based questions in HWs
- Questions in exams
  - Problem solving, e.g., classification, clustering, association rules, etc. They are similar questions in your assignments (calculation-based ones)
  - Concept questions
    - Basic concepts, e.g., what's the difference between supervised and unsupervised learning
    - Algorithms or solutions, e.g., describe how to build a decision tree (describe general workflow, no equations or formula)

Knowledge Review

- Data, Data Types
- Data Preprocessing
- Classification
  - KNN Classifier, Naïve Bayes Classifier
  - Decision Trees, SVM, Logistic Regression
  - Ensemble Classifier
  - Binary, ~~Multi-Class and Multi-Label classifications~~
- Clustering
  - K-Means, ~~Density-based Clustering~~, Hierarchical Clustering

Knowledge Review

- Data, Data Types
  - Nominal
    - General nominal
    - Binary
    - Ordinal
  - Numerical
    - Discrete
    - Continuous

Knowledge Review

- Data, Data Types
- Data Preprocessing
- Classification
  - KNN Classifier, Naïve Bayes Classifier
  - Decision Trees, SVM, Logistic Regression
  - Ensemble Classifier
  - Binary, ~~Multi-Class and Multi-Label classifications~~
- Clustering
  - K-Means, ~~Density-based Clustering~~, Hierarchical Clustering

Knowledge Review

- Data Preprocessing
  - Reduce noise in numerical variables by smoothing
  - Reduce redundancy by correlation/dependency
  - Reduce dimensions by feature selection/reduction
  - Data normalization
  - Data discretization
    - Numerical to Nominal
    - Nominal to Numerical

Knowledge Review

- Data, Data Types
- Data Preprocessing
- Classification
  - KNN Classifier, Naïve Bayes Classifier
  - Decision Trees, SVM, Logistic Regression
  - Ensemble Classifier
  - Binary, ~~Multi-Class and Multi-Label classifications~~
- Clustering
  - K-Means, ~~Density-based Clustering~~, Hierarchical Clustering

Knowledge Review

- Classification
  - Supervised vs Unsupervised Learning
  - Classification: definition, tasks
  - Classification Algorithms
    - KNN Classifier, Naïve Bayes Classifier, Decision Trees, SVM, Logistic Regression
    - Evaluation metrics
    - Understand how each algorithms work, pros, cons
    - Special issues, general issues and solutions
  - Ensemble Methods

Knowledge Review

- Data, Data Types
- Data Preprocessing
- Classification
  - KNN Classifier, Naïve Bayes Classifier
  - Decision Trees, SVM, Logistic Regression
  - Ensemble Classifier
  - Binary, ~~Multi-Class and Multi-Label classifications~~
- Clustering
  - K-Means, ~~Density-based Clustering~~, Hierarchical Clustering

Knowledge Review

- Clustering
  - Why it is unsupervised learning
  - Differences: partitional clustering, density-based clustering, hierarchical clustering
  - Know how KMeans and hierarchical clustering work
  - Know how to evaluate clustering outputs

Schedule

- Midterm Review
- Instructions for Final Project

About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do

About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do

About Final Projects

- Goals by the Final Project

  Examine your practical skills

  Train your research and experimental capabilities

  Train your presentation skills

  Train your paper or report writing skills

  Encourage you to learn going beyond the class

  Encourage your to solve problems by yourself

  Encourage you to work individually

  Encourage you to work in a team

About Final Projects

- Work In Team or Individually
  You can choose to work individually
  You can work together in a team
  A team can have up to 4 students
  More students in a team, better work expected
- Shared Project for ITMD 524
  You can work on a single project
  Or, you can work on two projects

About Final Projects

- Grading Details
  I will grade your final projects, not TA
  It is different from grading assignments.
- In terms of grading assignments
  You get 100 as long as your answers are correct
- In terms of grading final projects
  Your project will be compared with other groups
  For example, you get 80. It does not mean your project is not good, but there are better ones
- At the end of your presentation, I will give you feedbacks

About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do

About Final Projects

- Requirements (The minimal requirements)
  - Data size is at least with 100K rows
  - Well-defined problems
    - What are the questions in your data, or the problems you want to solve
    - What are the data mining tasks you want to perform
  - Appropriate solutions
  - Correct evaluations among multiple models
  - By using Python only
  - The degree of difficulty
    - It cannot be easier than the practice in the class or assignment
    - You should work on large data and void tiny data sets

About Final Projects

- Different Data Mining Tasks
  - Classification/Clustering/Association Rules/Outlier
- Different Applications
  - Information Retrieval/Web Mining/Recommender Sys
- For all of these options, you must have clear evaluations and comparisons. You cannot run a single solution without comparing with others

About Final Projects

- Notes: how to find data and define research problems
  - Where to find the data? See the following slides
  - You must find a data you are interested in
  - You may need more than one data and integrate them together
  - You need to understand the data first, and then think about what are the problems you want to solve
  - Then figure out whether you can use the techniques or models you learnt to solve the problem
  - If the current data is not ideal, you need to find another data and go through the steps again until you are satisfied with one data

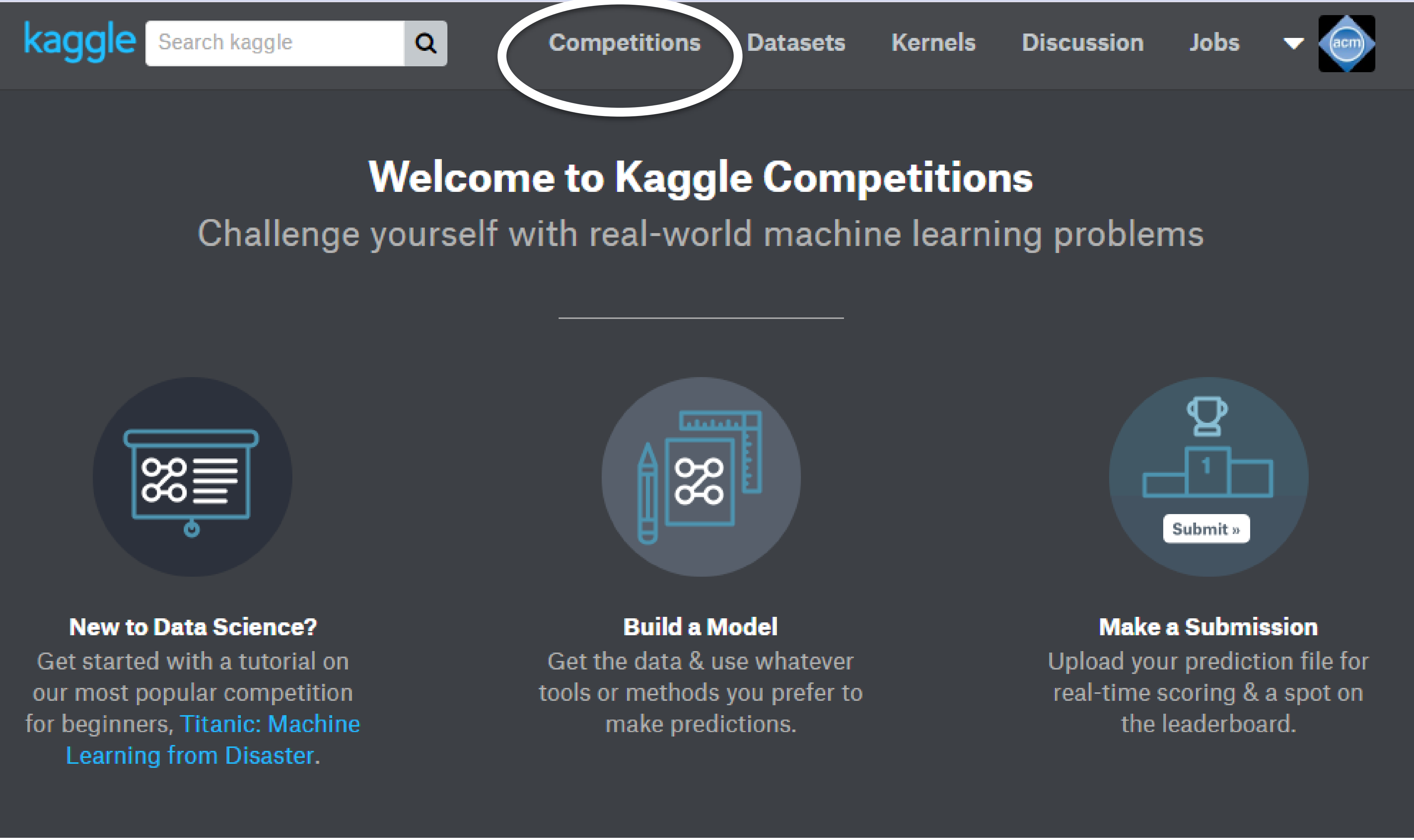
About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do

About Final Projects

- You should find data by yourself
  Kaggle: https://www.kaggle.com/
  Or other places to find appropriate data sets

Example: Data on Kaggle.com



Welcome to Kaggle Competitions

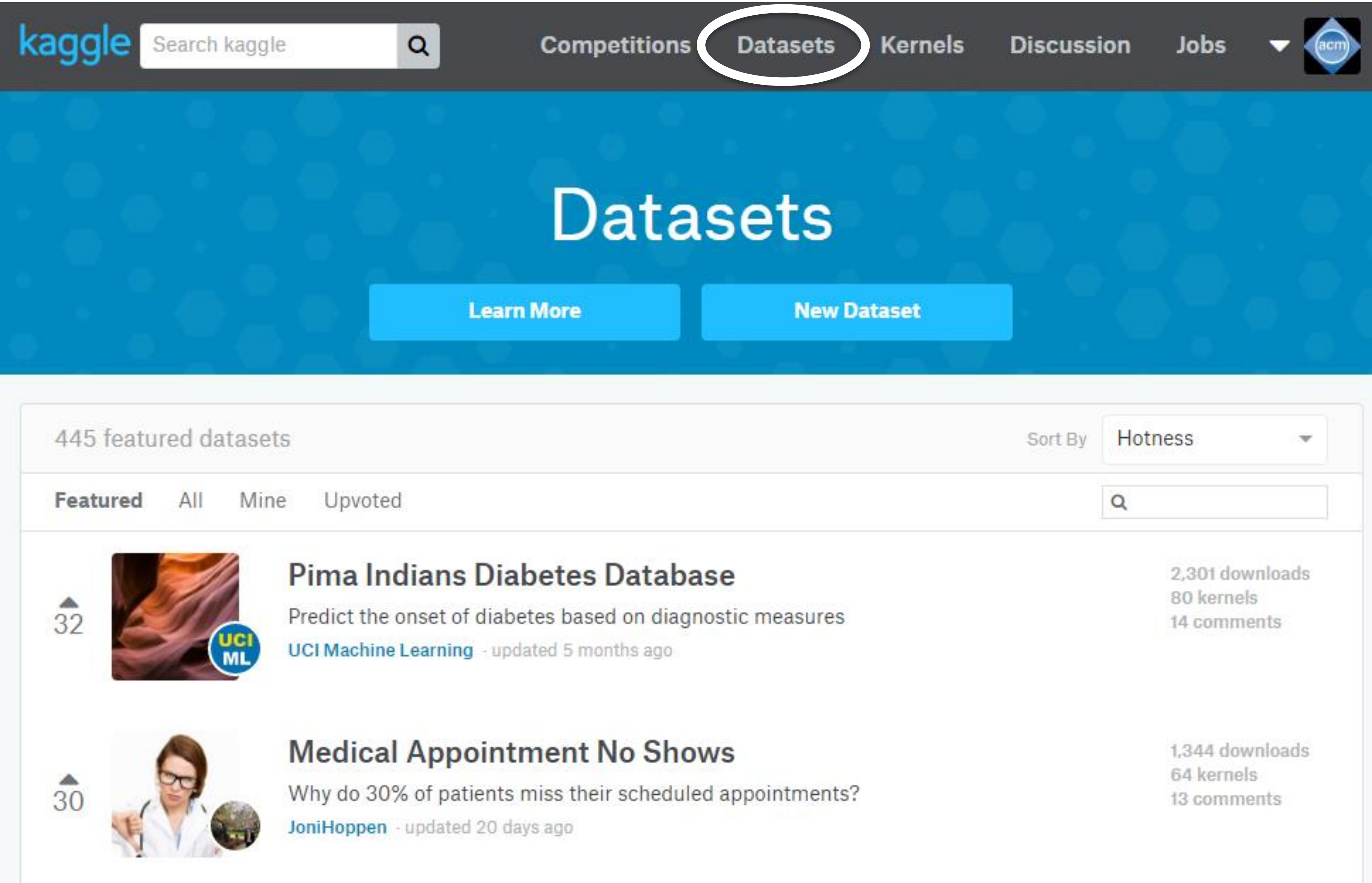
Challenge yourself with real-world machine learning problems

New to Data Science?
Get started with a tutorial on our most popular competition for beginners, Titanic: Machine Learning from Disaster.

Build a Model
Get the data & use whatever tools or methods you prefer to make predictions.

Make a Submission
Upload your prediction file for real-time scoring & a spot on the leaderboard.

Example: Data on Kaggle.com



Datasets

445 featured datasets

32 **Pima Indians Diabetes Database**
Predict the onset of diabetes based on diagnostic measures
UCI Machine Learning · updated 5 months ago
2,301 downloads
80 kernels
14 comments

30 **Medical Appointment No Shows**
Why do 30% of patients miss their scheduled appointments?
JoniHoppen · updated 20 days ago
1,344 downloads
64 kernels
13 comments

Raise some questions

- Do they make sense?
- Can we use data science techniques to solve them
- How to evaluate them?
- Any interpretations or explanations?
- How useful your outcomes are? Or, how can you use them in the future?

Rules

- You cannot use the following data
  - Weather data
  - Crime data
  - Housing data
  - Hotel data
  - Airline data
  - Game/Football/Basketball data
- They were frequently used by other students in previous years

About Final Projects

- General Idea
- Requirements
- Where to find the data
- Steps to do

About Final Projects

- Step 1: fill in a survey
- Step 2: decide to work individually or by team
- Step 3: write your project proposal
- Step 4: once your proposal is qualified, you can start working. Otherwise, you need to revise and resubmit the proposal
- Step 5: final project presentations and final report

About Final Projects

- Step 1: User Survey http://depaul.qualtrics.com/jfe/form/SV_3OGCoMY7U0Q997T
- Benefits
  - Give you a list of potential data/topics
  - Note: the data you can use in your project is not limited to this list
  - Collect your tastes on topics of final projects
  - Build better predictive or recommendation models to help students find the appropriate topics for projects

About Final Projects

- Step 2: decide to work individually or by team
  - There should be no more than 4 students in a team

About Final Projects

- Step 3: write the proposal
  - The template has been uploaded
  - Each team should ONLY submit one copy. For example, if there are 3 students in your team, it is enough for one student to submit the proposal. Do NOT submit multiple copies by different students!
  - Note: you may start working on it as soon as possible. Every semester, there are many teams whose proposal is disqualified.

About Final Projects

- Step 1: fill in a survey
- Step 2: decide to work individually or by team
- Step 3: write your project proposal
- Step 4: once your proposal is qualified, you can start working. Otherwise, you need to revise and resubmit the proposal
- Step 5: final project presentations and final report

About Final Projects

- Step 5: final project presentations and final report
  - Once you complete your presentation, I will give you feedbacks
  - It's your own choice to revise your projects in the final reports. If you did provide revisions, you should use RED fonts to mark them in your reports
  - Your score may be improved if you provide revisions in final reports
  - Your score, most likely, depends on your presentations

About Final Projects

- Proposal due: Nov 4
  - Must be submitted by one member of a group
  - Only one member, just put the name/email of all group members in the proposal
- Presentation: Nov 29 or/and Dec 1
  - It depends on how many groups we have
  - Remote students: send me a video
- Report Due: one week after presentation