2020 Spring Data Mining Syllabus

Instructor: Deokgun Park

Class hours: Tue/Thu 12:30-1:50pm
· Office hours: Tue/Thu 2:00-2:30pm

I am usually available after class for questions and discussions.

- Office: ERB 533
- E-mail: deokgun.park.uta.edu
- Homepage: http://crystal.uta.edu/~park/

TA: Md Ashaduzzaman Rubel Mondol

- Office hours: Tue/Thu 5-6 pm
- Office: ERB 509
- Email: mdashaduzzaman.mondol@mavs.uta.edu

Course Description

This is an introductory course on data mining. Data Mining refers to the process of automatic discovery of patterns and knowledge from large data repositories, including databases, data warehouses, Web, document collections, and data streams. The major topic we study includes the following:

- The fundamentals of the text mining
  - TF-IDF
  - Vector representation of words
  - Word Embedding
- Classifier
○ kNN
○ Naive Bayes
○ Support Vector Machines
○ Dimensionality reduction
○ Word embedding
  ● Association Analysis
  ● Clustering

Student Learning Outcomes:
A solid understanding of the basic concepts, principles, and techniques in data mining; an ability to analyze real-world applications, to model data mining problems, and to assess different solutions; an ability to design, implement, and evaluate data mining software. As a concrete outcome, each student will implement an app that can do following things:
  ● Build a classifier
  ● Conduct a clustering analysis
  ● Conduct Association Analysis

Textbook
Introduction to Data Mining by Pang-ning Tan, Michael Steinbach, and Vipin Kumar. 3rd edition.  

References (optional)
Introduction to Information Retrieval (IR)  
Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze  

Mining of Massive Datasets (MMDS)  
Jure Leskovec, Anand Rajaraman, Jeff Ullman  
Available freely at http://www.mmds.org/

Word Embedding  
https://www.tensorflow.org/tutorials/representation/word2vec

CNN  
https://www.tensorflow.org/tutorials/estimators/cnn

Image Captioning
Using Google colab
https://hackernoon.com/begin-your-deep-learning-project-for-free-free-gpu-processing-free-storage-free-easy-upload-b4dba18abebc

Creating an image recognition app that runs on the browser
https://medium.com/tensorflow/train-on-google-colab-and-run-on-the-browser-a-case-study-8a45f9b1474e

References
http://facweb.cs.depaul.edu/mobasher/classes/ect584/lecture.html
https://www.cs.purdue.edu/homes/lsi/CS473_Fall_2013/CS490W.html

Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Course Content</th>
<th>Reading &amp; Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 21</td>
<td>Course introduction.</td>
<td></td>
</tr>
<tr>
<td>Jan 23</td>
<td>Ch1. Data Mining process</td>
<td>Profile</td>
</tr>
<tr>
<td>Jan 28</td>
<td>Ch2. Data</td>
<td></td>
</tr>
<tr>
<td>Jan 30</td>
<td>Ch3. Exploring data</td>
<td>Project Preliminary due</td>
</tr>
<tr>
<td>Feb 4</td>
<td>Ch3. Exploring Data</td>
<td></td>
</tr>
<tr>
<td>Feb 6</td>
<td>Tutoria Lab - Jupyter notebook</td>
<td></td>
</tr>
<tr>
<td>Feb 11</td>
<td>Decision Tree</td>
<td></td>
</tr>
<tr>
<td>Feb 13</td>
<td>Decision Tree</td>
<td></td>
</tr>
<tr>
<td>Feb 18</td>
<td>Evaluation</td>
<td></td>
</tr>
<tr>
<td>Feb 20</td>
<td>kNN</td>
<td></td>
</tr>
<tr>
<td>Feb 25</td>
<td>kNN</td>
<td></td>
</tr>
<tr>
<td>Feb 27</td>
<td>Naive Bayes</td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------</td>
<td></td>
</tr>
<tr>
<td>Mar 3</td>
<td>Naive Bayes</td>
<td></td>
</tr>
<tr>
<td>Mar 5</td>
<td>Deep Learning</td>
<td></td>
</tr>
<tr>
<td>Mar 10-12</td>
<td>No class (Spring break)</td>
<td></td>
</tr>
<tr>
<td>Mar 17</td>
<td>Deep Learning</td>
<td></td>
</tr>
<tr>
<td>Mar 19</td>
<td>Deep Learning</td>
<td></td>
</tr>
<tr>
<td>Mar 24</td>
<td>SVM</td>
<td></td>
</tr>
<tr>
<td>Mar 26</td>
<td>SVM</td>
<td></td>
</tr>
<tr>
<td>Mar 31</td>
<td>Ensemble Methods</td>
<td></td>
</tr>
<tr>
<td>Apr 2</td>
<td>Class imbalance, multiclass</td>
<td></td>
</tr>
<tr>
<td>Apr 7</td>
<td>Association Analysis</td>
<td></td>
</tr>
<tr>
<td>Apr 9</td>
<td>Association Analysis</td>
<td></td>
</tr>
<tr>
<td>Apr 14</td>
<td>Clustering I</td>
<td></td>
</tr>
<tr>
<td>Apr 16</td>
<td>Clustering II</td>
<td></td>
</tr>
<tr>
<td>Apr 21</td>
<td>Buffer</td>
<td></td>
</tr>
<tr>
<td>Apr 23</td>
<td>Buffer</td>
<td></td>
</tr>
<tr>
<td>Apr 28</td>
<td>Visual Analytics I</td>
<td></td>
</tr>
<tr>
<td>Apr 30</td>
<td>Visual Analytics II</td>
<td></td>
</tr>
<tr>
<td>May 4</td>
<td>Artificial General Intelligence</td>
<td></td>
</tr>
<tr>
<td>May 7</td>
<td>Final exam review</td>
<td></td>
</tr>
</tbody>
</table>

**Expectations**

This will be a challenging course. I don't recommend students taking more than two challenging courses per semester. One common mistake that newly admitted ambitious master students make is taking too much challenging courses just because they would
like to quickly get the marketable skills. It can actually ruin your grade and quality of life especially if you just arrived in the USA from other places like India or China. Instead distribute challenging courses wisely.

Many class hours will be devoted to the project progress report including project idea pitching and final presentation. The rationale is that it is the best use of the course and the instructor resource, because it can provide personalized care and feedback to the individual students. And it is perfectly okay for students to point when the instructor is not right. For students from foreign cultures such as indian or chinese, they might not be accustomed to this. But the instructor is not perfect and by pointing the mistake, you are helping the class as a whole including the instructor learn better. In this point, I would like to give extra credit of 1 for pointing me wrong.

I will not check the attendance. But there will be a quiz about the previous class before every class.

No laptop or mobile phone use is allowed during class.

Please use the slack channel to ask questions, too. Please use the public message if the message is not sensitive. The rationale is that many students will have similar questions and they can get information, too. It is encouraged to help other students with slack.

Grades

- Term project 30%
  - There will be a semester long project. It will have multiple parts that will be graded separately.
- Final exam 30%
  - (Time and place: TBA)
- Quiz 20%
  - There will be an online-quiz before every lecture. The questions will be from the previous class.
- Assignment 20%

The final letter grades will be based on students' performance. There is no pre-defined cutoffs or distribution of grades. Undergraduate and graduate students are compared in separate groups. About 30% get ‘A’, 40% get ‘B’. The percentage is changing every semester.
Term Project

The goal of the term project is building a classifier that you can show to someone with your homepage that will help you look competent. We will participate in a Kaggle challenge (https://www.kaggle.com/c/nlp-getting-started/overview). I assume that you will work hard to get a job after graduation. If you want to go to academia and want to work on more academic project, please let me know. Please let me know if you want different term project. Coding the right solution is only half of the output. Your well-written report will be 50% of your grade. It means if you struggle the coding part, there is still a chance by writing a good report. Below are the modules for the term project

1. Preliminary (10%)
   a. Build your personal homepage
      i. You can use any format you want or reuse the homepage you already have. The only requirement is that you should have all the contents in your resume in your homepage. Put a download link to your resume. Use below template and mimic my homepage if you want.
      ii. https://sourcethemes.com/academic/
      iii. http://crystal.uta.edu/~park/
   b. Create a Linkedin profile. You should have all the contents in your resume and the link to your homepage.
   c. Add your photo and one sentence to the directory
      i. https://docs.google.com/document/d/1v5gQbMysQT7tKsVPfwYIrRHmudYVUP4TueD80qav4KU/edit?usp=sharing
   d. Submit
      i. Put a link for your homepage in this google doc. It should have a link to your homepage and Linkedin profile.
   e. Grading rubric
      i. Your resume is reasonably professional (2 point)
      ii. Your homepage has all the info in your resume (3 point)
      iii. Your homepage has a download link for your resume (1 point)
      iv. Your LinkedIn profile has all the info in your resume and (3 point)
      v. Your LinkedIn profile has a link to your homepage (1 point)
      vi. During the class, we will have a voting for the following award. The award winner will get 5 extra credits.
         1. Best homepage award
         2. Best LinkedIn profile award
   f. Example home pages
      i. based on following the rubric,
      ii. Presentation,
      iii. keeping it simple but contain all required information, and uniqueness
2. Practice Classifier (10%)
   a. Participate in the Kaggle competition
      i.  https://www.kaggle.com/c/nlp-getting-started
      ii. Write jupyter notebook
      iii. Put the notebook in your homepage
   b. Submit
      i. You don’t have to submit the result. We will visit your homepage to check it.
   c. Grading rubric
      i. You have ranking (50%)
      ii. You have jupyter notebook (50%)

3. Main Competition (70%)
   a. Use the board game geek review data
      i.  https://www.kaggle.com/jvanelteren/boardgamegeek-reviews
   b. Your goal is given the review, predict the rating. You can refer the code or tutorial internet. But main question you have to answer is what improvement you made over the existing reference.
   c. Documentation is the half of your work. Write a good blog post for your work and step-by-step how to guide for github readme.md
   d. Grading Criteria
      i. Demo: Developed predictor of reviews
         1. no localhost allowed (If you cannot deploy it in your homepage, ask TA how. If you cannot do, attach a good video. You will get some penalty)
         2. Show calculation step as much as possible
            a. For example, show probability scores for query and classes if you are using Naive Bayes.
      ii. GitHub: Have a readme.md for your github code to explain step by step deployment instruction
      iii. Report
         1. Upload your jupyter notebook to the Kaggle
         2. Add your jupyter notebook to your homepage
         3. Have your reference
         4. Explicitly state what is your contribution over the reference
            a. For graduate student, Engineering contribution such as changing the version of python, adapting to different server platform are not accepted as the contribution. Accepted contributions are things such as implementing optimization idea in the text book.
5. Describe what was your challenge and how you solved it 1 point
6. Have some experiments and explain your finding 1 point
   a. Hyper parameter tuning
   b. Overfitting
7. Explain the basic algorithms
8. Evaluation score
iv. Extra credit
   1. Use word embedding
   2. Good visualization

4. Project show video (10%)  
   a. Build a 1 minute project video advertising your app and upload it to the youtube. Put a video link in this google doc.
   b. Grading rubric
      i. We will watch the video during the class together.
      ii. I will grade your video myself.
      iii. We will do voting for the following award. Award winners will get 5 extra points
         1. Most professional video award
         2. Most beautiful video award
         3. Funniest video award
      iv. Make sure the video is easily discoverable in the homepage project description

5. Some of the presentation and implementation ideas from spring 2019
   d. https://tungpv.com/project/ted-recommender/

Academic Integrity

This is a graduate/senior level course. I take cheating very seriously. I will give 'F' to the FIRST cheating effort. Please note that some international students may not be familiar with how serious the plagiarism can harm your career. If you cannot follow the coursework, it is better to drop the course or to do your best and get 'C' than do cheating and get a 'F' or suspended for your professional career.
Some students got the D grade because they copied each other's work for the makeup assignment for an extra point. Some of them would have gotten 'B' if they did not submit the assignment at all. You are young and you might not know what is important and what is not yet. For example, GPA is not as important at graduate level. Your potential employer will not hire you because you have a perfect GPA. They will hire you when you demonstrate expert skills and when you are trustworthy. Your GPA will matter when you are going to get PhD. Still, if you get D in one course and get A in every other course, they will give you a chance. Here is my secret. My undergraduate GPA is 2.56 out of 4.3, which is ridiculously low. Still you can recover. Don’t think that you will ruin your life when you get a low grade.

However, let’s say that you cheat and don’t get a bad penalty today. Probably you will try similar behavior later. People do not change easily. Maybe you will not caught for long. But once you caught during the graduate school or during the professional career, you will get an irreversible damage and probably have to find another career in other field. The punishment will be irrationally severe than what you think you deserve. The rationale behind this is because the probability of getting caught is low and they would like to warn other people not to try such a behavior. Remember the expected outcome is the product of the probability of getting caught and the punishment you will get when you caught. Because the probability is low, the punishment should be higher to compensate. Otherwise people will cheat always because the expected benefit of cheating is higher than the expected penalty. By the way, medicine is the passphrase. You are still young to understand this fully, but as an educator I feel strongly obligated to teach this lesson before it is too late.

Similarly very common plagiarism mistake the students make is copying image from web without giving proper credit. In this case, because it is not strictly or intentionally copying, I gave the report part 0 point. It is still terrible mistake and you should not do that in professional report.

Profile

Even though I am quite bad remembering name, I am trying to get better. Please help me by uploading your name and photo and one sentence here. I put mine as an example. Profiles will be used for the main feedback area for the project.

https://docs.google.com/document/d/1v5gQbMysQT7lKsVPfwYIrRHmudYVUP4TueD80qav4KU/edit?usp=sharing

Slack channel

In this class, we will use slack as a main communication medium. If you send me an email, you get a penalty. It is for not reading the syllabus carefully. We're always experimenting
with how to structure our online discussions. I highly encourage you to be part of the conversation: speak up with thoughts, links, ideas, updates, and anything that comes to mind. Most importantly, relax and enjoy chatting with others, no pressure.¹ You can join at https://datamining2019spring.slack.com/ or using the following link.

https://join.slack.com/t/2020springuta-vl41572/shared_invite/enQtOTE3Mjg1OTAxODk1LTEExMGN1MTA0ZwE5OTihNTI3NzQ0MjU5N2IzOjc5NGJlMTc5YTUyMzdkMTliNzExMDFmNzFmMWJiNGI0MGQ1MjM

Resolving Grading issues

The course will be graded relatively. About 30% of the students will get A and 40% of the students will get B. The actual threshold or number of the grades will be depend on the instructor. If you have a question about the grade, please meet TA and resolve first. If it does not resolve the issue, then you can contact me.

Frequently asked questions

1. I would like to get research experience in my lab. May I join your lab as a master student?

You can contact me and discuss your interest. However, if you want to get a job after graduation, doing research with me might not be helpful. You might get a better chance for the job by preparing the coding interview. The research takes time to produce a meaningful outcome. And the research is a full-time job. For example, it is common that it takes more than two years before a phd student can produce a paper or meaningful software. However, this can be too late for the master students who need something to show to potential employer during first semester and third semester. Doing research is more beneficial for those who consider phd program after graduation.

¹ Paragraph cited from MIT 6.8099: Artificial General Intelligence Syllabus. https://docs.google.com/document/d/1ZqgghxV1lpZeWu5zNK0gMUBHfYTtw9n6eYzzx918nok/edit#heading=h.4qyhpup1515w
**Attendance:** At The University of Texas at Arlington, taking attendance is not required but attendance is a critical indicator in student success. Each faculty member is free to develop his or her own methods of evaluating students’ academic performance, which includes establishing course-specific policies on attendance. As the instructor of this section, I require all students to attend lectures. However, while UT Arlington does not require instructors to take attendance in their courses, the U.S. Department of Education requires that the University have a mechanism in place to mark when Federal Student Aid recipients “begin attendance in a course.” UT Arlington instructors will report when students begin attendance in a course as part of the final grading process. Specifically, when assigning a student a grade of F, faculty report the last date a student attended their class based on evidence such as a test, participation in a class project or presentation, or an engagement online via Blackboard. This date is reported to the Department of Education for federal financial aid recipients.

---

**Announcements:** Stay tuned and make sure to check Blackboard frequently. Important announcements will be posted there.

**Assignments and Deadlines**

- All the assignments must be submitted through Blackboard. **We will NOT take hardcopy or email submission, unless the university verifies that Blackboard was malfunctioning or unavailable. If you are not able to submit through Blackboard due to its technical failure, you can email your assignment to us, together with a screenshot showing the technical failure. We will verify with the university.**
- Everything is due by 11:59pm on the due date. The deadline is automatically managed by Blackboard. You can still turn in assignment after the deadline. However, you automatically lose 5 points per hour after the due time, till you get 0. (Each individual assignment is 100 points.) **We cannot waive the penalty, unless there was a case of illness or other substantial impediment beyond your control, with proof in documents.**
- **Do not send email asking for an extension for the following reasons:**
Regrading: Regrading request must be made within 7 days after we post scores on Blackboard. TA will handle regrade requests. If student is not satisfied with the regarding results, you get 7 days to request again. The instructor will regrade, and the decision is final.

Drop Policy: Students may drop or swap (adding and dropping a class concurrently) classes through self-service in MyMav from the beginning of the registration period through the late registration period. After the late registration period, students must see their academic advisor to drop a class or withdraw. Undeclared students must see an advisor in the University Advising Center. Drops can continue through a point two-thirds of the way through the term or session. It is the student's responsibility to officially withdraw if they do not plan to attend after registering. By the way, “bazinga” is the word being asked in the quiz 1. Students will not be automatically dropped for non-attendance. Repayment of certain types of financial aid administered through the University may be required as the result of dropping classes or withdrawing. For more information, contact the Office of Financial Aid and Scholarships (http://wweb.uta.edu/aaofao/).

Disability Accommodations: UT Arlington is on record as being committed to both the spirit and letter of all federal equal opportunity legislation, including The Americans with Disabilities Act (ADA), The Americans with Disabilities Amendments Act (ADAAA), and Section 504 of the Rehabilitation Act. All instructors at UT Arlington are required by law to provide “reasonable accommodations” to students with disabilities, so as not to discriminate on the basis of disability. Students are responsible for providing the instructor with official notification in the form of a letter certified by the Office for Students with Disabilities (OSD). Only those students who have officially documented a need for an accommodation will have their request honored. Students experiencing a range of conditions (Physical, Learning, Chronic Health, Mental Health, and Sensory) that may cause diminished academic performance or other barriers to learning may seek services and/or accommodations by contacting:

The Office for Students with Disabilities, (OSD) www.uta.edu/disability or calling 817-272-3364. Information regarding diagnostic criteria and policies for obtaining disability-based academic accommodations can be found at www.uta.edu/disability.

Counseling and Psychological Services, (CAPS) www.uta.edu/caps/ or calling 817-272-3671 is also available to all students to help increase their
understanding of personal issues, address mental and behavioral health problems and make positive changes in their lives.

**Non-Discrimination Policy:** The University of Texas at Arlington does not discriminate on the basis of race, color, national origin, religion, age, gender, sexual orientation, disabilities, genetic information, and/or veteran status in its educational programs or activities it operates. For more information, visit uta.edu/eos.

**Title IX Policy:** The University of Texas at Arlington ("University") is committed to maintaining a learning and working environment that is free from discrimination based on sex in accordance with Title IX of the Higher Education Amendments of 1972 (Title IX), which prohibits discrimination on the basis of sex in educational programs or activities; Title VII of the Civil Rights Act of 1964 (Title VII), which prohibits sex discrimination in employment; and the Campus Sexual Violence Elimination Act (SaVE Act). Sexual misconduct is a form of sex discrimination and will not be tolerated. For information regarding Title IX, visit www.uta.edu/titleIX or contact Ms. Jean Hood, Vice President and Title IX Coordinator at (817) 272-7091 or jmhood@uta.edu.

**Academic Integrity:** Students enrolled all UT Arlington courses are expected to adhere to the UT Arlington Honor Code:

> I pledge, on my honor, to uphold UT Arlington’s tradition of academic integrity, a tradition that values hard work and honest effort in the pursuit of academic excellence.

> I promise that I will submit only work that I personally create or contribute to group collaborations, and I will appropriately reference any work from other sources. I will follow the highest standards of integrity and uphold the spirit of the Honor Code.

UT Arlington faculty members may employ the Honor Code in their courses by having students acknowledge the honor code as part of an examination or requiring students to incorporate the honor code into any work submitted. Per UT System Regents’ Rule 50101, §2.2, suspected violations of university’s standards for academic integrity (including the Honor Code) will be referred to the Office of Student Conduct. Violators will be disciplined in accordance with University policy, which may result in the student’s suspension or expulsion from the University. Additional information is available at https://www.uta.edu/conduct/.

**Electronic Communication:** UT Arlington has adopted MavMail as its official means to communicate with students about important deadlines and
events, as well as to transact university-related business regarding financial aid, tuition, grades, graduation, etc. All students are assigned a MavMail account and are responsible for checking the inbox regularly. There is no additional charge to students for using this account, which remains active even after graduation. Information about activating and using MavMail is available at [http://www.uta.edu/oit/cs/email/mavmail.php](http://www.uta.edu/oit/cs/email/mavmail.php).

**Campus Carry:** Effective August 1, 2016, the Campus Carry law (Senate Bill 11) allows those licensed individuals to carry a concealed handgun in buildings on public university campuses, except in locations the University establishes as prohibited. Under the new law, openly carrying handguns is not allowed on college campuses. For more information, visit [http://www.uta.edu/news/info/campus-carry/](http://www.uta.edu/news/info/campus-carry/).

**Student Feedback Survey:** At the end of each term, students enrolled in face-to-face and online classes categorized as “lecture,” “seminar,” or “laboratory” are directed to complete an online Student Feedback Survey (SFS). Instructions on how to access the SFS for this course will be sent directly to each student through MavMail approximately 10 days before the end of the term. Each student’s feedback via the SFS database is aggregated with that of other students enrolled in the course. Students’ anonymity will be protected to the extent that the law allows. UT Arlington’s effort to solicit, gather, tabulate, and publish student feedback is required by state law and aggregate results are posted online. Data from SFS is also used for faculty and program evaluations. For more information, visit [http://www.uta.edu/sfs](http://www.uta.edu/sfs).

**Final Review Week:** For semester-long courses, a period of five class days prior to the first day of final examinations in the long sessions shall be designated as Final Review Week. The purpose of this week is to allow students sufficient time to prepare for final examinations. During this week, there shall be no scheduled activities such as required field trips or performances; and no instructor shall assign any themes, research problems or exercises of similar scope that have a completion date during or following this week unless specified in the class syllabus. During Final Review Week, an instructor shall not give any examinations constituting 10% or more of the final grade, except makeup tests and laboratory examinations. In addition, no instructor shall give any portion of the final examination during Final Review Week. During this week, classes are held as scheduled. In addition, instructors are not required to limit content to topics that have been previously covered; they may introduce new concepts as appropriate.
**Emergency Exit Procedures:** Should we experience an emergency event that requires us to vacate the building, students should exit the room and move toward the nearest exit. When exiting the building during an emergency, one should never take an elevator but should use the stairwells. Faculty members and instructional staff will assist students in selecting the safest route for evacuation and will make arrangements to assist individuals with disabilities.

**Student Support Services:** UT Arlington provides a variety of resources and programs designed to help students develop academic skills, deal with personal situations, and better understand concepts and information related to their courses. Resources include tutoring, major-based learning centers, developmental education, advising and mentoring, personal counseling, and federally funded programs. For individualized referrals, students may visit the reception desk at University College (Ransom Hall), call the Maverick Resource Hotline at 817-272-6107, send a message to resources@uta.edu, or view the information at [http://www.uta.edu/universitycollege/resources/index.php](http://www.uta.edu/universitycollege/resources/index.php).
Data Mining Introduction

1. Big data, datamining, machine learning, deep learning, ai
2. Data

Raggad’s classification

Knowledge

Information

Data

Noise

accepted facts, principles, or rules of thumb that are useful for specific domains. Knowledge can be the result of inferences and implications produced from simple information facts.

Processed Data

raw facts with a known coding system

raw facts with an unknown coding system

4. Knowledge Discovery (KDD) Process

5. Big Data Value Generation Process

a.
7. Sense Making Loop

8. The 4Vs
   a. Volume
      i. Sampling
         1. Random sampling does not work well with relational database
   b. Variety
   c. Velocity
   d. Veracity
      i. Fake news
      ii. Manipulating public opinion
9. Structured vs Unstructured Data
   a. Structured data
      i. Relational database tables
      ii. CSV
      iii. SQL
   b. Semi-structured data
      i. XML, JSON
      ii. NoSQL
   c. Unstructured data
      i. text
      ii. image
      iii. Information Retrieval & Visual Analytics
11. What we deal in the class
   a. multi-dimensional data
      i. Transaction data
   b. Text
   c. Multimedia
      i. image
      ii. movie
      iii. sound
12. What we don’t deal in the class
   a. Graph
      i. World wide web
      ii. Social network
   b. Sequence
      i. Hospital medical record
      ii. stock trading
   c. Geographic data
      i. Google map
   d. Spatio-temporal
      i. Real-time event monitoring in social media
13. Which companies are the most valuable tech company?
   a. Google
   b. Amazon
Multi-dimensional data

1. Multi-dimensional data
   a. tabular data
   b. relational database
2. Attributes and Objects
   a. attributes
      i. variable, field, characteristic, feature, data column
   b. objects
      i. record, point, case, sample, entity, instance, or data row

3. Types of Attributes
   a. by measure theory
      i. Categorical (Qualitative) attribute
         1. nominal ==, !=
         2. ordinal <>
      ii. Numeric (Quantitative) attribute
         1. interval + -
            a. The temperature in San Francisco is 50 Degree Fahrenheit and 100 degree Fahrenheit in Arlington. Is the following sentence scientifically correct?
            b. The temperature in Arlington is two times higher than SF.
2. ratio * /
   b. By number of values
      i. Discrete attribute
      ii. continuous attribute
4. Transformation
   a. Nominal
      i. permutation
   b. Ordinal
      i. order preserving changing \{1,2,3\} -> \{A, B, C\}
      ii. Likert scale
   c. Interval
      i. Fahrenheit to celsius
   d. Ratio
      i. Length meters to feet
5. Characteristics of Structured Data
   a. Dimensionality
      i. Curse of Dimensionality
   b. Sparsity
   c. Data Quality
      i. Noise and outliers
      ii. Absent values
         1. One mistake to fill the missing values for 0
      iv. duplicate data
Statistical Summary

1. Descriptive statistics
   a. Mode
   b. median
   c. mean
   d. geometric mean
      i. 
      ii. When the numbers have different range
      iii. When the events are related
         1. ex) Five years annual return on investment
         2. 90%, 10%, 20%, 30% and -90%
   e. harmonic mean
      \[ H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left( \frac{\sum_{i=1}^{n} x_i^{-1}}{n} \right)^{-1} \]
      i. 
      ii. Mitigate the large outlier
      iii. Useful when there are trade off between two values
         1. ex) Precision and Recall
   f. percentiles
   g. standard deviation
   h. entropy
   i. contingency
      
      |     | Dog | Cat | Total |
      |-----|-----|-----|-------|
      | Male| 42  | 10  | 52    |
      | Female| 9  | 39  | 48    |
      | Total| 51  | 49  | 100   |
   j. 
   k. Confusion matrix
ii.

2. Inferential statistics
   a. Null hypothesis test
      i. Null hypothesis is default state
         1. cancer detecting machine
            a. null hypothesis is one does not have a cancer
      ii. Type I error
         1. rejection of a true null hypothesis
         2. When one does not have a cancer (true null hypothesis),
            the system says one has a cancer (rejection of a true null hypothesis)
            a. False positive
      iii. Type II error
         1. acception of a false null hypothesis
   iv. one side test
   v. two side test

<table>
<thead>
<tr>
<th>Decision based on sample</th>
<th>Truth about the hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>H₀ is true</td>
<td>Fail to reject H₀</td>
</tr>
<tr>
<td>H₀ is false</td>
<td>Correct Decision (probability = 1 - α)</td>
</tr>
<tr>
<td></td>
<td>Type II Error - fail to reject H₀ when it is false (probability = β)</td>
</tr>
<tr>
<td>Reject H₀</td>
<td>Type I Error - rejecting H₀ when it is true (probability = α)</td>
</tr>
<tr>
<td></td>
<td>Correct Decision (probability = 1 - β)</td>
</tr>
</tbody>
</table>
b. Rational
   i. T-test vs Z-test
      1. Means of two groups are significantly different?
      2. Sample variance vs population variance
      3. Number of samples
   ii. Sample variance vs p

   Population variance
   1. Population variance = \( \frac{N}{N-1} \times \text{sample variance} \)

c. Nominal
   i. Chi-square test
      1. For example, in test, all four option A, B, C, D is equal probability?
      2. Sample: A: 20, B: 20, C: 25, D: 35


d. Correlation
   i. Nominal
      1. Cramer’s V
         \[
         V = \sqrt{\frac{\varphi^2}{\min(k - 1, r - 1)}} = \sqrt{\frac{\chi^2/n}{\min(k - 1, r - 1)}}
         \]
   ii. Ordinal
1. Spearman’s correlation
   a. [https://www.spss-tutorials.com/spearman-rank-correlation/](https://www.spss-tutorials.com/spearman-rank-correlation/)

iii. Rational
   1. Pearson’s correlation
      \[
      \rho_{X,Y} = \frac{\text{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}
      \]
   2. \[
   \rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}
   \]

iv. Correlation

<table>
<thead>
<tr>
<th>1</th>
<th>0.8</th>
<th>0.4</th>
<th>0</th>
<th>-0.4</th>
<th>-0.8</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

v.

vi.

3.
Natural language processing overview

1. n-gram or bag of words assumption
   a. Search / sentiment analysis
      i. I eat a fish
      ii. Fish eat me.
   b.
2. knowledge graph
   a. I eat a red apple at the restaurant.
      i. (I, eat, apple)
      ii. (apple, color, red)
      iii. (I, be at, restaurant)
   b. question and answer
      c. https://openie.allenai.org/
3. grounded language
   a. symbol grounding problem
   b. grounded language
Information retrieval

- **IR vs RDBMS**
  - Relational Database or structured data
    - Semantics (meaning of data attributes) are well defined
    - Complex query language
    - exact retrieval for what you ask
    - emphasis on efficiency
  - Information Retrieval
    - unstructured data
    - simple query language
    - You should get what you want, even the query is bad
    - Effectiveness is primary issue

- **SQL - Google - Visual Analytics**
  - Known unknown
  - Unknown unknown
  - Can you detect anomaly using search?

- Search and Rank
- Query
- Document
- Database | corpus
- Corpora
- ad-hoc IR vs filtering
  - Filtering: Queries are stable
  - IR: collection changes

- Search
  - Similarity or distance measure
  - Relevance

- Rank
  - similarity
  - importance

- Indexing
  - zipf’s law
Stopwords removal

Original Text: Information retrieval deals with the representation, storage, organization of, and access to information items

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1130021</td>
<td>market</td>
<td>52110</td>
</tr>
<tr>
<td>of</td>
<td>547311</td>
<td>bank</td>
<td>47940</td>
</tr>
<tr>
<td>to</td>
<td>516636</td>
<td>stock</td>
<td>47401</td>
</tr>
<tr>
<td>a</td>
<td>464736</td>
<td>trade</td>
<td>47310</td>
</tr>
<tr>
<td>in</td>
<td>390819</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>and</td>
<td>387703</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Statistics collected from Wall Street Journal (WSJ), 1987

Pointwise mutual information

Original Text: Information retrieval deals with the representation, storage, organization of, and access to information items
- Porter Stemmer (Stopwords removed): Inform retrieve deal represent storag organ access inform item
- Online example:
  http://maya.cs.depaul.edu/~classes/ds575/porter.html

- Lemmatization
- Inverted indexing
- Exact Boolean Search
- Ranked Boolean
- TF/IDF
- Vector Space Model

<table>
<thead>
<tr>
<th></th>
<th>Java</th>
<th>Sun</th>
<th>Starbucks</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Query</td>
<td>1</td>
<td>0.2</td>
<td>1</td>
</tr>
</tbody>
</table>

- Vector Coefficients
- Latent Semantic Indexing
<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>information</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>retrieval</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>machine</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>learning</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>protein</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>mutation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>expression</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Using singular value decomposition (SVD) to find the small set of concepts/topics

\[ X = U S V^T \]

\[ x_k = \sum_{i \leq k} u_i s_i v_i^T \]
Machine Learning Overview

- Supervised learning
  - classifier
  - Predictor
  - Ground truth data
- Unsupervised learning
  - cluster
  - Example
    - “Lion king” belongs to same category as
      - “tiger”
      - “avengers”
- Semi-supervised learning (Reinforcement learning)
Classifier

- Process overview

<table>
<thead>
<tr>
<th>Tid</th>
<th>Attrib1</th>
<th>Attrib2</th>
<th>Attrib3</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Large</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Medium</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Small</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Medium</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Large</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Medium</td>
<td>50K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Large</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Small</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Medium</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Small</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- Examples
  - Spam classifier
  - Medical image
  - Image classifier
  - OCR

- Trainset, testset, developer set

- Technique
  - Decision tree
  - kNN
  - Naive Bayes
  - SVM
  - Neural net

- Decision Tree
  - Attribute test for impurity
    - GINI index
    - Entropy

- kNN
  - Concept of Similarity
    - Euclidean distance
    - Scaling
    - Curse of dimensionality
    - Can produce counter-intuitive results
Cosine similarity

- Runtime Traintime performance

**Naive Bayes**
- Start with Dice example
- Tomorrow machine

https://xkcd.com/1132/

- Cancer diagnosis example

  - 1% of women have breast cancer (and therefore 99% do not).
  - 80% of mammograms detect breast cancer when it is there (and therefore 20% miss it). (Sensitivity = 80%)
  - 9.6% of mammograms detect breast cancer when it’s not there (and therefore 90.4% correctly return a negative result). (Specificity = 9.6%)

<table>
<thead>
<tr>
<th></th>
<th>Cancer (1%)</th>
<th>No Cancer (99%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Pos</td>
<td>80%</td>
<td>9.6%</td>
</tr>
<tr>
<td>Test Neg</td>
<td>20%</td>
<td>90.4%</td>
</tr>
</tbody>
</table>

- Bayes Theorem

  **Conditional Probability:**
  \[
P(C \mid A) = \frac{P(A, C)}{P(A)}
  \]

  \[
P(A \mid C) = \frac{P(A, C)}{P(C)}
  \]

  **Bayes theorem:**
  \[
P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}
  \]

- Bayesian Classifier
○ compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of $C$ using the Bayes theorem

$$P(C \mid A_1, A_2, \ldots, A_n) = \frac{P(A_1, A_2, \ldots, A_n \mid C) P(C)}{P(A_1, A_2, \ldots, A_n)}$$

○ Choose value of $C$ that maximizes $P(C \mid A_1, A_2, \ldots, A_n)$

○ Equivalent to choosing value of $C$ that maximizes $P(A_1, A_2, \ldots, A_n \mid C) P(C)$

○ Conditional Independence

$X, Y, Z$ denote three random variables.

$X$ is said to be conditionally independent of $Y$, given $Z$, if $P(X \mid Y, Z) = P(X \mid Z)$

Ex) $X$ : arm length

$Y$ : reading skill

$Z$ : age

○ Naive Bayes

Assume independence among attributes $A_i$ when class is given:

$P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C) P(A_2 \mid C) \ldots P(A_n \mid C)$

○ Smoothing

If one of the conditional probability is zero, then the entire expression becomes zero

Probability estimation:

Original: $P(A_i \mid C) = \frac{N_{ic}}{N_c}$
c: number of classes

Laplace: $P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$
p: prior probability

$m$-estimate: $P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$
m: parameter

○ Practice
<table>
<thead>
<tr>
<th>Name</th>
<th>Give Birth</th>
<th>Can Fly</th>
<th>Live in Water</th>
<th>Have Legs</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>python</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>non-mammals</td>
</tr>
<tr>
<td>salmon</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>non-mammals</td>
</tr>
<tr>
<td>whale</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>mammals</td>
</tr>
<tr>
<td>frog</td>
<td>no</td>
<td>no</td>
<td>sometimes</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>komodo</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>bat</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>pigeon</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>cat</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>leopard shark</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>non-mammals</td>
</tr>
<tr>
<td>turtle</td>
<td>no</td>
<td>no</td>
<td>sometimes</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>penguin</td>
<td>no</td>
<td>no</td>
<td>sometimes</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>porcupine</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>eel</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>non-mammals</td>
</tr>
<tr>
<td>salamander</td>
<td>no</td>
<td>no</td>
<td>sometimes</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>gila monster</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>platypus</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>mammals</td>
</tr>
<tr>
<td>owl</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
<tr>
<td>dolphin</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>non-mammals</td>
</tr>
<tr>
<td>eagle</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>non-mammals</td>
</tr>
</tbody>
</table>

- Runtime Traintime performance
Evaluation

- Accuracy
- Precision and Recall
- F-measure
  - F0.5 F2
- Mean average precision
- Train, Dev, Test set
- overfit
  - In Decision Tree
    - Using ID as an attribute
- underfit
  - Decision Tree for All the tax payer in the US
  - Can we achieve 100% accuracy?
- Limitation of the search evaluation
  - independence relevance
  - binary relevance
- SQL - Google - Visual Analytics
  - Known unknown
  - Unknown unknown
  - Can you detect anomaly using search?