Big Data Intelligence Project

To better understand Big Data Analysis (and Intelligence), the course project aims to enhance your intuitive “feeling” on various aspects of data analysis and expects you to apply what you have learned to practical problems.

Optional Topics

1. Online Competitions.

Specifically, we select some hot and interesting topics from online competition websites (Kaggle, AliTianchi, etc.). You are required to participate in these competitions, compete with all participants, and try to improve your rank based on what you have learned from this course. The optional topics are as follows (Specifically,**Green**/ **Black** /**Red** titles represents **Easy** / **Medium**/**Hard** topics for your reference):

Natural Language Processing with Disaster Tweets

https://www.kaggle.com/competitions/nlp-getting-started

Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they’re observing in real-time. Because of this, more agencies are interested in programatically monitoring Twitter (i.e. disaster relief organizations and news agencies). But, it’s not always clear whether a person’s words are actually announcing a disaster. Take this example: The author explicitly uses the word “ABLAZE” but means it metaphorically. This is clear to a human right away, especially with the visual aid. But it’s less clear to a machine.In this competition, you’re challenged to build a machine learning model that predicts which Tweets are about real disasters and which one’s aren’t.

House Prices - Advanced Regression Techniques

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

Ask a home buyer to describe their dream house, and they probably won't begin with the height of

the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Repeat Buyers Prediction-Challenge the Baseline

https://tianchi.aliyun.com/competition/entrance/231576?lang=en-us

Merchants sometimes run big promotions (e.g., discounts or cash coupons) on particular dates (e.g., Boxing-day Sales, "Black Friday" or "Double 11 (Nov 11th)" , in order to attract a large number of new buyers. Unfortunately, many of the attracted buyers are one-time deal hunters, and these promotions may have little long lasting impact on sales. To alleviate this problem, it is important for merchants to identify who can be converted into repeated buyers. By targeting on these potential loyal customers, merchants can greatly reduce the promotion cost and enhance the return on investment (ROI). It is well known that in the field of online advertising, customer targeting is extremely challenging, especially for fresh buyers. However, with the long-term user behavior log accumulated by Tmall.com, we may be able to solve this problem. In this challenge, we provide a set of merchants and their corresponding new buyers acquired during the promotion on the "Double 11" day. Your task is to predict which new buyers for given merchants will become loyal customers in the future. In other words, you need to predict the probability that these new buyers would purchase items from the same merchants again within 6 months. a data set containing around 200k users is given for training, while the other of similar size for testing. Similar to other competitions, you may extract any features, then perform training with additional tools. You need to only submit the prediction results for evaluation.

The Purchase and Redemption Forecasts-Challenge the Baseline

https://tianchi.aliyun.com/competition/entrance/231573

Ant Financial Services Group (AFSG) processes cash inflow and outflow for millions of its members. As one can imagine, predicting future cash flows based on historical data is an important part of

AFSG's business. Participants will be challenged to predict future cash flows based on users' historical purchase and redemption data to help Ant Financial Services Group (AFSG) improve its funds management abilities. (Purchases refers to funds inflow, while redemptions refers to funds outflow.)

Sina Weibo Interaction-prediction

https://tianchi.aliyun.com/competition/entrance/231574/introduction

Weibo is a Chinese microblogging website. It is one of the most popular sites in China, with a market penetration similar to the Twitter. “Weibo” is a Chinese word for “microblog”. User behaviors such as forwarding, commenting and liking are important factors that can be used to estimate the quality of a certain weibo and implement the recommendation and feed controlling strategy. In this competition, participants are required to predict the forwarding, commenting and liking amount of a weibo based on the historical interaction data.

I’m Something of a Painter Myself

https://www.kaggle.com/competitions/gan-getting-started

We recognize the works of artists through their unique style, such as color choices or brush strokes. The “je ne sais quoi” of artists like Claude Monet can now be imitated with algorithms thanks to generative adversarial networks (GANs). In this getting started competition, you will bring that style to your photos or recreate the style from scratch! Computer vision has advanced tremendously in recent years and GANs are now capable of mimicking objects in a very convincing way. But creating museum-worthy masterpieces is thought of to be, well, more art than science. So can (data) science, in the form of GANs, trick classifiers into believing you’ve created a true Monet? That’s the challenge you’ll take on!

Child Mind Institute — Problematic Internet Use

https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use/overview

In today’s digital age, problematic internet use among children and adolescents is a growing concern. Better understanding this issue is crucial for addressing mental health problems such as depression

and anxiety. Current methods for measuring problematic internet use in children and adolescents are often complex and require professional assessments. This creates access, cultural, and linguistic barriers for many families. Due to these limitations, problematic internet use is often not measured directly, but is instead associated with issues such as depression and anxiety in youth. Conversely, physical & fitness measures are extremely accessible and widely available with minimal intervention or clinical expertise. Changes in physical habits, such as poorer posture, irregular diet, and reduced physical activity, are common in excessive technology users. We propose using these easily obtainable physical fitness indicators as proxies for identifying problematic internet use, especially in contexts lacking clinical expertise or suitable assessment tools. This competition challenges you to develop a predictive model capable of analyzing children's physical activity data to detect early indicators of problematic internet and technology use. This will enable prompt interventions aimed at promoting healthier digital habits. Your work will contribute to a healthier, happier future where children are better equipped to navigate the digital landscape responsibly.

Google - Unlock Global Communication with Gemma

https://www.kaggle.com/competitions/gemma-language-tuning

With over 7,000 languages and countless cultural differences, AI has the potential to foster global understanding. In a step towards broader linguistic inclusion, we're launching a Kaggle competition focused on adapting Gemma 2, Google's open model family, for 73 eligible languages. These languages were selected to represent a diverse range and to align with the expertise of our judging panel for effective evaluation. Our initial focus on these languages will allow us to establish a robust foundation of techniques and resources that will later enable us to support under-resourced languages.

Jane Street Real-Time Market Data Forecasting

https://www.kaggle.com/competitions/jane-street-real-time-market-data-forecasting

When approaching modeling problems in modern financial markets, there are many reasons to believe that the problems you are trying to solve are impossible. Even if you put aside the beliefs

that the prices of financial instruments rationally reflect all available information, you'll have to grapple with time series and distributions that have properties you don't encounter in other sorts of modeling problems. Distributions can be famously fat-tailed, time series can be non-stationary, and data can generally fail to satisfy a lot of the underlying assumptions on which very successful statistical approaches rely. Layer on all of this the fact that the financial markets are ultimately a human endeavor involving a large number of individuals and institutions that are constantly changing with advances in technology and shifts in society, and responding to economic and geopolitical issues as they arise - and you can start to get a sense of the difficulties involved!

2. Project of Big Data Intelligence applied in specific discipline

As we introduced in the course, how to integrate big data analysis with a specific discipline is also a very important topic. Therefore, students not from the department of information science can choose the combination of big data analytics and a specific domain application as a project. For this purpose, the students who choose this topic should first submit a Project Proposal with the following contents (no more than 3 A4 pages).

2.1 Problem definition.

2.1.1 Please clearly define the problem to be solved.

2.1.2 Please explain how the problem is related to Big Data/Data Analytics.

2.1.3 Please indicate the data you will use (data source, data content, data size, etc.).

2.2 Related work.

2.2.1 Please briefly summarize and describe the relevant work for this project.

2.2.2 Please analyze the similarities and differences between the problem you will be working on and the related work, as well as the innovation points.

2.3 Solution.

2.3.1 Please briefly explain the solution ideas, algorithms, models, etc. that you plan to use.

2.3.2 Please briefly explain how you plan to evaluate and test your solution.

The teaching assistant will judge and give feedback based on the project proposal, and contact the students to confirm the chosen topic.

Note: The project can be a topic you are currently working on in the lab, but you need to clearly explain how it relates to big data/data analytics, where it fits in, etc.

3. Research survey on literature related to big data intelligence

Due to the time limit of the course, it is not possible to cover all the frontiers of Big Data Intelligence. You can also choose a frontier area of big data analytics to conduct literature research survey as a major assignment with the following requirements.

1. Your survey should include no less than three papers you like in the field, but they must be of high quality in the field, such as seminal papers, highly cited papers, test cutting-edge papers (e.g. published in well-known journals and magazines or computer-related fields), etc.

2. Some computer experiments should be conducted based on the research results, including but not limited to reproducing the experiments in the paper, finding/constructing the data set by yourself for comparison, testing different algorithms, etc.

3. Optional topics includes:1) hashing; 2) network embedding; 3) graph neural networks;

4) multimodal analysis; 5) multi-view learning/multi-task learning; 6) recommendation;

7) transfer learning; 8) social network analysis; 9) urban computing; 10) sequence mining/time series analysis; 11) causal inference 12) data privacy; 13) reinforcement learning; 14) generative adversarial networks; 15) meta-learning; 16) continual learning; 17) AutoML/neural architecture search/hyper-parameter optimization; 18) adversarial attacks and defenses/robustness; 19) self-supervised learning; 20) big data analysis+ covid-19 related. 21) Diffusion Models for AIGC; 22) Large language models;

Note that: in order to avoid duplication, each topic above can only be chosen by 1 group. If you find that one team has already chosen the topic before you, you need to change your topic. Fill the table

we give with your topic information.

Some suggestions from TA: If you choose this survey topic, we expect that you should at least include the following contents in your presentation:

a. The background, basic concepts, and the importance of the surveyed topic.

b. Current mainstream techniques of the surveyed topic, how they relate to the topic.

c. Representative works under this topic.

d. Comparison between the different methods and your thinking about this topic.

Project requirements

1) The assignment is a free team assignment with best 3 members per team in principle (if there are special circumstances, please contact the TA. A team with 2~4 members is acceptable, but make sure that everyone could have adequate work for the presentation.).

2) For the group information, please fill out the team leader, team member information (name, student number) and the selected topic in the document in the WeChat group by **October 29th**. If you choose Project of Big Data Intelligence applied in specific discipline, the project proposal should be uploaded to the web learning assignment submission window before **October. 29th**.

3) About the topic selection: students from the department of Information Technology are required to choose only data competition (Note: a student from the department of Information Technology in the team is counted as the department of Information Technology), while non-department of Information Technology students can choose data competition or application. At the same time, please consciously ensure that the chosen topic is not used as other class assignments (i.e., you cannot participate in the same competition and use it as an assignment in two different classes).

4) Assessment method: We will assess the major assignments through presentations and reports (the grade accounts for 40% of the overall course evaluation). The presentation is tentatively scheduled for **14th-16th** weeks**(Dec.10th to Dec.24th)**. The scores of this assignment will take into

account the results of the presentation, report, and competition (if you choose data competition), and the specific details will be announced later.

Note: We encourage you to think deeply, explore more, experiment and innovate in the project.

5) For students participating in the competition, since different competitions are divided into different stages and have different deadlines, we only ask you to submit the results at the deadline of our assignment. For example, if a competition is over at the time of the ddl (deadline) presentation/report, you need to submit the final result; if a competition is still in progress, you only need to submit the result and ranking before the ddl.

6) Please start as early as possible because of the limited time.

FAQ

Q: How to balance the scores of students from different departments and projects?

A: We will make a comprehensive evaluation based on the presentation, report, and competition results (if you choose data competition), taking into account the students' backgrounds and the difficulty of the competition to ensure fairness as much as possible. The presentation score is composed of your group score and your personal score, where the group score will consider the ranking and innovation of your method, and the personal score will consider how your perform in the presentation, how much workload you have in the project, and how your perform in the QA stage, so it is better for each of you to present the part that you are responsible for in the project.

Q: I have no previous experience in data competitions, will I get a low score?

A: As mentioned in the assignment requirements, we encourage you to think deeply, explore, experiment and innovate more, and do not want students who participate in the competition to just "ranking high". Therefore, we will pay more attention to the process of the project rather

than just the final result of the competition. As long as you are willing to put your effort, many students who participated in the competition for the first time in previous years also achieved good scores. On the contrary, some experienced students who took their original models from similar competitions and brought them directly without any modification or new attempts, even though the competition results were okay, they did not get high marks in the course.

Q: Can you provide computing resources for the course?

A: At this moment, we do not have any computational resources available, so we hope you can consider this issue when you choose a topic. If you really have difficulties, you can contact the teaching assistant.