THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Topics for STAT4798 Statistics and Actuarial Science Project (6 credits)
(Offered in both 1st and 2nd semesters of 2024 - 2025 for STAT4798)

1. Natural hedging between longevity and mortality risk

A life insurer faces longevity risk, which arises from a systematic improvement in future life expectancies and the subsequent increase in the actuarial value of their annuity liabilities. On the other hand, the life insurer may face mortality risk, which arises from a systematic decrease in future life expectancies and the subsequent increase in the actuarial value of their death benefit (life insurance) liabilities. To manage those two risks, the life insurer may seek an optimal balance between annuities and death benefit insurances. This project seeks to find an optimal mix. To obtain this mix, an insurance pricing strategy can be analyzed, or risk-sharing arrangements can be sought with an appropriate counterparty. The data is available open source, and the literature will be provided. Students are expected to have good knowledge in programming languages such as R.

References:

- Cox, S. H., & Lin, Y. (2005). Natural hedging of life and annuity morality risks. Journal of Risk and Insurance , 11, 1-15.

Supervisor: **Prof. T.J. Boonen,** tjboonen@hku.hk, Dept of Statistics and Actuarial Science

2. Privacy Protection in Machine Learning

In recent years, machine learning has become an important tool for data analysis in various domains, including healthcare, finance, and marketing. However, as ML models rely on large amounts of data, there is a growing concern about the privacy risks associated with the collection, storage, and processing of sensitive data. In this project, students will explore different techniques and tools for privacy-preserving machine learning, which aim to enable data analysis while protecting the privacy of individuals and organizations.

The target students are senior undergraduate students with a strong background in deep learning and python (PyTorch/TensorFlow) programming.

Supervisor: **Prof. Y. Cao,** yuancao@hku.hk, Dept of Statistics and Actuarial Science

3. Dependence Structures in Multiple Life Insurances and Annuities

The price of a multiple-life insurance/annuity product depends not only on the marginal distributions of the underlying future lifetimes, but also on their dependence structure. In this project, the effect of dependence structure on the actuarial present values will be studied. In the course of the research, the student will learn some basic theory of dependence structures.

Supervisor: **Prof. K.C. Cheung**, kccg@hku.hk, Dept of Statistics and Actuarial Science

4. Stock market forecasting and stock investment in practice

The stock market is known for its inherent volatility and complexity, making it a challenging environment to predict with certainty. To become a successful stock investor, one needs to have the macro level understanding of the stock and financial markets, their trends and movements as well as the micro level understanding of individual stocks, related businesses and accounting measures. This project aims to provide students with hands-on virtual experiences of stock investment. Students are expected to build the following three types of models. First, based on the macro level information only, the macro model that predicts the overall market movements and produces clear buying and selling signals in light of the long term market movement cycle. Second, based on the principles of value investment and the development cycle of the industries, the portfolio selection models that design combinations of stocks suitable for the purposes of short-, medium- and long-term investments. Third, trading models that provide guidance for daily, weekly, monthly and quarterly buying and selling. In- and out-of-sample validations should be conducted to verify the quality of these models, and real-world virtual trading of no less than one month should be conducted to test the profitability of the daily/weekly, and monthly trading strategies.

Pre-requisite knowledge: STAT1603/STAT2601, STAT2602, STAT3600, STAT4601
Software required: R or Python, Excel

Supervisor: **Mr. Harrison Y.Y. Cheung**, hcheung4@hku.hk, Dept of Statistics and Actuarial Science

5. Queuing Theory

The study of waiting lines, called queuing theory, is one of the oldest and most widely used quantitative analysis techniques. The analytical models of waiting lines can help management evaluate the cost and effectiveness of service systems.

In this project, students will conduct a case study and explore the queuing systems with various statistical models, and apply simulation models to enhance the efficiency of the queuing system.

Supervisor: **Dr. Olivia Choi**, ochoi@hku.hk, Dept of Statistics and Actuarial Science

6. Project Management Analysis

Most realistic projects that corporates/organizations like Microsoft, General Motors undertake are large and complex. Project management is a process that allows project managers to plan, execute, track and complete projects effectively.

In this project, students will conduct a case study and evaluate the efficiency of the selected project by identifying the crucial activities, developing the Work Breakdown Structure and critical path, and hence provide recommendation for the management through the analysis of the Program evaluation and review technique (PERT).

Supervisor: **Dr. Olivia Choi**, ochoi@hku.hk, Dept of Statistics and Actuarial Science

7. Financial markets co-integration analysis

Financial market microstructure theory, a new research branch in financial market analysis, attracted attention from researchers (Busato and Handorf, 2001; Nagel 2016).

Over the years, international cross-listing has arisen a great deal of academic focus particularly on the segmented markets theory and the failure of the law of one price in multiple markets. The existence of long term co-movements of cross-listing company shares performance has importance implication on the Efficient Market Hypothesis as well as portfolio diversification (Taylor & Tonks, 1989; Kasa 1992).

Given the close economic and rapid growth relationships between many major financial markets, students are expected to look into market co-integration and shares performance of these dually listed companies.

Supervisor: **Dr. Olivia Choi**, ochoi@hku.hk, Dept of Statistics and Actuarial Science

8. Matrix Completion with Application to Recommender System

Developing efficient recommender system to track users' preference and provide tailored recommendations are becoming increasingly important in modern society. This project focuses on using matrix completion techniques to construct a recommender system, with the MovieLens dataset as the primary data source. Students are expected to have good knowledge in programming languages such as Python or R.

Supervisor: **Prof. L. Feng**, lfeng@hku.hk, Dept of Statistics and Actuarial Science

9. Approximate Inference for Bayesian Models

Markov chain Monte Carlo (MCMC) methods are considered the gold standard for inference in Bayesian models. However, in modern settings like machine learning, large datasets and high-dimensional models have become the norm. This presents a challenge to MCMC, as it is inherently serial and computational demanding. As a result, alternative scalable approximate methods for Bayesian inference are being developed. These include variational Bayes, expectation propagation, Laplace's approximation, the Bayesian bootstrap, and others. The aim of the project is to investigate the application of these methods in complex settings and evaluate their respective merits and weaknesses.

Requirement: Experience in Python or R; familiar with Bayesian inference

Supervisor: **Prof. Edwin C.H. Fong**, chefong@hku.hk, Dept of Statistics and Actuarial Science

10. Transfer Learning for Survival Analysis

Description: Transfer learning has gained considerable attention in machine learning and data science. By leveraging knowledge from related datasets, transfer learning can overcome the limitations of small sample sizes and/or short study durations, and improve learning performance on target data. Survival analysis is a powerful tool for risk prediction and identifying high-risk subjects, allowing for personalized interventions and improved patient outcomes. In this project, students will study transfer learning methods for survival analysis and apply them to real-world breast cancer genomic studies, aiming to develop a more accurate risk prediction model for the target population. Students are expected to have basic knowledge of survival analysis and be familiar with C++/R/Python.

Supervisor: **Prof. Y. Gu**, yugu@hku.hk, Dept of Statistics and Actuarial Science

11. Open-world object discovery with deep learning

Deep learning has achieved remarkable success in many tasks, even surpassing humans, for example in image classification. However, the success comes at the cost of intensively labeled data, e.g., ImageNet which contains over 1.2 million manually annotated images. When a trained classification model meets an image from an unseen class, it often mistakenly predicts the image as one of the seen classes with high confidence. In other words, current learning models struggle to handle open-world problems where there are unseen or unfamiliar objects. In this project, the students will study the open-world object discovery problem with deep learning and develop solutions to enable the model to deal with unseen or unfamiliar objects.

Requirement: Knowledge and hands-on experience in computer vision and deep learning; familiar with Python; preferably also familiar with PyTorch/TensorFlow/JAX.

Supervisor: **Prof. K. Han**, kaihanx@hku.hk, Dept of Statistics and Actuarial Science

12. Content Generation with Diffusion Models

Diffusion models have shown promising results in visual content creation, driving the intriguing development of new generation image generation platforms such as Stable Diffusion and Midjourney. Though encouraging advancements have been achieved, the full potential of diffusion models is yet to be discovered. Despite generating visually appealing images, diffusion models can also be applied to generate other types of content, such as 3D models and videos.

In this project, students will study and explore diffusion models for different content generation tasks, showcasing their versatility and potential for different types of content creation.

Requirement: Knowledge and hands-on experience in computer vision and deep learning; familiar with Python; preferably also familiar with PyTorch/TensorFlow/JAX.

Supervisor: **Prof. K. Han**, kaihanx@hku.hk, Dept of Statistics and Actuarial Science

13. Efficient Adaptation of Foundation Models

Many foundation models have been made available, such as CLIP, SAM and the GPT family. The training of such models is immensely expensive, necessitating industry-level computing power and Internet-scale training data. Therefore, it is intriguing to investigate their capabilities without requiring retraining, but with efficient adaptation that allows them to accomplish other tasks they were not trained on. In this project, students will examine existing foundation models, showcase their applications, and explore efficient adaptation methods for these models.

Requirement: Knowledge and hands-on experience in computer vision and deep learning; familiar with Python; preferably also familiar with PyTorch/TensorFlow/JAX.

Supervisor: **Prof. K. Han**, kaihanx@hku.hk, Dept of Statistics and Actuarial Science

14. Implicit Neural Representations

With the advance of deep learning in computer vision, implicit neural representations appear to be a novel way to parameterize all kinds of signals. Unlike the conventional discrete signal representations (e.g., images are discrete grids of pixels, 3D shapes are often discrete grids of voxels/point clouds/meshes, and audio signals are discrete samples of amplitudes), implicit neural representations parameterize a signal as a continuous function that maps the domain of the signal (i.e., a coordinate, such as a pixel coordinate for an image) to whatever is at that coordinate (for an image, an R, G, B color). In this project, students will study and develop methods for using implicit neural representations to process visual information, such as images and 3D shapes, with potential applications including image super-resolution and 3D shape reconstruction.

Requirement: Knowledge and hands-on experience in computer vision and deep learning; familiar with Python; preferably also familiar with PyTorch/TensorFlow/JAX.

Supervisor: **Prof. K. Han**, kaihanx@hku.hk, Dept of Statistics and Actuarial Science

15. Random matrices under dependence

Random matrix theory typically assumes the random entries of a high-dimensional, square matrix to be iid. Of interest is then the empirical distribution of all eigenvalues of this random matrix, with applications to estimated covariance matrices. The goal of this project is to present various ways of introducing dependencies in random matrices and to numerically investigate their influence on the distribution of eigenvalues (of the random matrices or their corresponding covariance matrices).

Requirement: Knowledge in R, LaTeX

Supervisor: **Prof. M. Hofert**, mhofert@hku.hk, Dept of Statistics and Actuarial Science

16. Comparative Analysis of Various Deep Learning Models for Latent Structure Exploration: Simulated and Real Data Evaluation

This project aims to compare the effectiveness of various techniques for exploring latent structures in data: factor analysis and deep learning models. We will use both simulated and real datasets to evaluate the performance of exploratory and confirmatory factor analysis as well as deep learning models. The study will focus on exploring various factor structures to identify the most suitable method for extracting meaningful structures from the data. The results of this project will provide valuable insights into the strengths and limitations of these methods and their potential applications in various fields, including psychology, economics, and marketing.

Requirement: Knowledge in Python
Knowledge in factor analysis and deep learning.

Supervisor: **Dr. C.W. Kwan**, cwkwan@hku.hk, Dept of Statistics and Actuarial Science

17. Analysis of correlated zero-inflated count data

In many medical and public health investigations, the count data encountered often exhibit an excess of zeros, and very frequently this type of data are collected on clusters of subjects or by repeated measurements on each subject. For example, in the analysis of medical expenditure, members in the same family may exhibit some correlation possibly due to housing locality, genetic predisposition, similar dietary and living habit. Ignoring such correlation may lead to misleading statistical inference. This project will survey the models and methods in the literature and apply them to a real data set.

Requirement: Knowledge in R or Python.

Supervisor: **Prof. Eddy K.F. Lam**, hrntlkf@hku.hk, Dept of Statistics and Actuarial Science

18. ESG ontology (Company Project)

This project aims to develop an ESG ontology. Students will learn how to implement an ontology (a taxonomy of words with semantic meaning), and apply data science, social media and AI models to create the ontology. Students will learn different open source AI tools for text analysis and ontology building, and learn how to develop an ontology database. Students who have basic knowledge in statistics, AI, machine learning, text analysis are preferred, and have a minor in computer science and finance background will take an advantage.

Supervisor: **Dr. Adela S.M. Lau**, adelalau@hku.hk, Dept of Statistics and Actuarial Science

19. Building an ontology-based AI chatbot (Company Project)

This project aims to develop an AI chatbot. Students will learn how to implement an ontology (a taxonomy of words with semantic meaning), and apply different language models to create a chatbot. Students will learn different open source AI tools for NLP and text analysis, and learn how to develop a knowledge base. Students who have basic knowledge in statistics, AI, machine learning, text analysis are preferred, and have a minor in computer science will take an advantage.

Supervisor: **Dr. Adela S.M. Lau**, adelalau@hku.hk, Dept of Statistics and Actuarial Science

20. Emotion and Mind Detection using EEG device (Company Project)

This project aims to develop an AI program to detect human emotion and the mind. Students will evaluate the existing EEG devices and develop an AI program for human emotion and mind detection. Students will learn different EEG technologies and open source AI tools for innovative application. Students who have basic knowledge in statistics, AI, machine learning, and blockchain technologies are preferred, and have a minor in computer science and/or neurocognitive science knowledge will take an advantage.

Supervisor: **Dr. Adela S.M. Lau**, adelalau@hku.hk, Dept of Statistics and Actuarial Science

21. A 3D scene reconstruction (Company Project)

This project aims to construct a 3D model from a 360° Panoramic View and an image file. Students will review existing open source software to implement a 3D model for a scene and objects. Students will learn various AI open source software for 3D objects and the scene reconstruction. If it is a group project format, students will also learn the IoT devices and software development in unity. Students who have basic knowledge in statistics, AI, machine learning, and 3D generation technologies are preferred, and have a minor in computer science will take an advantage.

Supervisor: **Dr. Adela S.M. Lau**, adelalau@hku.hk, Dept of Statistics and Actuarial Science

22. Video Analytics (Company Project)

This project aims to analyze the video for detect a set of speech, movement and emotion. Students will review existing open source software to implement for object detection and develop AI models and algorithms to analyze the scene. Students will learn various AI open source software for object detection, pose detection, and text analysis. Students who have basic knowledge in statistics, AI, and machine learning background are preferred, and have a minor in computer science will take an advantage.

Supervisor: **Dr. Adela S.M. Lau**, adelalau@hku.hk, Dept of Statistics and Actuarial Science

23. Applications of Extreme Value Models

Extreme value theory concerns the behaviour of maxima or minima, and has been used extensively in areas such as finance, hydrology, engineering and meteorology where the occurrence of extremes may have catastrophic consequences. In this project, the student will learn the basic modelling techniques for data of extremes and will apply such models to data sets of practical interest. The emphasis is on conceptual understanding of the underlying theory and interpretation of the fitted models.

Requirement: The student should be competent in computer programming. Knowledge in or willingness to learn the R programming language is essential.

Supervisor: **Dr. David Lee**, leedav@hku.hk, Dept of Statistics and Actuarial Science

24. Resampling Methods for Regression

Recent years have found increasing use of resampling methods in regression studies. Examples include the paired bootstrap, the residual bootstrap, the wild bootstrap, random perturbation, bagging, etc. In this project we explore their potential applications in contemporary regression settings where statistical inference remains prohibitively difficult.

Supervisor: **Prof. Stephen M.S. Lee**, smslee@hku.hk, Dept of Statistics and Actuarial Science

25. Applications of Secure Blockchain Solution

In this project we begin with a review of the basic architecture for blockchain in Python. This includes state transition rules, method for creating blocks, mechanisms for checking the validity of transactions, blocks, and the full chain. Next, we will create new blocks from data, validate the new blocks and add them to the existing blockchain.

Security is of the utmost importance in any blockchain architecture, in this project we will discuss 3 popular verification methods: public key cryptography, digital signature algorithm and trusted time-stamping. Finally, we will construct practical blockchain solutions to current fintech problems.

Supervisor: **Dr. Eric A.L. Li**, ericli11@hku.hk, Dept of Statistics and Actuarial Science

26. Introduction to Quantum Computing Algorithms

First we begin with a basic understanding of quantum computing (QC). Then we move on to some popular QC algorithms, written in Javascript and Python. In addition to constructing these QC codes, we will also provide the meanings, purposes and theoretical bases of these QC codes.

The QC algorithms we will cover include: Deutsch-Jozsa Algorithm, Simon's Algorithm, Super Dense Coding, Period Finding, and Shor's Factoring Algorithm. The last one is particularly important in modern cryptography: given an integer which is a product of two distinct prime numbers, this algorithm finds one of its prime factors.

Supervisor: **Dr. Eric A.L. Li**, ericli11@hku.hk, Dept of Statistics and Actuarial Science

27. Statistical Inference for Tensor Data

Tensors have been used in many fields and have provided powerful applications in various practical domains. They generalize vectors and matrices and have been studied from different viewpoints. The study of tensor methods has a long history in statistics. In the era of big data, tensor data appear frequently in the forms of video data, spatio-temporal expression data, relationship data in recommending and mining, and latent variable models, from a vast range of statistical applications. However, the extension of methods for dealing with matrices to tensors is much more difficult than those from vectors to matrices. This project targets to several tensor-based statistical methods.

Supervisor: **Prof. G. Li**, gdli@hku.hk, Dept of Statistics and Actuarial Science

28. Deep Learning Approach for Stochastic Control Problems

A stochastic optimal control problem deals with uncertainties when making decisions to maximize or minimize an objective function. It is widely used in deriving the optimal trading strategy in the financial field and the optimal insurance strategy in actuarial science. However, the ``curse of dimensionality'' can quickly rise when solving a high dimensional stochastic control problem (e.g., a portfolio with a bunch of stocks, bonds, and insurances). Although no rigorous proof exists, some studies show that the deep learning approach can effectively reduce the ``curse of dimensionality'' phenomenon.

How to use a neural network to compute the optimal trading and insurance strategies for the high dimensional stochastic control problem? This is a promising direction worth of study.

You may need the following theories and techniques to conduct this research:

1. The basic theories of Mathematical Finance to model a portfolio optimization problem (like Financial Economics I & II).
2. Monte Carlo approach to simulate stochastic market scenarios.
3. Neural network algorithm to maximize or minimize an objective function.

Supervisor: **Prof. W. Li**, wylsaas@hku.hk, Dept of Statistics and Actuarial Science

29. Privacy preservation for federated learning in healthcare

Artificial intelligence (AI) approaches have shown great promise for augmenting clinical workflows. However, access to large quantities of diverse training data is needed to develop robust models. Notably, sharing data across institutions is not always feasible due to security and privacy concerns. As such, Federated Learning (FL) approaches allow for multi-institutional training of deep learning models without the need to share data. However, FL comes with security and privacy concerns as well. Specifically, the data insights exchanged during FL training can leak information about institutional data. In addition, the collaborative nature of the FL workflow can introduce new issues when there is a lack of trust among the entities performing the distributed compute.
In this project, the students will study the current privacy threats and associated threat mitigations for FL workflows. Students are also encouraged to design new and robust privacy preserving models for FL in healthcare.

Requirement: knowledge in machine learning/deep learning, proficient in python (PyTorch/TensorFlow) programming.

Supervisor: **Prof. L. Qu,** liangqqu@hku.hk, Dept of Statistics and Actuarial Science

30. Diffusion models for medical image restoration and synthesis

Medical imaging is an essential element for biomedical research and has demonstrated tremendous success in a wide range of areas, such as disease diagnosis, monitoring, or treatment. However, most existing medical imaging equipment is often cost-prohibitive and not always accessible in clinic. Thus, it is it is crucial to develop methods to reconstruct/synthesize high-quality medical images from low-cost, facilitating doctors with high diagnostic image quality for diagnostic decision. Recently, Denoising Diffusion Probabilistic Models have achieved remarkable success in various image generation tasks compared with Generative Adversarial Nets (GANs). In this project, the students will study and explore the diffusion models and apply it to medical image restoration and synthesis.

Requirement: knowledge in machine learning/deep learning, proficient in python (PyTorch/TensorFlow) programming.

Supervisor: **Prof. L. Qu,** liangqqu@hku.hk, Dept of Statistics and Actuarial Science

31. Multimodal large language model for medical image analysis

Medical image analysis is a critical component in healthcare, contributing substantially to accurate diagnoses and effective treatment decisions. Despite its importance, the analysis of medical images presents considerable challenges due to their inherent complexity and variability. Large language models, such as ChatGPT and Qwen, have made significant strides in natural language processing tasks by mastering the contextual representation of words and sentences. However, their application in the domain of medical image analysis is still largely unexplored, with existing multimodal medical image analysis methods predominantly based on traditional general AI strategies.

This project offers students an opportunity to delve into the development and application of a multimodal large language model specifically for medical image analysis. This includes tasks like medical image segmentation, report generation, visual question answering, among others. Students will be exposed to the latest techniques in natural language processing and medical image analysis, fostering a deeper understanding of how to integrate these cutting-edge language processing techniques to enhance medical image analysis.

This project is particularly well-suited for students interested in exploring the intersection of AI and healthcare, providing them with practical experience in applying AI techniques in a critical and rapidly evolving field.

References:

Tong S, Brown E, Wu P, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. arXiv 2024

Zhang S, Xu Y, Usuyama N, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv 2024

Requirement: knowledge in machine learning/deep learning, proficient in python (PyTorch/TensorFlow) programming.

Supervisor: **Prof. L. Qu,** liangqqu@hku.hk, Dept of Statistics and Actuarial Science

32. Tackling data heterogeneity challenge in federated learning

Federated learning is an emerging research paradigm enabling collaborative training of machine learning models among different organizations while keeping data private at each institution. Despite recent progress, there remain fundamental challenges such as non-convergence and the risk of catastrophic forgetting, particularly when dealing with real-world heterogeneous devices and non-IID (independent and identically distributed) data. In this project, students will have the opportunity to study the impact of data heterogeneity on federated learning performance. They will explore various techniques and strategies to mitigate the negative effects of non-IID data on model convergence and learning. Moreover, students are encouraged to design new and robust federated learning algorithms that can effectively tackle the challenges posed by non-IID data distribution across participating organizations. By engaging in this project, students will gain valuable insights into the complexities of federated learning and develop critical skills in designing and implementing advanced machine learning solutions in real-world, heterogeneous environments.

Requirement: knowledge in machine learning/deep learning, proficient in python (PyTorch/TensorFlow) programming.

Supervisor: **Prof. L. Qu,** liangqqu@hku.hk, Dept of Statistics and Actuarial Science

33. Cointegration in Financial Analysis

The goal of this project is to test cointegration in financial time series. Students are required to have basic understanding of cointegration and some knowledge of computer programming.

Supervisor: **Prof. C. Wang**, stacw@hku.hk, Dept of Statistics and Actuarial Science

34. A Simulation Study on an Extended Optimal Approach to Bühlmann Credibility Theory

In the recent paper by Yan and Song (2022), the classical Bühlmann credibility theory was extended to include non-linear Bayesian credibility estimators. Various numerical examples and simulation studies were performed in their work.

This project aims to study the benefits and effectiveness of this framework by establishing further examples and numerical studies. Monte Carlo experiments are to be performed to assess the performance of the proposed method in finite sample cases. Emphasis can be put on the examples for heavy-tailed excess claims distributions. Students taking this project are expected to have fundamental knowledge in credibility theory and programming skills for extensive simulation study.

Requirement: Pass in STAT3908, or equivalent

References:

- Yan, Y. and Song, K.-S. (2022). A General Optimal Approach to Bühlmann Credibility Theory. *Insurance: Mathematics and Economics*, 104, 262-282. (full paper accessible online from HKUL by HKU students)
- Bühlmann, H. and Gisler, A. (2005). *A Course in Credibility Theory and its Applications*. Springer.

Supervisor: **Dr. K.P. Wat**, watkp@hku.hk, Dept of Statistics and Actuarial Science

35. A/B testing

A/B testing has been growing immensely in online platform industry, such as Amazon, Google, Meta, Taobao, Wechat, Baidu, etc.
It is an online randomized experiment to evaluate which strategy is the best. There are many new challenges in online experiments due to fast recruitment, large sample size, multiple tests as well as multi-arm bandit problems. This project will explore Bayesian approaches to enhancing efficiency of A/B tests while controlling the type I error rate.

Supervisor: **Prof. G. Yin**, gyin@hku.hk, Dept of Statistics and Actuarial Science

36. Large Language Models (LLMs) for Healthcare Applications

Recent advancements in Large Language Models (LLMs) have significantly revolutionized natural language processing tasks. Owing to the massive training text data and billions of model parameters, LLMs have demonstrated a strong ability for general-purpose language understanding, generation, and question answering. In the healthcare domain, there is a vast amount of text data that records critical medical conditions and complex relationships. Thus, exploring how to leverage the knowledge from LLMs to solve healthcare tasks is of great interest and importance. For example, LLMs can be applied in medical image report generation, clinical decision prediction, etc.

This project will study and explore the applications of large language models in healthcare tasks, such as medical image analysis, electronic health record analysis.

Requirement: The student needs to have experience with Python programming and be familiar with basic machine learning/deep learning.

Supervisor: **Prof. L. Yu**, lqyu@hku.hk, Dept of Statistics and Actuarial Science

37. Generative AI with Applications in Healthcare

Generative AI models, such as diffusion models and autoregressive models, have shown remarkable capabilities in synthesizing high-quality data across various domains. In healthcare, these models can be leveraged to generate synthetic medical data, which can be used for training other AI models, augmenting datasets, and even simulating rare medical conditions. This can significantly enhance the robustness and generalizability of AI applications in healthcare.
This project will study and explore generative AI models, and demonstrate their applications in the healthcare domain by synthesizing medical images, text, and genomic data.

Requirement: The student needs to have experience with Python programming and be familiar with basic machine learning/deep learning and some techniques including diffusion models and decoder-only transformers.

Supervisor: **Prof. L. Yu**, lqyu@hku.hk, Dept of Statistics and Actuarial Science

38. Multimodal AI with applications in healthcare

Most of the current applications of AI in medicine have addressed narrowly defined tasks using one data modality, such as a computed tomography (CT) scan or retinal photograph. In contrast, clinicians process data from multiple sources and modalities when diagnosing, making prognostic evaluations and deciding on treatment plans. The development of multimodal AI models that incorporate data across modalities such as medical images, EHRs, and genomic data can partially bridge this gap and enable broad applications in healthcare.

This project will study and explore multimodal AI models and demonstrate its applications in healthcare domain by analysing image, text, or even genomic data.

Requirement: The student needs to have experience with Python programming and be familiar with basic machine learning/deep learning.

Supervisor: **Prof. L. Yu**, lqyu@hku.hk, Dept of Statistics and Actuarial Science

39. Causality in Healthcare

Current learning approaches primarily rely in the statistical correlations of the covariates and the response. The reliance limits their generalization when higher-order recognition is necessitated. A foundational approach to overcome these limitations involves incorporating principles of causality into the development of machine learning algorithms. Developing causality-aware algorithms is crucial in healthcare as it highlights the genuine factors and rules out the noisy variables in clinical decision making.

This project aims to develop causal machine learning algorithms in the healthcare domain (e.g., medical images and electronic health records), which are not only high-performant, but also interpretable and generalizable.

Requirement: The student needs to have experience with Python programming and be familiar with basic machine learning/deep learning and statistics.

Supervisor: **Prof. L. Yu**, lqyu@hku.hk, Dept of Statistics and Actuarial Science

40. Bayesian Change Point Detection in Financial Time Series

Time series data are commonly observed in the real world, of which the patterns and trends are of great interest, especially in the financial industry. Fluctuations are frequently observed in financial time series data. Statistical approaches to locate abrupt variations driven by changes in policy, event, and market sentiment have raised great concerns. In this project, students will study various Bayesian change point detection algorithms and learn how to implement those techniques in real financial time series data.

Requirement: Knowledge in R or Python

Supervisor: **Dr. C. Zhang**, zhangcys@hku.hk, Dept of Statistics and Actuarial Science

41. Phase II Clinical Trial Design with Time-to-event Outcomes

Clinical trial design plays a crucial role in drug development, with the primary objective being to establish the effect of the investigated intervention. Following the assessment of safety and toxicity in Phase I trials, Phase II trial focuses on the effectiveness of the intervention for patients under specific conditions. In this project, students will learn and develop Phase II clinical trial designs with time-to-event outcomes. Students with fundamental knowledge in biostatistics are preferred.

Requirement: Knowledge in biostatistics and R programming

Supervisor: **Dr. C. Zhang**, zhangcys@hku.hk, Dept of Statistics and Actuarial Science

42. Statistical Modelling for Biological/Medical Data

In this project, the students will implement statistical methods to analyse real biological/medical data set to understand/interpret biology/disease etiology. Statical methods include Bayesian methods, variable selection, network analysis, etc.

Requirement: Students need to know at least one programming language (such as R, Python, etc) and basic data analysis skills.

Supervisor: **Prof. Dora Y. Zhang**, doraz@hku.hk, Dept of Statistics and Actuarial Science

43. Multiple Output Online Non-stationary GPs

The goal of this project is to implement an online algorithm for multiple output Gaussian processes. The student will extend a Sequential Monte Carlo sampler for online Gaussian processes by writing a linear co-regionalization kernel to model multiple time series signals. Possible applications include medical settings or financial settings. Strong programming ability in Python and prior experience in Bayesian inference is required.

Supervisor: **Prof. Michael M.Y. Zhang,** mzhang18@hku.hk, Dept of Statistics and Actuarial Science

44. Online Spectral Mixture Kernel

The goal of this project is to implement a method to estimate the parameters in the flexible "Spectral Mixture Kernel" in an online setting using a Sequential Monte Carlo algorithm. Applications of this method include medical or financial settings. Strong programming ability in Python and prior experience in Bayesian inference is required.

Supervisor: **Prof. Michael M.Y. Zhang,** mzhang18@hku.hk, Dept of Statistics and Actuarial Science

45. Online Student-t Process Algorithm

The goal of this project is to implement an online inference algorithm to learn a heavy tailed Student-t process for time series analysis. Strong programming ability in Python and prior experience in Bayesian inference is required.

Supervisor: **Prof. Michael M.Y. Zhang,** mzhang18@hku.hk, Dept of Statistics and Actuarial Science

46. Non-linear Network Embedding

The goal of this project is to model relational data as a non-linear decomposition of a lower dimensional representation of the relations between observations. Strong programming ability in Python and prior experience in Bayesian inference is required.

Supervisor: **Prof. Michael M.Y. Zhang,** mzhang18@hku.hk, Dept of Statistics and Actuarial Science

47. A Bayesian Hypothesis Testing Approach for Generative Adversarial Networks

This project involves combining the popular Generative Adversarial Network with various forms of Bayesian hypothesis testing. If successful, the Bayesian hypothesis testing GAN could have stronger classification abilities and could possible reduce the risk of mode collapse. Prior knowledge of deep learning and strong programming ability in Python and deep learning packages like PyTorch, Tensorflow or Keras are required.

Supervisor: **Prof. Michael M.Y. Zhang,** mzhang18@hku.hk, Dept of Statistics and Actuarial Science

48. Forecasting Time Series: with Application to Stocks Trading

This project aims to forecast forward behavior of stock prices using neural networks. Simulated trading strategies based on the forecast results are also required.

Requirement: Knowledge of course STAT3612 or STAT8017, AI/machine learning/deep learning, and skills in statistical programming using either SAS, R, or C++.

Supervisor: **Dr. Z. Zhang**, zhangz08@hku.hk, Dept of Statistics and Actuarial Science

49. Financial data analysis

This project aims to analyze the financial data by using the time series models, causal semantics, or machine learning techniques. Students are expected to use these methodologies to analyze real data sets, and develop useful trading algorithms.

Requirement: At least one programming language and knowledge about financial time series analysis

Supervisor: **Prof. K. Zhu**, mazhuke@hku.hk, Dept of Statistics and Actuarial Science

********** END **********