THE UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Topics for STAT3799 Directed Studies in Statistics (6 credits)
(Offered in both 1st and 2nd semesters of 2021 - 2022 for STAT3799)

1. Statistical learning with fewer labeled data

Modern statistical learning models such as deep neural networks can achieve very good performances in a wide range of learning tasks such as speech recognition and image processing. However, these learning models typically rely on the availability of a large number of labeled training data, and in many learning tasks, labeled training data may be very difficult to obtain. For example, in image classification problems, the labels of the images in the training data set are usually given manually by human labelers, and thus can be expensive to obtain. To investigate this issue, this project aims at studying learning methods that work even when there is only a limited number of labeled data.

Possible directions include:
(1) Active learning. Active learning is a learning paradigm that actively queries the data label during training. It is known to be effective in reducing human labeling efforts by actively selecting the most informative examples to label.
(2) Semi-supervised learning. Semi-supervised learning is an approach to machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training.
(3) Transfer learning/domain adaptation. Transfer learning applies the knowledge gained while solving one problem to a different but related problem. In this way, the learning of the related problem is possible even with limited data.

In this project, students will survey methods in the literature and run simulations/real-world experiments to compare the label complexity of different methods.

Requirement: Students are required to have basic knowledge in statistical learning and programming.

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

2. Copulas in Risk Management

Copulas are functions that join multivariate distribution functions to their one-dimensional marginal distribution functions. The student who takes this project is expected to study the basic theory of copula and some of its applications in risk management. All the related literature will be provided.

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

3. Chronological age prediction based on DNA methylation

Over the years, the correlation between DNA methylation levels and chronological age has been discovered in different species. In this project, we are going to use the 450K Human Methylation Beadchip data for human age prediction. High dimensional variable selection methods and machine learning models will be attempted. Students are expected to have good knowledge in programming languages such as R or Python.

Supervisor: **Prof. Tony W.K. Fung**, wingfung@hku.hk, Dept of Statistics and Actuarial Science

4. Test for genetic association accounting for X-chromosome inactivation

X-chromosome inactivation (XCI) is an important epigenetic factor in the study of genetic disease or phenotype. XCI is the mechanism by which gene dosage compensation is achieved between male mammals with a single copy of the X-chromosome and female mammals with two copies of the X-chromosome by inactivation of one of the two copies of the X-chromosome during early embryonic development. In this project, we will propose statistical tests for association on the X-chromosome accounting for XCI. Based on the constructed models, we aim to establish a theoretical and empirical framework for X-chromosome association testing of common variants, allowing for different degrees of X-inactivation. The performance of the proposed tests will be investigated by simulation and the analysis of a data set. Students are expected to write computer programs using R. Knowledge in genetics is preferred but not necessary.

Supervisor: **Prof. Tony W.K. Fung**, wingfung@hku.hk, Dept of Statistics and Actuarial Science

5. Machine learning methods for analysing single-cell genomic data

In recent years, the rapid development of single-cell sequencing technologies brings unprecedented opportunities to disentangle the heterogeneity in cell populations, including different immune cells, differentiation trajectory or cancer mutations. However, it remains highly challenging to decipher how a biological system functions and the underlying patterns of the data, not only because of high technical noise but also the high dimensions of gene or other molecular feature space.

Therefore, this project aims to develop machine learning methods, likely in a form of probabilistic models or deep learning methods, to analyse single-cell genomic data. Experimental data is available for both model validation and biological exploration. The student is expected to have interests in biomedical data, but no previous experience is required.

Requirement: Experience with Python or R programming.

Supervisor: **Dr. Y.H. Huang**, yuanhua@hku.hk, School of Biomedical Sciences & Dept of Statistics and Actuarial Science

6. Applications of unsupervised learning

Unsupervised learning aims at representing structure in the input data, often by means of features. The resulting features can be used as input for classification tasks or as initialization for further supervised learning. Traditional principal component analysis, factor analysis and independent component analysis are some examples. Deep learning methods including autoencoders, variational inference, variational autoencoders and GAN models are also developed.

The objective of the project is to explore and compare various unsupervised methods.

- Literature review of various unsupervised learning in the recent years.
- Apply to a real data set to identify any hidden features.
- Conduct simulation of various scenarios and evaluate the accuracy of models with various measures.

Requirement: Knowledge in Python.
Knowledge in multivariate statistics and machine learning.

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

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

8. Big Data Analytics in Securities Market

This project aims to discover new knowledge of market volatility of US securities and Hong Kong securities. Students will learn some skills of financial data analysis, and use big data to analyze the securities markets. Some statistical techniques, social media analysis, and/or artificial intelligence methods will be explored in this project. Students with some basic knowledge in statistics, data mining, text mining, and/or programming will take an advantage. Students are expected to be willing to learn new methods and skills.

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

9. Building a Predictive Model to Determine Epidemic Disease Outbreak Risk

This project aims to develop a predictive model of epidemic disease outbreak risk. Students will research the environmental factors of epidemic disease outbreak, and explore to use a new AI method for building a predictive model. Students will learn the methods of contact analysis, network analysis, and artificial intelligence in this project. Students are expected to be interested in research, have some basic knowledge in programming, and use statistics and data mining software.

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

10. Data Visualization in Global Market Analysis

This project aims to discover new business opportunities between countries by analyzing the import and export trading data. Some analysis, such as network analysis, demand and supply analysis, etc., will be done in this project. Students will develop a knowledge map to visualize the import and export trading patterns, and new discovery. Students will learn how to apply statistical methods and data analysis theories in business analysis. Students who have basic knowledge in using statistics and data mining software are preferred.

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

11. 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

12. Post-model-selection Inference

When a correct statistical model is not known a priori, as is common in modern statistical analysis, one often needs to select a data-driven model before proceeding to statistical inference. Classical statistical theory is, however, developed under the assumption of a correct model and may no longer hold if the model is data-driven and is therefore random. This extra level of uncertainty induced by model selection should be taken into serious consideration for subsequent inferences, which requires almost a complete revision of existing statistical practice. This project investigates this problem and its potential solutions.

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

13. 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

14. 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

15. 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

16. Modeling of Social Media Data

In this project, the students will implement latent Dirichlet allocation models to analyse social media data to discover hidden semantic structure in the social media. The students need to know python programming languages and data crawling skills.

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

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

18. The Gambler's and Hot-Hand Fallacies Continued…

Probability fallacies have always been appealing to students and sometimes even practical in real life. In particular, the gambler's fallacy (or called Monte Carlo fallacy) as well as hot-hand fallacy in basketball shooting were widely studied in different fields such as sports, economics and finance, psychology, and others. The seemingly classical topic with controversial views from different people does not lose its popularity in the modern days.

This project aims to investigate the related fallacies with simulation work as well as applications in reality such as behavioural finance in investment or financial risk management.

Students taking this project are expected to do simulation study on top of the theoretical review of the concepts involved.

Some papers below can be accessed online in public domain or from HKUL by HKU students.

References:

- Miller, J. B. and Sanjurjo, A. (2018). Surprised by the Hot Hand Fallacy? A Truth in the Law of Small Numbers. *Econometrica*, 86(6), 2019-2047.
- Rabin, M. and Vayanos, D. (2010). The Gambler's and Hot-Hand Fallacies: Theory and Applications. *The Review of Economic Studies*, 77(2), 730-778.
- Croson, R. and Sundali, J. (2005). The Gambler's Fallacy and the Hot Hand: Empirical Data from Casinos. *Journal of Risk and Uncertainty*, 30(3), 195-209.
- Gilovich, T., Vallone, R., and Tversky, A. (1985) The Hot Hand in Basketball: On the Misperception of Random Sequences. *Cognitive Psychology*, 17(3), 295-314.

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

19. Modelling Inline Warrants in Hong Kong

In July 2019, Hong Kong Exchanges and Clearing Limited (HKEX) launched a new type of financial derivative called inline warrant (界內證), which is essentially a product of financial engineering in a way to capture the potential price fluctuation of the underlying stock or index within a particular range. Like other financial derivatives or structured products, the inline warrants can serve the market speculators as well as the need for risk management for hedging purpose.

This project aims to study the features of inline warrants currently issued in the Hong Kong market and to model their price using option pricing formulae and/or simulation. Comparison with other synthetic products or real products in other markets can be performed and simulation study based on stock price modelling should be carried out. Applications of inline warrants in risk management are also to be investigated.

Students taking this project are expected to have fundamental knowledge in financial derivatives and option pricing (e.g., a pass in STAT3618, STAT3910 or FINA2322). Basic modelling skills such as programming and/or spreadsheet modelling are needed.

References:

- https://www.hkex.com.hk/Products/Securities/Structured-Products/Overview?sc_lang=en#iw
- https://www.hkex.com.hk/-/media/HKEX-Market/Products/Securities/Naming-Conventions-of-Stock-Short-Name-by-Product-Types/English-Inline-Warrant-Factsheet-v2,-d-,0.pdf

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

20. A Statistical Study on Financial Market Anomalies

Market anomalies can refer to strange patterns in financial data which violate the efficient market hypothesis (EMH). Some famous market anomalies include weekend effect, January effect and size effect. Traders using technical analysis and trading strategies may earn abnormal profits from market inefficiency.

This project aims to study various market anomalies based on statistical analysis. Investigations should be made on the existence or significance of the effects of any market anomalies in various financial markets.

Students taking this project are expected to study the relevant literature and conduct statistical tests using real market data. Elementary programming skills may be required to process large amount of data.

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

21. Further Topics in Portfolio Theory

This project aims to study some further topics in portfolio theory which may be more advanced or less mentioned in a traditional study of the Markowitz's model and related topics.

A list of potential topics is given as follows:

- Black-Litterman model
- Kelly criterion
- Advanced topics in asset allocation, e.g., active asset management
- Beta estimation, e.g., Blume's adjustment method
- Tracking portfolio analysis

Students taking this project must first have fundamental knowledge in mean-variance analysis, or modern portfolio theory (MPT), as well as pricing models like the capital asset pricing model (CAPM). Basic modelling skills such as programming or spreadsheet modelling are needed.

Requirement: Pass in STAT3609 or FINA2320

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

22. On Randomized Algorithm to Graph Problems

Most graph problems are known to be NP complete or #P complete. While a deterministic linear time algorithm in solving those problem is not optimistic, there often exists linear time approximation algorithm by making use of randomization. The analysis of complexity and correctness often involves the use of probability theory. The student who takes this course is expected to study the basic theory in relation to randomized algorithm design and its application to some famous graph problems. All related literature will be provided.

Supervisor: **Dr. Jeff T.Y. Wong**, jefftywong@hku.hk, Dept of Statistics & Actuarial Science

23. Statistical Learning of Recurrent Events Data

Recurrent events data is an important type of survival data, which is frequently encountered in practice. There has been a vast literature on recurrent events data analysis. For example, well-known methods include modeling the intensity process of recurrent events and modeling the marginal hazard of each recurrent event or the gap time between recurrent events. In addition, recurrent events problems can also be fit into the paradigm of multi-state models, for which transition probability or transition intensity between states can be estimated to characterize event progression. Another popular approach for recurrent events data is to specify covariate effects on mean or rate functions of recurrent events. This type of approach is attractive because mean or rate functions are more intuitive to interpret than intensity or hazard functions. This project will assess and compare the above-mentioned approaches. In addition, we will also employ various machine learning/deep learning techniques and develop powerful predictive models for recurrent events data.

Requirement: R or Python.

Supervisor: **Dr. J. Xu**, xujf@hku.hk, Dept of Statistics & Actuarial Science

24. Change Measure: Survey and Applications

Change measure is a useful and powerful tool in a number of areas, including mathematical finance, actuarial science and probability theory. In this course, we will study various change measure techniques, such as Girsanov theorem, Esscher transform, and applications in option pricing, premium calculation. The student needs to have some probability background in order to take this course. An advanced probability course (with some measure theory included) is preferred.

Supervisor: **Prof. H. Yang**, hlyang@hku.hk, Dept of Statistics & Actuarial Science

25. Deep Learning with Application in Artificial Intelligence

This project will focus on extracting useful information from structured and unstructured data and formulating statistical models for inference and prediction. In particular, we will develop deep learning, including deep neural networks for imaging analysis and computer vision and natural language processing for text data analysis. Extensive computation will be needed and real data will be used for analysis and illustration.

Requirement: The student needs to have experience with Python and R programming.

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

26. Generalizable machine learning technology with application in medical image analysis

Medical imaging is a critical step in modern healthcare procedures and automatic medical image analysis is beneficial to computer-aided diagnosis, assessment, and therapy. While deep learning has achieved remarkable success in medical image analysis in recent years, most of the deep models lack the sense of reliability and generalization when facing new cases, which prohibits deep-learning-based methods from being translated into clinical practice. This project will explore and develop generalizable deep learning/machine learning techniques for medical image analysis. Real public benchmark data will be used for analysis and illustration.

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

27. Multi-task machine learning for joint diagnosis and prognosis of human cancers

With the tremendous development of artificial intelligence, many machine learning algorithms have been applied to the diagnosis of human cancers. Recently, rather than predicting categorical variables (e.g., stages and subtypes) as in cancer diagnosis, several prognosis prediction models basing on patients' survival information have been adopted to estimate the clinical outcome of cancer patients. However, most existing studies treat the diagnosis and prognosis tasks separately. This project will explore and develop multi-task multi-modal machine learning techniques for joint diagnosis and prognosis of cancers from multi-modality medical data, such as histopathological image and genomics data. Public real dataset will be used for analysis and illustration.

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

28. Optimality Studies with Dependent Risks

Due to the complexity of modern insurance and financial products, contemporary insurance risk models have taken many realistic features into consideration. In the actuarial literature, the incorporation of realistic features such as dividends, investment and reinsurance into the basic insurance risk process has generated a lot of interesting research on optimality in the past two decades. This project aims at studying optimal dividends, investment and/or reinsurance for an insurance risk models with dependent risks.

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

29. Statistical modelling for biological/medical data

(This project will only be offered in Semester 2 only.)

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

30. Non-stationary Multiple Output Gaussian Processes

This project focuses on developing a non-linear Gaussian process regression model with multiple outputs that can capture changes in the functional behavior over time (non-stationary behavior). Students will learn about Bayesian machine learning. The project will involve extensive amounts of programming and computation. Interested students are required to have experience in Python.

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

31. 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

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

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