

A Short Course

Machine Learning for Data Mining

organized by Pattern Recognition and Machine Intelligence Association (PREMIA)

11 & 12 May 2010 (Tue & Wed), 9.00 am – 5.30 pm

Seminar Room 2, School of Computing, Computing Drive, NUS, Singapore 117417

This course comprises a tutorial on the fundamentals of machine learning with data mining in mind, and a series of invited talks on various practical applications of data mining. Some mathematical ability is assumed.

Registration Fees

Members of PREMIA: S\$400.00

Non-members: S\$500.00

Student members of PREMIA: S\$150.00 (Limited seats)

Student non-members: S\$200.00 (Limited seats)

The registration fee includes course notes, refreshments, and a one-year free PREMIA membership subscription. For non-members, entrance fee to PREMIA membership is waived.

Registration Procedure

Please register online at www.premia-sg.org. Make your cheque payable to PREMIA and send it to PREMIA's treasurer Dr. Zhou Jiayin as follows: Dr. Zhou Jiayin, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632.

If you wish to do online fund transfer, please email Dr. Zhou Jiayin (jzhou@i2r.a-star.edu.sg) for PREMIA account information.

Registration will close on 30 April 2010. However, due to space limitation, the registration may close before the deadline if the class limit is reached. Please register early to avoid disappointment. If your organization needs an invoice for the course fee registration, or if you encounter any problem during registration, please contact Ms Gong Tianxia at gong_tianxia@comp.nus.edu.sg

PREMIA reserves its right to cancel the course due to circumstances beyond its control.

Day 1 (11 May 2010): Machine Learning for Data Mining (by Dr. David Hardoon)

9:00 am – 10:30 am	Intro to Machine Learning and Kernel Methods (Linear Discriminant Learning; Perceptron Algorithm)
10:30 am – 10.45 am	Tea Break
10:45 am – 12:00 noon	Intro to Machine Learning and Kernel Methods – continued (SVM and Kernel Methods and non linearity)
12:00 noon – 1:00 pm	Lunch (on your own)
1:00 pm – 2:30 pm	Features (feature selection and how they relate to real-world problems, dimensionality reduction, subspace representation, learning in subspace)
2:30 pm – 3:30 pm	Applications of Machine Learning (Biology, Aerospace, etc)
3 :30 pm – 3:45 pm	Tea Break
3:45 pm – 5:30 pm	Data Mining Fundamentals (Goals of Data Mining, Classification, Association Analysis, Cluster Analysis, Anomaly Detection)

Biodata: David Hardoon completed his B.Sc. in Computer Science and Artificial Intelligence with first class honours at Royal Holloway, University of London within the Department of Computer Science. After completion of his undergraduate, he started at Royal Holloway to study for a PhD in Neural and Computational Learning, where he also worked part time on the KerMIT and thereafter full time on the

LAVA European projects. He later transferred at the start of his second PhD year to University of Southampton at the Information: Signals, Images, Systems research group, where he has completed his PhD. He is registered as a PASCAL researcher. As of October 2009, he is working as a Research Fellow at the Data Mining Department in the Institute for Infocomm Research, I²R, A*STAR. He is also an Honorary Senior Research Associate at University College, London in the Computer Science Department, Intelligent Systems group, at the Centre for Computational Statistics and Machine Learning.

Day 2 (12 May 2010): Invited Talks on Data Mining Applications

9:00 am – 10:10 am	Privacy Preserving Data Mining (by Dr. Mafruz Zaman Ashrafi)
10:10 am – 10:25 am	Tea Break
10:25 am – 11:35 am	Data Mining for Sentiment Analysis (by Mr. Mak Mun Thye)
11:35 am – 12:45 pm	Data Mining for Recommender Systems (by Dr. Yap Ghim Eng)
12:45 pm – 1:45 pm	Lunch (on your own)
1:45 pm – 2:55 pm	Exploratory data mining for Hypothesis Generation (by Dr. Feng Mengling)
2:55 pm – 4:05 pm	Image and video data mining (by Dr. Yuan Junsong)
4:05 pm – 4:20 pm	Tea Break
4:20 pm – 5:30 pm	Analytics in Retail Consumer Finance (by Mr. Nathaniel B. Noriel)

Talks' Synopses and Speakers' Biodata

1. Privacy Preserving Data Mining (by Dr. Mafruz Zaman Ashrafi)

Data mining techniques have capability to explore and fully exploit enormous quantities of data. As the advances in hardware technology have increased the capability to store and record personal data about consumers and individuals, data mining techniques raise concerns regarding privacy. The privacy preserving data mining seminar will highlight a number of techniques to perform the data mining tasks in a privacy-preserving way.

Biodata: "Mafruz Zaman Ashrafi is currently a Research Fellow at the Data Mining Department, Institute for Infocomm Research. He received his Ph.D. degree in Computer Science from Monash University, Melbourne in 2006. His research interests include Data mining, Privacy and Security, Social Networking, E-commerce. He has been serving as the members of Editorial Review Board (ERB) of the "Advances in Data Warehousing and Mining" Book Series published by IGI Press, Inc (USA)."

2. Data Mining for Sentiment Analysis (by Mr. Mak Mun Thye)

Sentiment analysis is also known as opinion mining and is a recent subdiscipline at the crossroads of information retrieval and computational linguistics. Sentiment analysis is not wholly concerned with the topic of a text but the opinion or sentiment that it expresses instead. Sentiment analysis finds a place in the industry as a means of judging opinions made on products by critics, or even as a way of tracking public opinion on various political issues. In this talk, we will take a look at the three subtasks of sentiment analysis, namely determining subjectivity/objectivity of text, determining positive/negative opinion of text, and determining the strength of positive/negative-polarity of text. We will also look at an application of sentiment analysis in a real world scenario as a demonstration of what sentiment analysis can offer.

Biodata: Mak MunThye obtained his bachelor's degree in Computer Science with University Honours from the School of Computer Science in Carnegie Mellon University under the National Science Scholar (Bachelor's) Scholarship awarded by the Agency for Science, Technology and Research

(A*STAR). His research focus has been on machine learning and data mining over the text domain. His is currently focused on teaching machines to learn the sentiments of people who contribute to the rising social media. Mak is currently working in the Institute for Infocomm Research (I²R), A*STAR.

3. Data Mining for Recommender Systems (by Dr. Yap Ghim Eng)

We face limited resources and numerous decisions every day. From simple meals and leisure decisions like what to eat for lunch or what shows, music and books to occupy our free time, to costly travel, housing and study plans, there are often more available options than we can digest. Wouldn't life be simpler if we have a personal assistant to filter and shortlist the top candidates in each case, so we can get the best deals without having to consider everything ourselves? Recommender systems, which automatically rank the set of possible choices to highlight the ones that best matches our needs, are specifically designed to serve this purpose. Over time, the recommenders accumulate large amount of information about what different users have chosen. Studying these interactions enables the systems to learn about items' suitability for various user types. How do the recommender systems automatically discover such knowledge from data? Data mining, the process of finding relevant and useful patterns or relationships from large data, is therefore an important component for effective recommendation systems.

Biodata: Yap Ghim-Eng received his Ph.D. degree from the Nanyang Technological University in 2009, under a full-time graduate scholarship awarded by the Agency for Science, Technology and Research (A*Star). He received his Bachelor of Computer Engineering degree – with First Class Honors – from the same university in 2004. He is currently a Research Engineer in the Data Mining Department at the Institute for Infocomm Research (I²R). Dr. Yap's main research areas include context-awareness, recommendation systems, reasoning under uncertainty, as well as causal interpretation.

4. Exploratory Data Mining for Hypothesis Generation (by Dr. Feng Mengling)

More and more data have been accumulated and stored in digital format in various applications. These data provide rich sources for making new discoveries. Data mining has become an important tool to transform data into knowledge. Finding useful and actionable knowledge is the main objective of diagnostic data mining. Most existing works tackle the problem by discovering patterns and rules and then studying their interestingness. In this talk, we explore a different paradigm which represents the discovered knowledge in the form of hypotheses. A hypothesis involves a comparison of two or more samples, and it is much closer to how human obtain knowledge. Compared with patterns and rules, hypotheses provide the context in which a piece of information is interesting, thus hypotheses are more intuitive and informative than patterns and rules. More importantly, users can take actions more easily based on what a hypothesis indicates. We will also investigate the reasons behind the discovered significant hypotheses, so that users not only get to know what is happening but also have some rough ideas on when or why it is happening. This new data mining paradigm has the potential to make diagnostic data mining as successful as predictive data mining in real-life applications.

Biodata: Feng Mengling studied both his bachelor and PhD degrees from the school of Electrical & Electronic Engineering, Nanyang Technological University. His research focus has been on fundamental data mining, clinical data mining, graph mining & bioinformatics. He is currently focusing on extracting knowledge that is really practical, useful and meaningful. His research work has been published in various reputable international conferences and journals, such as PODS,

PAKDD, ICME, Computational Intelligence, etc. Feng is currently working in Institute for Infocomm Research (I²R), A*STAR.

5. Image and Video Data Mining (by Dr. Yuan Junsong)

The recent advances in the image data capture, storage and communication technologies have brought a rapid growth of image and video contents. It thus becomes an emerging task to discover useful knowledge from image and video data. Despite a lot of previous work, data mining techniques that are successful in text and transaction data cannot simply apply to image and video data that contain much more complex structure. Due to the structure and content variations of the visual patterns, it is not a trivial task to discover meaningful patterns from images and videos. This talk concentrates the problem of mining interesting patterns from the image and video data, which are crucial for the research and applications of indexing and understanding them. Three effective data mining methods toward knowledge discovery from image and video data will be discussed: common object discovery from images, semantically meaningful visual pattern discovery, and recurring events mining from videos. The difficulties of structure and content variations for mining complex visual patterns are addressed and efficient algorithms are proposed to handle the large image and video dataset.

Biodata: Yuan Junsong received his Ph.D. and M.Eng from Northwestern University and National University of Singapore respectively, both in electrical engineering. Before that, he graduated from the special program for the gifted young in Huazhong University of Science and Technology, Wuhan, P.R.China. He is currently a Nanyang assistant professor at the school of EEE, Nanyang Technological University. During the summer 2008, 2007 and 2006, he was a research intern with Microsoft Research, Redmond, Kodak Research Laboratories, Rochester, and Motorola Laboratories, Schaumburg, respectively. From 2003 to 2004, he was a research student at the Institute for Infocomm Research, Singapore. He was a recipient of the Doctoral Spotlight Award from IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'09), and a recipient of the Nanyang Assistant Professorship from Nanyang Technological University in 2009. When he was in P.R.China, he was awarded the National Outstanding student by the Ministry of Education in 2001. His current research interests include computer vision, image and video data mining and content analysis, machine learning, multimedia search, etc.

6. Analytics in Retail Consumer Finance (by Mr. Nathaniel B. Noriel)

The development and maintenance of analytic models is now a mainstream business function in many of the world's leading banks and consumer finance companies. The use of analytics within this particular industry is quite mature due to the nature of the products involved and the amount and type of data collected from its customers. The speaker will discuss some of the main types of models used to support both the risk and reward aspects of the business, respectively for credit risk management and for customer relationship management.

Biodata: Nathaniel B. Noriel is currently Manager of Customer Analytics in NTUC Income. His career in data mining/statistics/analytics has spanned a number of industries including banking/consumer finance, telecoms and the public sector. He holds a BSc (Hons) in Mathematics and Economics from the University of Warwick, an MSc in Operational Research and Management Science from the University of Edinburgh, and an MSc in Statistics from the National University of Singapore. He was President of the Singapore Institute of Statistics from 2004- 2006 and is a Fellow of the Royal Statistical Society.